

# Cross-Selling in a Call Center with a Heterogeneous Customer Population

## Technical Appendix

Itay Gurvich<sup>†</sup>      Mor Armony\*      Costis Maglaras<sup>‡</sup>

### A Introduction

This is the technical appendix accompanying the paper, “Cross-Selling in a Call Center with a Heterogeneous Customer Population,” [3]. The organization of this appendix is as follows: we begin in §B with the completion of the proof of Proposition 1, whose sketch was given in §A of [3]. We continue in §C with some preliminaries required for the performance analysis of (S)-(C). Specifically, we provide a sample-path construction that uses a collection of independent rate-1 Poisson processes. We also discuss some strong approximation tools. Finally, in §D we prove the main performance-analysis results which are used in the proof in §A of the main paper [3]. Some auxiliary results are proved in §E.

### B A Detailed Proof of Proposition 1

This section is dedicated to the completion of the proof of Proposition 1 in [3]. Following our proof sketch in §A of the main paper [3] we show that there exists  $\delta^0 \geq 0$  such that for any sequence of initial states  $\{\xi^n\} \subseteq \Xi$  with  $|\xi^n| \rightarrow \infty$ :

$$\limsup_{n \rightarrow \infty} \frac{1}{|\xi^n|} E_{\xi^n} [|\Xi(|\xi^n|(1 + \delta^0))|] = 0. \quad (\text{A1})$$

Whenever this holds we can always find a positive number  $K$ , such that for all  $|\xi| > K$ , (20) holds. Toward that end, let  $V(t) = \sum_{j \geq 1} v_j(t)$  be the amount of residual work in the system. Also, let

---

<sup>†</sup>Columbia Business School, 41 Uris Hall, 3022 Broadway, NY, NY 10027. ([ig2126@columbia.edu](mailto:ig2126@columbia.edu))

\*Stern School of Business, NYU, 44 West 4th Street, NY, NY 10012. ([marmony@stern.nyu.edu](mailto:marmony@stern.nyu.edu))

<sup>‡</sup>Columbia Business School, 409 Uris Hall, 3022 Broadway, NY, NY 10027. ([c.maglaras@gsb.columbia.edu](mailto:c.maglaras@gsb.columbia.edu))

$V_s(t)$  and  $V_q(t)$  be, respectively, the residual work for the customers in service at time  $t$ , i.e.  $V_s(t) = \sum_{j \leq N} v_j(t)$ , and the residual work for the customers in queue at time  $t$ , i.e.  $V_q(t) = V(t) - V_s(t)$ . Set

$$\frac{1}{\bar{\mu}} = \sum_{i=1}^K \frac{\lambda_i}{\Lambda} \left( \frac{1}{\mu^s} + \frac{1}{\mu_i^{cs}} \right),$$

so that  $1/\bar{\mu}$  is the mean customer handling time in case cross-selling is performed. The following Lemma is the analogue of Lemma 4.3 in Dai [2]. Its proof is the same and is hence omitted.

**Lemma 4** *Fix a sequence of initial conditions  $|\xi^n|$  with  $|\xi^n| \rightarrow \infty$ . Then,*

(a) *Almost surely, uniformly on compact sets,*

$$\limsup_n \frac{V_s(|\xi^n|t)}{|\xi^n|} \leq N(1-t)^+, \quad (\text{A2})$$

and

$$\limsup_n \frac{V_q(|\xi^n|t)}{|\xi^n|} \leq 1 + \frac{\Lambda}{\bar{\mu}}t, \quad (\text{A3})$$

(b) *For each  $t \geq 0$ , the sequences  $\left\{ \frac{V_s(|\xi^n|t)}{|\xi^n|}, n \geq 1 \right\}$  and  $\left\{ \frac{V_q(|\xi^n|t)}{|\xi^n|}, n \geq 1 \right\}$ , are uniformly integrable.*

Finally, for any fixed  $t \geq 0$

$$\limsup_n \frac{1}{\mu^s} \frac{Q(|\xi^n|t)}{|\xi^n|} \leq \limsup_n \frac{V_q(|\xi^n|t)}{|\xi^n|} \leq \limsup_n \frac{1}{\bar{\mu}} \frac{Q(|\xi^n|t)}{|\xi^n|}, \quad (\text{A4})$$

and

$$\limsup_n \frac{1}{\mu^s} \frac{E_{\xi^n}[Q(|\xi^n|t)]}{|\xi^n|} \leq \limsup_n \frac{E_{\xi^n}[V_q(|\xi^n|t)]}{|\xi^n|} \leq \limsup_n \frac{1}{\bar{\mu}} \frac{E_{\xi^n}[Q(|\xi^n|t)]}{|\xi^n|}. \quad (\text{A5})$$

For the following let  $D(t)$  be the cumulative number of customer departure from the system up to time  $t$ . Also, let  $W_i(t)$  be the virtual waiting time for type- $i$  customers at time  $t$ . Since customers are assumed to be served FCFS we have that  $W_i(t) = W_j(t)$  for all  $i \neq j$ . Hence, we can define  $W(t)$  to be the virtual waiting time for all the customers, regardless of their type. Using the standard notation, we let  $D^d[0, \infty)$  be the space of functions  $f(\cdot) : [0, \infty) \mapsto \mathbb{R}^d$  that are *Right Continuous with Left Limits* (RCLL). A sequence of processes,  $X^n$ , in  $D^d$  is said to be

**C-tight** if, in addition to being tight (as random elements of  $D^d$ ), every convergent subsequence converges to a limit that is a.s. continuous. The following lemma establishes C-tightness of the different processes in consideration and identifies important properties of the limit points. We let  $\mu^{max} = \max\{\mu^s, \mu_1^{cs}, \dots, \mu_K^{cs}\}$ .

**Lemma 5** *For any sequence of initial conditions  $\{\xi^n\} \subseteq \Xi$  with  $|\xi^n| \rightarrow \infty$  as  $n \rightarrow \infty$ , and on compact subsets of  $[1, \infty)$ , the sequence  $\left\{ \left( \frac{Q(|\xi^n|\cdot)}{|\xi^n|}, \frac{D(|\xi^n|\cdot)}{|\xi^n|}, \frac{V_q(|\xi^n|\cdot)}{|\xi^n|}, \frac{W(|\xi^n|\cdot)}{|\xi^n|} \right) \right\}$  is C-tight. Moreover, any limit point  $(\bar{Q}(t), \bar{D}(t), \bar{V}_q(t), \bar{W}(t))$  satisfies*

$$\bar{D}(t) \leq N\mu^{max}t, \tag{A6}$$

$$\bar{W}(t) \geq \frac{\bar{Q}(t)}{N\mu^{max}}, \tag{A7}$$

$$\frac{1}{\bar{\mu}}\bar{Q}(t) \geq \bar{V}_q(t), \tag{A8}$$

for all  $t \geq 0$ . Finally,

$$\bar{V}_q(t) \leq \left[ \bar{V}_q(1) + (t-1) \left( \frac{\Lambda}{\mu^s} - N \right) \right]^+, \tag{A9}$$

for all  $t \geq 1$ .

**Proof:** The proof of C-tightness for the sequence  $\left( \frac{Q(|\xi^n|t)}{|\xi^n|}, \frac{D(|\xi^n|t)}{|\xi^n|} \right)$  is reminiscent of the proof of Theorem 4.1 in Dai [2] and we omit it. The inequality (A6) is trivial to establish. We proceed then to prove (A7). The virtual waiting time satisfies the representation

$$W(t) = \inf\{s \geq 0 : D(t+s) - D(t) \geq Q(t)\}, \tag{A10}$$

and in particular,

$$\frac{W(|\xi^n|t)}{|\xi^n|} = \inf \left\{ s \geq 0 : \frac{D(|\xi^n|(t+s)) - D(|\xi^n|t)}{|\xi^n|} \geq \frac{Q(|\xi^n|t)}{|\xi^n|} \right\}.$$

Considering a convergent subsequence  $\{n_j\}$  of  $(Q(|\xi^n|t)/|\xi^n|, D(|\xi^n|t)/|\xi^n|)$ , we can now apply the corollary in [9], to conclude that  $W(|\xi^{n_j}|t)/|\xi^{n_j}|$  also converges to a limit  $\bar{W}(t)$ . Consequently, the sequence  $(Q(|\xi^n|t)/|\xi^n|, D(|\xi^n|t)/|\xi^n|, W(|\xi^{n_j}|t)/|\xi^{n_j}|)$  is also C-tight. Moreover, any limit point satisfies  $\bar{W}(t) \geq \bar{Q}(t)/N\mu^{max}$ . The inequality (A8) follows from Lemma 4. Finally, to establish (A9), fix  $\tilde{\epsilon} > 0$  and set  $\tau^n := \inf\{t \geq 1 : V_q(|\xi^n|t) \leq \tilde{\epsilon}|\xi^n|\}$ . Then, for  $1 \leq t \leq \tau^n$  and by work

conservation

$$V_q(|\xi^n|t) + V_s(|\xi^n|t) = V_q(|\xi^n|1) + V_s(|\xi^n|1) - N|\xi^n|(t-1) + \sum_{l=A(|\xi^n|)}^{A(|\xi^n|t)} s_l(W(\tau_l)). \quad (\text{A11})$$

Here  $\tau_l$  is the time of the  $l^{\text{th}}$  arrival. Consider a convergent subsequence

$$\left( \frac{Q(|\xi^{n_j}|t)}{|\xi^{n_j}|}, \frac{D(|\xi^{n_j}|t)}{|\xi^{n_j}|}, \frac{W(|\xi^{n_j}|t)}{|\xi^{n_j}|} \right).$$

Then, we claim that uniformly on compact subsets of  $[1, \tau^{n_j})$ ,

$$y^{n_j}(t) := \frac{\sum_{l=A(|\xi^{n_j}|)}^{A(|\xi^{n_j}|t)} s_l(W(\tau_l))}{|\xi^{n_j}|} \rightarrow \frac{\Lambda}{\mu^s}(t-1), \text{ in probability, as } j \rightarrow \infty. \quad (\text{A12})$$

To prove this claim, fix  $T > 1$  and consider the set

$$\Omega^{n_j} := \left\{ \omega \in \Omega : \inf_{1 \leq t \leq T} W(|\xi^{n_j}|t) \leq \frac{1}{2} \frac{\bar{\mu} V_q(|\xi^{n_j}|t)}{N \mu^{max}} \right\}.$$

By inequalities (A7) and (A8),  $P(\Omega^{n_j}) \rightarrow 0$  as  $j \rightarrow \infty$ . Then,

$$P \left\{ \sup_{1 \leq t \leq T \wedge \tau^{n_j}} \left| y^{n_j}(t) - \frac{\Lambda}{\mu^s}(t-1) \right| > \epsilon \right\} \leq P \left\{ \sup_{1 \leq t \leq T \wedge \tau^{n_j}} \left| y^{n_j}(t) - \frac{\Lambda}{\mu^s}(t-1) \right| > \epsilon, (\Omega^{n_j})^c \right\} + P\{\Omega^{n_j}\} \rightarrow 0, \text{ as } j \rightarrow \infty. \quad (\text{A13})$$

for any  $\epsilon > 0$ , where the convergence

$$P \left\{ \sup_{1 \leq t \leq T \wedge \tau^{n_j}} \left| y^{n_j}(t) - \frac{\Lambda}{\mu^s}(t-1) \right| > \epsilon, (\Omega^{n_j})^c \right\} \rightarrow 0, \text{ as } j \rightarrow \infty,$$

follows from the definition of  $\Omega^{n_j}$  and a careful application of the strong law of large numbers using the fact that on  $(\Omega^{n_j})^c$ , for  $t \leq \tau^{n_j}$  and for any  $\epsilon > 0$  there exists  $j$  large enough so that for all  $k \geq j$ ,  $E[s_l(W(\tau_l))] \leq \frac{1}{\mu^s} + \epsilon$ . Now, Let

$$\tilde{y}(t) := \left[ \bar{V}_q(1) + (t-1) \left( \frac{\Lambda}{\mu^s} - N \right) \right] \vee \frac{3}{2} \tilde{\epsilon}.$$

Combining (A11) and (A13) we then have that for any  $\epsilon > 0$ ,

$$P \left\{ \sup_{0 \leq t \leq \tau^{n_j} \wedge T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\} \rightarrow 0, \quad \text{as } j \rightarrow \infty. \quad (\text{A14})$$

Define now two random times as follows:

$$\tilde{\tau}^n = \inf\{t \geq \tau^n : V_q(|\xi^n t|) > 2\tilde{\epsilon}|\xi^n|\} \quad \text{and} \quad \tau'^n = \sup\{t \leq \tilde{\tau}^n : V_q(|\xi^n t|) \leq \tilde{\epsilon}|\xi^n|\}.$$

Then, we can extend the arguments we used above to show that for any  $\epsilon > 0$ ,

$$P \left\{ \sup_{\tau'^{n_j} \wedge T \leq t \leq \tilde{\tau}^{n_j} \wedge T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\} \rightarrow 0, \quad \text{as } j \rightarrow \infty.$$

Since for  $\epsilon < \tilde{\epsilon}/2$ ,

$$\{\tilde{\tau}^{n_j} < T\} \subset \left\{ \sup_{\tau'^{n_j} \wedge T \leq t \leq \tilde{\tau}^{n_j} \wedge T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\},$$

we have that  $P\{\tilde{\tau}^{n_j} < T\} \rightarrow 0$ , as  $j \rightarrow \infty$ . In particular, since

$$\left\{ \sup_{0 \leq t \leq T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\} \subseteq \left\{ \sup_{0 \leq t \leq \tilde{\tau}^{n_j} \wedge T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\} \cup \{\tilde{\tau}^{n_j} < T\},$$

we conclude that for arbitrarily small  $\epsilon$ ,

$$P \left\{ \sup_{0 \leq t \leq T} \left[ \frac{V_q(|\xi^{n_j} t|)}{|\xi^{n_j}|} - \tilde{y}(t) \right]^+ > \epsilon \right\} \rightarrow 0, \quad \text{as } j \rightarrow \infty. \quad (\text{A15})$$

The result of the Lemma now follows since  $\tilde{\epsilon}$  was arbitrary. ■

To complete the proof of Proposition 1, note that with  $\delta^0 = (1 + \Lambda/\bar{\mu})(N - R)$ , and since by Lemma 4,  $\bar{V}_q(1) \leq 1 + \Lambda/\bar{\mu}$ , we have that

$$\bar{V}_q(1) + (\delta^0) \left( \frac{\Lambda}{\mu^s} - N \right) \leq 0.$$

To conclude the argument, we fix an arbitrary sequence of initial conditions  $\{\xi^n\}$ . Consider a convergent subsequence  $\{n_j\}_{j \geq 1}$ , of  $\frac{E[V_q(|\xi^n|(1+\delta^0))]}{|\xi^n|}$ . By Lemma 5 each subsequence  $\{n_{j_k}\} \subset \{n_j\}$

such that  $\frac{V_q(|\xi^{n_{j_k}}|(1+\delta^0))}{|\xi^{n_{j_k}}|}$  converges satisfies

$$\lim_{k \rightarrow \infty} \frac{E[V_q(|\xi^{n_{j_k}}|(1+\delta^0))]}{|\xi^{n_{j_k}}|} = 0.$$

In particular,

$$\limsup_{n \rightarrow \infty} \frac{E[V_q(|\xi^n|(1+\delta^0))]}{|\xi^n|} = 0.$$

Since the sequence  $\{\xi^n\}$  was arbitrary the argument is complete. ■

## C Sample Path Construction, Strong Approximations and Other Preliminaries

For the rest of our results we replace the sample path construction from Proposition 1 with a different one that takes advantage of the properties of the Poisson process. This alternative construction follows an approach that is, by now, quite common; see Whitt [4] for an overview. Having constructed the sample paths using Poisson processes we can use strong approximations to bound the distance of the underlying Poisson process from their respective rates. This section is divided then to two subsections. First, in §C.1 we construct the sample path. Then, in §C.2 we introduce the strong approximation tools that will be required for our proofs.

### C.1 Sample-Path Construction

We base the sample-path construction on independent unit-rate Poisson processes  $A$ ,  $S_1$  and  $S_{i,2}$ ,  $i = 1, \dots, \bar{k}$ , where  $\bar{k}$  is the last customer class that is cross-sold under the control (C); see §3.1 in [3]. We generate arrivals of customers using the Poisson process  $A^\Lambda(t) := A(\Lambda t)$ . The arrival times are registered in a dynamic array in which the customers are ordered in order of their arrival times. Denote the value of this array at time  $t \geq 0$  by the vector  $\mathcal{T}(t)$ . Let  $\tilde{\mathcal{T}}(t)$  be the process obtained by setting  $\tilde{\mathcal{T}}_k(t) = t - \mathcal{T}_k(t)$ . Whenever a customer leaves the queue to be served he is erased from this array. We hold an  $N^\Lambda$  dimensional vector  $w(t) = (w_1(t), \dots, w_{N^\Lambda}(t))$  with  $w_j(t)$ ,  $j = 1, \dots, N^\Lambda$  standing for the registered waiting time of the customer currently in service with agent  $j$ . This value is registered immediately when the customer is admitted to service.

We let  $\mathcal{S}(t)$  be the set of agents giving service at time  $t$ , i.e.,  $\mathcal{S}(t)$  contains all the agents that are

not idle and not cross-selling at time  $t$ . We generate the process of phase 1 (service) completions with type- $i$  customers using a time change of a unit-rate Poisson process. Specifically, let  $D_1^\Lambda(t)$  be the number of phase 1 completion up to time  $t$ . Then, we write

$$D_1^\Lambda(t) := S_1 \left( \mu^s \int_0^t Z_1^\Lambda(s) ds \right), \quad (\text{A16})$$

where  $Z_1^\Lambda(t)$  is the number of agents giving service at time  $t$ . Whenever the process  $D_1^\Lambda(t)$  jumps we generate a discrete random variable with values in  $\mathcal{S}(t-)$  to identify the number of the agent that completed service. Formally, for a subset  $J$  of  $\{1, \dots, N^\Lambda\}$  we let  $\{e_s(J, l), l \in \mathbb{N}\}$  be a sequence of i.i.d random variables distributed uniformly over  $J$ . We will use these to determine the actual agents that completes services. Also, we let  $\{e_c(l), l \in \mathbb{N}\}$  be a sequence of i.i.d discrete random variables on  $\{1, \dots, K\}$  with  $P\{e_c(1) = i\} = \lambda_i/\Lambda$ . Finally, we let  $\{e_{cs}(l), l \in \mathbb{N}\}$  be i.i.d uniform random variables on  $[0, 1]$ .

The number of service completions followed by a cross-selling phase with a type- $i$  customer,  $\tilde{D}_i(t)$ , for  $i \leq \bar{k}$  is then given by

$$\tilde{D}_i^\Lambda(t) = \sum_{l=1}^{D_1^\Lambda(t)} \sum_{j \in \mathcal{S}(t-)} \mathbf{1}\{e_s(\mathcal{S}(t-), l) = j, e_c(l) = i, e_{cs}(l) \leq q_i(w_j(t-)), Q^\Lambda(t-) \leq \eta_i \Lambda\}.$$

Here we used the fact that an agent that completes service will attempt cross-selling to a type- $i$  customer only if at the time of completion the queue is less than the threshold  $\eta_i^\Lambda$ . The process  $\tilde{D}_i^\Lambda(t)$  is then a non-homogenous Poisson process with an instantaneous rate that is bounded from above by

$$\sum_{i=1}^{\bar{k}} \frac{\lambda_i}{\Lambda} \mu^s q_i(0) Z_1^\Lambda(t) \mathbf{1}\{Q^\Lambda(t) \leq \eta_i\}.$$

Finally, for  $i \leq \bar{k}$ , we generate cross-selling completions by a non-homogenous Poisson process with an instantaneous rate  $\mu_i^{cs} Z_{i,2}^\Lambda(t)$  at time  $t$ , i.e, we write

$$D_{i,2}^\Lambda(t) = S_{i,2} \left( \mu_i^{cs} \int_0^t Z_{i,2}^\Lambda(s) ds \right),$$

where  $Z_{i,2}^\Lambda(t)$  is the number of type- $i$  customers undergoing cross-selling at time  $t$ .

The system state is then captured by the multi-dimensional Markov process

$$\Xi^\Lambda(t) := (Z_1^\Lambda(t), Z_{i,2}^\Lambda(t), \mathcal{S}^\Lambda(t), w^\Lambda(t), Q^\Lambda(t), \tilde{T}^\Lambda(t); i = 1, \dots, K). \quad (\text{A17})$$

We let  $\mathcal{X}$  be the domain of this process. The following Lemma is a corollary of Proposition 1 and its proof is easily obtained by expanding the state-space of the Markov process constructed in the proof of Proposition 1. Although the proof would use a different sample path construction, the uniqueness of the stationary distribution is invariant to this construction as the probability law is the same under both constructions.

**Lemma 6** *Fix  $\Lambda$ . Then the process  $\Xi^\Lambda(t)$  admits a unique stationary distribution  $\pi^\Lambda$  which coincides with the unique limit distribution.*

We let  $\xi$  be a general element in  $\mathcal{X}$ , and for given  $\xi$  we let  $\mathfrak{q}(\xi)$ ,  $z_1(\xi)$ , and  $z_{i,2}(\xi)$ ,  $i = 1, \bar{k}$  be respectively the queue length, the number of agents giving service and the number of agents cross-selling to a type- $i$  customer in state  $\xi$ . The notation  $\mathfrak{q}(\xi)$  should not be confused with the integer  $q$  that we will use as a general power nor with the delay sensitivity functions  $q_i(\cdot)$ .

We conclude this section with some additional notation. For fixed  $T > 0$ , a positive integer  $d$  and a function  $y \in D^d[0, \infty)$ , we define  $\|y(\cdot)\|_T := \sup_{0 \leq t \leq T} \sum_{i=1}^d |y(t)|$ . We let  $(\Omega, \mathcal{F}, P)$  be the probability space which will remain fixed throughout and use the notation  $\omega$  to denote an element in  $\Omega$ . We let  $P_\xi$  be the probability distribution under which  $P_\xi(\Xi^\Lambda(0) = \xi) = 1$ , and put  $E_\xi[\cdot]$  be the expectation operator with respect to to the probability distribution  $P_\xi$ . We let  $P_{\pi^\Lambda}$  be the probability distribution under which  $\Xi^\Lambda(0) \sim \pi^\Lambda$  where  $\pi^\Lambda$  is the unique stationary distribution from Lemma 6 and we define  $E_{\pi^\Lambda}[\cdot]$  accordingly. Finally for  $x, y \in \mathbb{R}$ , we use the standard notations  $x \vee y = \max\{x, y\}$  and  $x \wedge y = \min\{x, y\}$  as well as  $x^+ = \max\{x, 0\}$  and  $x^- = \max\{-x, 0\}$ .

## C.2 Strong Approximations

Time changes of unit-rate Poisson processes, as the ones we have used above, can be approximated by time-changed Brownian motion plus a logarithmic error term. We refer the reader to Mandelbaum et. al. [8] and the references therein for a detailed discussion of strong approximations and their application to Markovian queueing networks. For our purposes, it suffices to know that given a unit-rate Poisson process  $\mathcal{N}(\cdot)$  and an instantaneous bounded rate function  $0 \leq \lambda(t) \leq \bar{\lambda}$ , for

some  $\bar{\lambda} > 0$ , we have that

$$\sup_{t \geq 0} \frac{\mathcal{N} \left( \int_0^t \lambda(u) du \right) - \int_0^t \lambda(u) du - B \left( \int_0^t \lambda(u) du \right)}{\log(\bar{\lambda}t \vee 2)} \leq C,$$

where  $C$  is a non-negative random variable with

$$P\{C > \gamma + \beta x\} \leq c_1 e^{-c_2 x}, \quad (\text{A18})$$

for some strictly positive constants  $\gamma, \beta, c_1$  and  $c_2$ ; see Lemma 9.4 in [8]. Since the rate of service (or cross-selling completions) is bounded by  $N^\Lambda(\mu^s \vee \max_{i=1, \dots, K} \mu_i^{cs})$  and since the staffing rule (S) dictates  $N^\Lambda = R(1 + \bar{z})$  we can find  $m$  large enough so that all the instantaneous rates in the system are bounded  $m\Lambda$ . We henceforth fix  $m$  to be such a value. We can then define a  $3K$ -dimensional Brownian motion,  $B(t)$ , such that

$$\left( \left\| A^\Lambda(\cdot) - \Lambda \cdot \right\|_T + \left\| D_1^\Lambda(\cdot) - \mu^s \int_0^\cdot Z_1^\Lambda(s) ds \right\|_T + \sum_{i=1}^{\bar{k}} \left\| D_{i,2}^\Lambda(\cdot) - \mu_i^{cs} \int_0^\cdot Z_{i,2}^\Lambda(s) ds \right\|_T \right) \leq \|B(m\Lambda \cdot)\|_T + C \log(m\Lambda T \vee 2). \quad (\text{A19})$$

Finally, we will be using some basic facts about Brownian motion. Specifically, for a  $d$ -dimensional Brownian motion and a constant  $m > 0$ , it can be easily shown that for all  $x \geq 0, T > 0$ ,

$$P \left\{ \|B(m\Lambda \cdot)\|_T \geq x\sqrt{\Lambda} \right\} \leq c_3 \sqrt{\frac{T}{x}} e^{-c_4 \frac{x^2}{T}},$$

for some strictly positive constants  $c_3$  and  $c_4$ ; see e.g. Problem 2.8.2 in [7]. Using (A18) we then have that (A18) for all  $x > 0$

$$P \left\{ \|B(m\Lambda \cdot)\|_T + C \log(m\Lambda T \vee 2) \geq x\sqrt{\Lambda} \right\} \leq c_5 \left( 1 \vee \sqrt{\frac{T}{x}} \right) e^{-c_6 \min\{\frac{x^2}{T}, \frac{x}{\sqrt{T}}\}}, \quad (\text{A20})$$

for some strictly positive constants  $c_5, c_6$ . Denote by

$$\Omega^*(\Lambda, T, x) = \{\omega \in \Omega : \|B(m\Lambda \cdot)\|_T + C \log(m\Lambda T \vee 2) \leq x\}. \quad (\text{A21})$$

Then, by (A20)

$$P \left\{ \Omega^*(\Lambda, T, x\sqrt{\Lambda}) \right\} \geq \left( 1 - c_5 \left( 1 \vee \sqrt{\frac{T}{x}} \right) e^{-c_6 \min\left\{ \frac{x^2}{T}, \frac{x}{\sqrt{T}} \right\}} \right)^+.$$

## D Performance Analysis for (S)-(C)

The aim of this section is two-fold. First, we want to establish that the queue length is, in some sense, bounded by the smallest threshold  $\eta_k^\Lambda$ . Then, fixing  $\epsilon > 0$ , we want to show that the steady-state expected number of agents cross-selling type- $i$  customers satisfies

$$\left| E[Z_{i,2}^\Lambda(\infty)] - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right| \leq \epsilon \Lambda,$$

for all  $\Lambda$  large enough. Since  $\epsilon$  is arbitrary, we will have that

$$E[Z_{i,2}^\Lambda(\infty)] = \frac{\lambda_i q_i(0)}{\mu_i^{cs}} + o(\Lambda),$$

which is what is used in the proof of Theorem 1 in §A of the main paper [3].

All the subsequent proofs share the same basic ideas. Using the strong approximation construction we examine the behavior of the system on a subset of the sample paths (such as  $\Omega^*(\Lambda, T, x)$ ) where the stochastic fluctuations generated by the Brownian Motion are bounded. This allows us to examine a deterministic version of the system dynamics. For the deterministic version, and with arguments reminiscent of Lyapunov function tools used for stability proofs, we show that the system is in some sense attracted back into a certain domain. Finally, we remove the conditioning and apply some arguments from [5] to obtain the steady state bounds. Some of the arguments are common to several proofs. We abbreviate the proofs whenever this is the case. Throughout we fix  $T > 0$  and  $\epsilon > 0$ .

Most of the Propositions in this section share a common structure. The first part of each such proposition states a steady-state bound. The second part essentially states that if the process is initialized at time 0 close enough to its steady-state distribution (in a sense to be made precise), it actually stays there.

Our first result towards our stated goals shows that the number of agents busy giving service (not cross-selling) is close to the load  $R$ . It also shows that the queue length will not exceed, in

some sense, the largest threshold,  $\eta_1^\Lambda$ . We will use these results to refine the bound on the queue length process in Proposition 7. For the following we define

$$\mathcal{X}_1 := \{\xi \in \mathcal{X} : |z_1(\xi) - R| \leq \epsilon\Lambda, \mathbf{q}(\xi) \leq \eta_1^\Lambda + \epsilon\sqrt{\Lambda}\} \quad (\text{A22})$$

**Proposition 5** *Fix  $\epsilon > 0$  and  $q \geq 2$ . Then,*

$$P \{|Z_1^\Lambda(\infty) - R| > \epsilon\Lambda\} \leq \frac{c_7}{(\sqrt{\Lambda})^{q-1}}, \quad (\text{A23})$$

and

$$\limsup_{\Lambda \rightarrow \infty} E \left[ \left( (Q^\Lambda(\infty) - \eta_1^\Lambda)^+ \right)^{q-1} \right] \leq C_Q, \quad (\text{A24})$$

for all  $\Lambda$  large enough and some strictly positive constants  $c_7$  and  $C_Q$ . Moreover,

$$\sup_{\xi \in \mathcal{X}_1} P_\xi \{ \|Z_1^\Lambda(\cdot) - R\|_T \geq 2\epsilon\Lambda \} \leq c_9 e^{-c_{10}\sqrt{\Lambda}}, \quad (\text{A25})$$

and

$$\sup_{\xi \in \mathcal{X}_1} P_\xi \left\{ \left\| (Q^\Lambda(\cdot) - \eta_1^\Lambda)^+ \right\|_T \geq 2\epsilon\sqrt{\Lambda} \right\} \leq c_9 e^{-c_{10}\epsilon\sqrt{\Lambda}} \quad (\text{A26})$$

for all  $\Lambda$  large enough and for some strictly positive constants  $c_9$  and  $c_{10}$ .

Note that with the exception of the customer cross-selling probability being delay sensitive, whenever the queue-length process exceeds the greatest threshold,  $\eta_1^\Lambda$ , it behaves just as the queue length process in the single class model of Armony and Gurvich [1] when it exceeds the unique threshold there. Moreover, when the queue is above  $\eta_1^\Lambda$  the delay sensitivity does not play any role as by the control mechanism (C), no cross-selling will be attempted until the queue goes below  $\eta_1^\Lambda$ . The proof of this result will then follow some of the results in [1]. A detailed proof of Proposition 5 would be obtained by expanding the proofs of Lemma B.1 and Proposition B.1 in [1] with the unique threshold there replaced by  $\eta_1^\Lambda$ . We postpone the proof of this result to §E. We will also need the following Proposition which establishes, the otherwise intuitive result, that the number of agents busy cross-selling to class- $i$  customers cannot exceed  $\lambda_i q_i(0)/\mu_i^{cs}$  significantly. For the following we define

$$\mathcal{X}_2 := \{\xi \in \mathcal{X}_1 : |z_{i,2}(\xi) - \lambda_i q_i(0)/\mu_i^{cs}| \leq \epsilon\Lambda, \ i = 1, \dots, \bar{k}\}, \quad (\text{A27})$$

where  $\mathcal{X}_1$  was defined in equation (A22).

**Proposition 6** *Fix  $\epsilon > 0$ . Then, for all  $\Lambda$  large enough and all  $x > 0$ ,*

$$P \left\{ \left( Z_{i,2}^\Lambda(\infty) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right)^+ > \epsilon \Lambda \right\} \leq c_{11} e^{-c_{12} \sqrt{\Lambda}}, \quad i = 1, \dots, K.$$

for some strictly positive constants  $c_7$  and  $c_8$ . Moreover,

$$\sup_{\xi \in \mathcal{X}_2} P_\xi \left\{ \left\| \left( Z_{i,2}^\Lambda(\cdot) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right)^+ \right\|_T \geq 2\epsilon \Lambda \right\} \leq c_{11} e^{-c_{12} \sqrt{\Lambda}}, \quad i = 1, \dots, K, \quad (\text{A28})$$

for all  $\Lambda$  large enough and for some strictly positive constants  $c_{11}$  and  $c_{12}$ .

The proof of Proposition 6 is very similar to the proofs of the estimate for  $Z_1^\Lambda(t)$  that were stated in Proposition 5. The detailed proof is omitted, and we turn now to the finer analysis of the queue length and waiting time processes.

Towards than end, we let  $W_i^\Lambda(t)$  be the virtual waiting time for type- $i$  customers at time  $t$ . Since customers are served in a FCFS fashion, we have that  $W_i^\Lambda(t) = W_j^\Lambda(t)$  for all  $i \neq j$ . Hence, we may let  $W^\Lambda(t)$  be the common virtual waiting time. Since, by Proposition 1, steady state exists, the PASTA property guarantees that waiting time as seen by arriving customers is equal in distribution to the steady-state virtual waiting time. The following proposition shows that using (S)-(C) the queue length and the waiting time are small and in particular the queue exceeds the threshold  $\eta_{\bar{k}}$  at most by an amount that is  $o(\sqrt{\Lambda})$ .

We now turn to a more refined analysis of the queue length process. Towards this end, define the set

$$\mathcal{X}_3 := \{\xi \in \mathcal{X}_2 : \mathfrak{q}(\xi) \leq \eta_{\bar{k}}^\Lambda + \epsilon \sqrt{\Lambda}\},$$

where  $\mathcal{X}_2$  was defined in equation (A27). The following proposition is the most complicated one in this appendix as it deals with a refined analysis of the queue length behavior and in particular one that considers the behavior of the queue at an  $O(1)$  level. The other results are  $o(\Lambda)$  result and are hence much simpler. The more refined analysis for the queue length is however necessary for the other proofs as well as for the asymptotic feasibility result for the constrained case, as given in Theorem 3 of the main paper [3].

**Proposition 7** Fix an integer  $q \geq 2$ . Then,

$$\limsup_{\Lambda \rightarrow \infty} E \left[ \left( (Q^\Lambda(\infty) - \eta_k^\Lambda)^+ \right)^{q-1} \right] \leq C_q, \quad \text{and} \quad \limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda} E[W^\Lambda(\infty)] < \infty, \quad (\text{A29})$$

for some constant  $C_q$ . In particular, for all  $x > 0$  and all  $\Lambda$  large enough,

$$P\{(Q^\Lambda(\infty) - \eta_k^\Lambda)^+ > x\sqrt{\Lambda}\} \leq \frac{C_q}{(x\sqrt{\Lambda})^{q-1}}. \quad (\text{A30})$$

Also,

$$\sup_{\xi \in \mathcal{X}_3} P_\xi \left\{ \|(Q^\Lambda(\cdot) - \eta_k^\Lambda)^+\|_T \geq 2\epsilon\sqrt{\Lambda} \right\} \leq c_{11} e^{-c_{12}\epsilon\sqrt{\Lambda}}, \quad (\text{A31})$$

and

$$\sup_{\xi \in \mathcal{X}_3} P_\xi \left\{ \|W^\Lambda(\cdot)\|_T \geq \frac{M_2 + \epsilon}{\sqrt{\Lambda}} \right\} \leq c_{11} e^{-c_{12}\epsilon\sqrt{\Lambda}}, \quad (\text{A32})$$

for all  $\Lambda$  large enough and some strictly positive constants  $c_{11}, c_{12}$  and  $M_2$ .

**Remark 3 (Lemma 1 and Theorem 3 in [3])** Note that Lemma 1 in the main paper [3] is covered as a special case of Proposition 7. Moreover, when the threshold  $\eta_k^\Lambda$  is chosen according to the recommendation in §3.2 in the main paper [3], Theorem 3 there is a consequence of Proposition 7 above.

**Proof of Proposition 7:** Fix a constant  $K > 1$  and define a function  $\Phi(x) : \mathbb{R}_+ \mapsto \mathbb{R}_+$  as follows:

$$\Phi(x) = (x - \eta_k^\Lambda)^+ + K.$$

The proof now proceeds as follows: We first fix the integer  $q \geq 2$  and establish that there exist  $t^* > 0$  and  $M > 0$  so that

$$\sup_{\xi \in \mathcal{X}_1: \mathfrak{q}(\xi) > \eta_k^\Lambda + M} E_\xi \left[ \Phi(Q^\Lambda(t^*/\Lambda))^q \right] - \Phi(\mathfrak{q}(\xi))^q \leq -\gamma \Phi(\mathfrak{q}(\xi))^{q-1}, \quad (\text{A33})$$

for some  $\gamma > 0$ . We will then use Lemma 8 from §E and adapt the argument used in the proof of Theorem 5.1 in [5] to obtain a bound for the steady-state queue length process. The bound for the steady-state waiting time will then follow from an application of Little's law. Finally, we will establish the bound in (A31) and (A31).

We start, then, by establishing (A33). Towards this end, fix  $M > 0$  and assume that  $Q^\Lambda(0) >$

$\eta_{\bar{k}}^\Lambda + M$ . Fix  $0 < \eta \leq M/2$  and define the random time  $\tau^\Lambda = \inf \{t \geq 0 : Q^\Lambda(t) \leq Q^\Lambda(0) - \eta\}$ . Note that on  $[0, \tau^\Lambda \wedge T]$ , the queue length process  $Q^\Lambda(t)$  satisfies

$$Q^\Lambda(t) > \eta_{\bar{k}}^\Lambda, \text{ and } Q^\Lambda(t) = Q^\Lambda(0) + A^\Lambda(t) - \sum_{i=1}^{\bar{k}} D_{i,2}^\Lambda(t) - \tilde{D}_0^\Lambda(t),$$

where  $\tilde{D}_0^\Lambda(t)$  is the process of service completions **not followed** by a cross-selling phase, i.e.,  $\tilde{D}_0^\Lambda(t) = D_1^\Lambda(t) - \sum_{i=1}^{\bar{k}} \tilde{D}_i^\Lambda(t)$ . On  $[0, \tau^\Lambda]$  the instantaneous rate of  $\tilde{D}_0^\Lambda(t)$  can be bounded from below by

$$\mu^s Z_1^\Lambda(t) \left( \sum_{i=1}^{\bar{k}-1} \frac{\lambda_i}{\Lambda} (1 - q_i(0)) + \sum_{i=\bar{k}}^K \frac{\lambda_i}{\Lambda} \right),$$

which corresponds to the fact that, by the control (C), as long as the queue length is greater than  $\eta_{\bar{k}}$  all service completions of customers from type  $i \geq \bar{k}$  are followed by the admission to service of a customer that is waiting in the queue. Fix now  $\epsilon > 0$  and  $\delta > 0$  and define the set

$$\hat{\Omega}(\Lambda) = \left\{ \omega \in \Omega : \|Z_1^\Lambda(\cdot) - R\|_T \leq 2\epsilon\Lambda, \left\| Z_{i,2}^\Lambda(\cdot) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right\|_T \leq 2\epsilon\Lambda, i = 1, \dots, \bar{k} \right\} \cap \Omega^*(\Lambda, T/\Lambda, \delta),$$

where  $\Omega^*$  is as defined in (A21). Then, on  $\hat{\Omega}(\Lambda)$  and for  $t \leq \tau^\Lambda \wedge T/\Lambda$ ,

$$Q^\Lambda(t) \leq Q^\Lambda(0) + \Lambda t - \sum_{i=1}^{\bar{k}} \mu_i^{cs} \int_0^t Z_{i,2}^\Lambda(u) du - \mu^s \left( \sum_{i=1}^{\bar{k}-1} \frac{\lambda_i}{\Lambda} (1 - q_i(0)) + \sum_{i=\bar{k}}^K \frac{\lambda_i}{\Lambda} \right) \int_0^t Z_1^\Lambda(u) du + \delta,$$

Observe that for all  $t \leq \tau^\Lambda$  all servers are busy and consequently  $\sum_{i=1}^{\bar{k}} Z_{i,2}^\Lambda = N^\Lambda - Z_1^\Lambda$ . Since, by definition,  $N^\Lambda - R = \sum_{i=1}^{\bar{k}} \lambda_i q_i / \mu_i^{cs}$ , we have, after some straightforward algebra, that on  $\hat{\Omega}(\Lambda)$ ,

$$Q^\Lambda(t \wedge \tau^\Lambda) \leq Q^\Lambda(0) + C\Lambda\epsilon(t \wedge \tau^\Lambda) - \lambda_{\bar{k}} q_{\bar{k}}(0) \cdot t \wedge \tau^\Lambda + \delta, \quad (\text{A34})$$

for  $t \leq T/\Lambda$  and for some constant  $C > 0$ . Equation (A34) is the crucial one. It reflects the fact that once cross-selling to class  $\bar{k}$  is stopped, the capacity of the system is large enough to attract the aggregate queue back to  $\eta_{\bar{k}}$ , and it will do so at a rate of approximately  $\lambda_{\bar{k}} q_{\bar{k}}(0)$ . We can now choose  $\epsilon$  small enough so that

$$Q^\Lambda(t \wedge \tau^\Lambda) \leq Q^\Lambda(0) + \delta - \frac{1}{2} \Lambda \lambda_{\bar{k}} q_{\bar{k}}(0) \cdot t \wedge \tau^\Lambda, \quad (\text{A35})$$

where for all  $i \leq K$ ,  $\bar{\lambda}_i := \lambda_i/\Lambda$ . In particular, on  $\hat{\Omega}(\Lambda)$  we have that

$$\Lambda\tau^\Lambda \leq t^* := \frac{\eta + \delta}{\frac{1}{2}\bar{\lambda}_{\bar{k}}q_{\bar{k}}(0)}.$$

Choosing  $\delta \leq \eta/2$ , the same argument shows that on  $\hat{\Omega}(\Lambda)$  and for all  $\tau^\Lambda \leq t \leq t^*/\Lambda$ ,  $Q^\Lambda(t) \leq Q^\Lambda(0) - \eta/2$ . Indeed, along the arguments leading to (A35) one can establish that on  $\hat{\Omega}(\Lambda)$ , the queue length is a linearly decreasing process as long as  $Q^\Lambda(t) > \eta_{\bar{k}}^\Lambda$ . We conclude then that if  $Q^\Lambda(0) \geq \eta_{\bar{k}}^\Lambda + M$ ,

$$\Phi(Q^\Lambda(t^*/\Lambda)) - \Phi(Q^\Lambda(0)) \leq -\eta/2 \quad (\text{A36})$$

on  $\hat{\Omega}(\Lambda)$ . Since  $Q^\Lambda(t) \leq Q^\Lambda(0) + A^\Lambda(t)$ , it is straightforward to show that

$$\limsup_{\Lambda \rightarrow \infty} \sup_{\xi \in \mathcal{X}_1} E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))\mathbf{1}\{\hat{\Omega}^c\}] - \Phi(Q^\Lambda(0)) = 0.$$

Consequently, we have from (A36) that for all  $\Lambda$  large enough

$$\sup_{\xi \in \mathcal{X}_1: \Phi(\mathbf{q}(\xi)) > M} E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))] - \Phi(Q^\Lambda(0)) \leq -\frac{\eta}{4}.$$

Let

$$L^\Lambda(t^*) := \sup_{\xi \in \mathcal{X}_1} \Phi^{-(q-2)}(\xi) E_\xi \left[ (\Phi(Q^\Lambda(t^*/\Lambda)) - \Phi(\mathbf{q}(\xi)))^2 (\Phi(\mathbf{q}(\xi)) + |\Phi(Q^\Lambda(t^*/\Lambda)) - \Phi(\mathbf{q}(\xi))|)^{q-2} \right]. \quad (\text{A37})$$

Using again the fact that  $Q^\Lambda(t) \leq Q^\Lambda(0) + A^\Lambda(t)$  as well as some basic properties of the Poisson process we have that  $\limsup_{\Lambda \rightarrow \infty} L^\Lambda(t^*) \leq C_1$  for some constant  $C_1$ . Hence, we have by Lemma 8 in §E that there exists a constant  $C_2$  such that

$$\sup_{\xi \in \mathcal{X}_1: \Phi(\mathbf{q}(\xi)) > C_2} E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))^q] - \Phi(\mathbf{q}(\xi))^q \leq -\frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1}. \quad (\text{A38})$$

Using again the fact that  $Q^\Lambda(t) \leq Q^\Lambda(0) + A^\Lambda(t)$ , we have that

$$\sup_{\xi \in \mathcal{X}} \frac{E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))^q] - \Phi(\mathbf{q}(\xi))^q}{\Phi(\mathbf{q}(\xi))^q} \leq C_3, \quad (\text{A39})$$

for some  $C_3$ . In particular,

$$\sup_{\xi \in \mathcal{X}: \Phi(\mathbf{q}(\xi)) \leq C_2} E_\xi [\Phi(Q^\Lambda(t^*/\Lambda))^q] \leq C_4,$$

for some  $C_4 > 0$ . We now adapt arguments from the proof of Theorem 5.1 in Gamarnik and Zeevi [5], to establish bounds for  $(Q^\Lambda(\infty) - \eta_k^\Lambda)^+$ . Specifically, by the definition of stationarity we have that

$$E_{\pi^\Lambda} [\Phi(Q^\Lambda(0))^q] = E_{\pi^\Lambda} [\Phi(Q^\Lambda(t^*/\Lambda))^q]. \quad (\text{A40})$$

and, in particular,

$$0 = \int_{\xi \in \Xi^\Lambda} (\Phi(\mathbf{q}(\xi))^q - E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))^q]) \pi^\Lambda(d\xi). \quad (\text{A41})$$

Combining equations (A38) and (A39) we have that

$$E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))^q] - \Phi(\mathbf{q}(\xi))^q \leq -\frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} \mathbf{1}\{\xi \in \mathcal{X}_1\} + C_4 + C_3 \Phi(\mathbf{q}(\xi))^q \mathbf{1}\{\xi \notin \mathcal{X}_1\},$$

and, consequently, that for all  $\xi \in \mathcal{X}$ ,

$$\Phi(\mathbf{q}(\xi))^q - E_\xi[\Phi(Q^\Lambda(t^*/\Lambda))^q] \geq \frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} - C_4 - C_3 \Phi(\mathbf{q}(\xi))^q \mathbf{1}\{\xi \notin \mathcal{X}_1\} + \frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} \mathbf{1}\{\xi \notin \mathcal{X}_1\}.$$

Plugging back into (A41), we have that

$$\int_{\xi \in \Xi^\Lambda} \left( \frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} - C_4 - C_3 \Phi(\mathbf{q}(\xi))^q \mathbf{1}\{\xi \notin \mathcal{X}_1\} + \frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} \mathbf{1}\{\xi \notin \mathcal{X}_1\} \right) \pi^\Lambda(d\xi) \leq 0. \quad (\text{A42})$$

Now, using the bounds on the steady-state queue length from Proposition 5 and applying the Cauchy-Schwarz inequality yields that both

$$E_{\pi^\Lambda} [\Phi(Q^\Lambda(0))^{q-1} (\mathbf{1}\{\Xi^\Lambda(0) \notin \mathcal{X}_1\})] \rightarrow 0, \text{ as } \Lambda \rightarrow \infty, \quad (\text{A43})$$

and

$$E_{\pi^\Lambda} [\Phi(Q^\Lambda(0))^q (\mathbf{1}\{\Xi^\Lambda(0) \notin \mathcal{X}_1\})] \rightarrow 0, \text{ as } \Lambda \rightarrow \infty. \quad (\text{A44})$$

Consequently, (A42) implies that for all  $\Lambda$  large enough

$$\int_{\xi \in \Xi^\Lambda} \left( \frac{\eta q}{8} \Phi(\mathbf{q}(\xi))^{q-1} - 2C_4 \right) \pi^\Lambda(d\xi) \leq 0.$$

so that

$$E[\Phi(Q^\Lambda(\infty))^{q-1}] \leq \frac{16C_2}{\eta q}.$$

Note that with  $q = 2$  we have that  $E[(Q^\Lambda(\infty) - \eta_k^\Lambda)^+] \leq C_4$ , for some  $C_4 > 0$  and for all  $\Lambda$  large enough. Consequently,  $\limsup_{\Lambda \rightarrow \infty} E[(Q^\Lambda(\infty) - \eta_k^\Lambda)^+]/\sqrt{\Lambda} = 0$ , so that the first part of (A29) is established. Note that we have actually established that  $E[(Q^\Lambda(\infty) - \eta_k^\Lambda)^+] = O(1)$ , which is stronger than the statement of the Proposition which requires only that  $E[(Q^\Lambda(\infty) - \eta_k^\Lambda)^+] = o(\sqrt{\Lambda})$ . The statement about the steady-state waiting time now follows from Little's law.

We now turn to the proof of equation (A31). To analyze the behavior of the queue length process over the interval  $[0, T]$ , fix  $x > 0$  and re-define the set

$$\hat{\Omega}(\Lambda) = \left\{ \omega \in \Omega : \|Z_1^\Lambda(\cdot) - R\|_T \leq 2\epsilon\Lambda, \left\| \left( Z_{i,2}^\Lambda(\cdot) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right)^+ \right\|_T \leq 2\epsilon\Lambda, \text{ for all } i = 1, \dots, K \right\} \\ \cap \Omega^{**}(\Lambda, T, \delta\Lambda),$$

where

$$\Omega^{**}(\Lambda, T, \delta\Lambda) := \left\{ \omega \in \Omega : \sup_{0 \leq \eta \leq T} \sup_{t-s \leq \eta} (|B(m\Lambda t) - B(m\Lambda s)| - \delta\Lambda(t-s) + O(\log(m\Lambda t \vee 2))) \leq \frac{\epsilon}{4}\sqrt{\Lambda} \right\}. \quad (\text{A45})$$

One can easily show that  $P\{\Omega^{**}(\Lambda, T, \delta\Lambda)^c\} \leq C_5 e^{-C_6 \epsilon \sqrt{\Lambda}}$  for all  $\Lambda$  large enough and for some constants  $C_5$  and  $C_6$ . Indeed, this probability bound follows from a combination of equation (A18) and a basic result for Brownian motion (see Example 4.3.12 in page 264 of [7]). We now define the random times

$$\tau'^\Lambda = \sup \left\{ t \leq T : Q^\Lambda(t) \leq \eta_k^\Lambda + \epsilon\sqrt{\Lambda} \right\}, \text{ and } \tau''^\Lambda = \inf \left\{ t \geq \tau'^\Lambda : Q^\Lambda(t) \geq \eta_k^\Lambda + 2\epsilon\sqrt{\Lambda} \right\}.$$

By the same arguments preceding equation (A34), we have for all  $t \geq \tau'^\Lambda$  and by choosing  $\epsilon$  small enough that

$$Q^\Lambda(t \wedge \tau''^\Lambda) \leq Q^\Lambda(\tau'^\Lambda) - \frac{1}{2} \lambda_k q_k(0) \cdot (t \wedge \tau''^\Lambda - \tau'^\Lambda) + \delta\Lambda (t \wedge \tau''^\Lambda - \tau'^\Lambda) + \frac{\epsilon}{4} \sqrt{\Lambda}. \quad (\text{A46})$$

Choosing small enough  $\delta$ , we have by that on  $\hat{\Omega}(\Lambda)$ ,  $Q^\Lambda(\cdot)$  is smaller than  $Q^\Lambda(\tau'^\Lambda) + \epsilon/4\sqrt{\Lambda}$  for all

$t \geq \tau' \Lambda 4$ . Consequently,  $\tau^{\mu \Lambda} > T$  on  $\hat{\Omega}(\Lambda)$  so that  $\sup_{0 \leq t \leq T} Q^\Lambda(t) > \eta_{\bar{k}}^\Lambda + 2\epsilon\sqrt{\Lambda}$ . Hence,

$$\sup_{\xi \in \mathcal{X}_3} P_\xi \left\{ \sup_{0 \leq t \leq T} Q^\Lambda(t) > \eta_{\bar{k}}^\Lambda + 2\epsilon\sqrt{\Lambda} \right\} \leq P\{\hat{\Omega}(\Lambda)^c\} \leq C_7 e^{-C_8 \sqrt{\Lambda}},$$

for all  $\Lambda$  large enough where the last inequality follows from Propositions 5 and 6 as well as the probability bound for  $\Omega^{**}(\Lambda, T, \delta\Lambda)$ . This establishes equation (A31). The proof of (A32) uses the following representation for the virtual waiting time:

$$W^\Lambda(t) = \inf \{s \geq 0 : D^\Lambda(t+s) - D^\Lambda(t) \geq Q^\Lambda(t)\},$$

where  $D^\Lambda(t)$  is the aggregate number of depletions of customers from the queue up to time  $t$ . Let  $\tau^\Lambda(t) = \inf\{s \geq 0 : Q^\Lambda(t+s) = 0\}$ . Since while the queue is non-empty every cross-selling completion is followed by an admission of a customer from the queue, we have that on  $[t, \tau^\Lambda(t)]$ ,

$$D^\Lambda(t+s) - D^\Lambda(t) \geq \sum_{i=1}^{\bar{k}} D_{i,2}^\Lambda(t+s) - D_{i,2}^\Lambda(t) \geq \sum_{i=1}^{\bar{k}} \mu_i^{cs} \int_t^{t+s} Z_{i,2}^\Lambda(u) du - |B(m\Lambda(t+s)) - B(m\Lambda t)|.$$

Then, on the set  $\hat{\Omega}(\Lambda)$  and for  $s \leq \tau^\Lambda(t)$ ,

$$D^\Lambda(t+s) - D^\Lambda(t) \geq \sum_{i=1}^{\bar{k}} \mu_i^{cs} \int_t^{t+s} Z_{i,2}^\Lambda(u) du - \delta\Lambda s \geq s \sum_{i=1}^{\bar{k}} \mu_i^{cs} \frac{\lambda_i q_i(0)}{\mu_i^{cs}} - C_9 \epsilon \Lambda s - \delta\Lambda s - \frac{\epsilon}{4} \sqrt{\Lambda},$$

for some constant  $C_9 > 0$ . Consequently, for all  $t \geq 0$ ,

$$W^\Lambda(t) \leq \frac{C_w}{\sqrt{\Lambda}} + \frac{\|Q^\Lambda(\cdot)\|_T}{\sum_{i=1}^{\bar{k}} \lambda_i q_i(0) - C\epsilon\Lambda - \delta\Lambda},$$

for some constant  $C_w$ . The bound for the waiting time in (A32) now follows by choosing  $\delta$  and  $\epsilon$  small enough and using the bound in (A31). Note that the virtual waiting time actually depends on the behavior of the departures slightly after time  $T$ . This problem is easily overcome, however, by re-defining  $\hat{\Omega}(\Lambda)$  using the interval  $[0, 2T]$  instead of  $[0, T]$ .  $\blacksquare$

In order to analyze the number of cross-sold customers from each class a finer analysis is required. In particular, we need a handle of the waiting time of the customers that are present in the system at time 0 (which is assumed to be distributed according to the stationary distribution). For the following results, we let  $\tilde{Z}^\Lambda(t)$  be the number of customers in the first phase of service or in queue

that found upon arrival a virtual waiting time that is longer than  $2M_2/\sqrt{\Lambda}$ , where  $M_2$  is as given in Proposition 7. Formally,

$$\tilde{Z}^\Lambda(t) = \sum_{j \in \mathcal{S}(t)} \mathbf{1}\{w_j(t) > 2M_2\sqrt{\Lambda}\}. \quad (\text{A47})$$

We then have the following Lemma where we use

$$\mathcal{X}_4 := \{\xi \in \mathcal{X}_3 : \tilde{z}(\xi) \leq \epsilon\Lambda\}, \quad (\text{A48})$$

and  $\mathcal{X}_3$  is as defined in (A29).

**Lemma 7** *Fix an integer  $q \geq 2$  and  $\epsilon > 0$ . Then,*

$$P \left\{ \tilde{Z}^\Lambda(\infty) > 2\epsilon\Lambda \right\} \leq c_{13} \frac{1}{\sqrt{\Lambda}^{(q-1)}}. \quad (\text{A49})$$

Moreover,

$$\sup_{\xi \in \mathcal{X}_3} P_\xi \left\{ \|\tilde{Z}^\Lambda(\cdot)\|_T > 2\epsilon\Lambda \right\} \leq c_{13} e^{-c_{14}\epsilon\sqrt{\Lambda}}, \quad (\text{A50})$$

for all  $\Lambda$  large enough and for some constant  $c_{13}$  and  $c_{14}$ .

**Proof:** We define  $\tilde{Q}^\Lambda(t)$  to be the number of customers in queue that found a virtual waiting time longer than  $2\eta_k^\Lambda/\Lambda$  upon arrival. Then, focusing on the number of customers with waiting times longer than  $2\eta_k^\Lambda/\Lambda$  in queue or in service we have the equation,

$$\begin{aligned} \tilde{Q}^\Lambda(t) + \tilde{Z}^\Lambda(t) &= \tilde{Q}^\Lambda(0) + \tilde{Z}^\Lambda(0) + \int_0^t \mathbf{1}\{W^\Lambda(t-) > 2M_2\sqrt{\Lambda}\} dA^\Lambda(t) \\ &\quad - \sum_{l=1}^{D_1^\Lambda(t)} \sum_{j \in \mathcal{S}(t-)} \mathbf{1}\{e_s(\mathcal{S}(t-), l) = j, w_j(t-) > 2M_2\sqrt{\Lambda}\}. \end{aligned} \quad (\text{A51})$$

Note that  $\sum_{l=1}^{D_1^\Lambda(t)} \sum_{j \in \mathcal{S}(t-)} \mathbf{1}\{e_s(\mathcal{S}(t-), l) = j, 2M_2\sqrt{\Lambda}\}$  is a non-homogeneous Poisson process with instantaneous rate equal to  $\mu^s \tilde{Z}^\Lambda(u)$ .

Consider the differential equation (initialized at  $\tilde{Z}^\Lambda(0)$ ),

$$\bar{\bar{Z}}^\Lambda(t) = \bar{\bar{Z}}^\Lambda(0) - \mu^s \int_0^t \bar{\bar{Z}}^\Lambda(u) du. \quad (\text{A52})$$

Since  $\tilde{Z}^\Lambda(0) \leq N^\Lambda$ , there exists a finite time  $t^*$  after which  $\tilde{Z}^\Lambda(t) \leq \epsilon\Lambda$ . Fix the set

$$\hat{\Omega}(\Lambda) := \Omega^*(\Lambda, T, \epsilon\Lambda) \cap \left\{ \omega \in \Omega : \|W^\Lambda(\cdot)\|_T \leq \frac{M_2 + \epsilon}{\sqrt{\Lambda}} \right\} \cap \left\{ \omega \in \Omega : \|Q^\Lambda(\cdot)\|_T \leq \eta_{\bar{k}}^\Lambda + 2\epsilon\sqrt{\Lambda} \right\}.$$

Subtracting  $\tilde{Z}^\Lambda(t)$  from  $\bar{\tilde{Z}}^\Lambda(t)$ , using the fact that  $\int_0^t \mathbf{1}\{W^\Lambda(t-) > 2\eta_{\bar{k}}^\Lambda/\Lambda\} dA^\Lambda(t) = 0$  on  $\hat{\Omega}(\Lambda)$  and finally applying Gronwall's inequality (see, e.g., Problem 5.2.7 on page 287 of Karatzas and Shreve [7]) we have that

$$\|\tilde{Z}^\Lambda(\cdot) - \bar{\tilde{Z}}^\Lambda(\cdot)\|_T \leq Ce^{CT} (\|B(m\Lambda \cdot)\|_T + \|Q^\Lambda(\cdot)\|_T), \quad (\text{A53})$$

on  $\hat{\Omega}(\Lambda)$ . We then have that

$$\begin{aligned} P_{\pi^\Lambda} \{ \tilde{Z}^\Lambda(t^*) > 2\epsilon\Lambda \} &\leq P_{\pi^\Lambda} \{ \Xi^\Lambda(0) \notin \mathcal{X}_3 \} + \sup_{\xi \in \mathcal{X}_3} P_\xi \{ C'^\Lambda > \epsilon\Lambda, \mathbf{1}\{\hat{\Omega}(\Lambda)\} \\ &\quad + \sup_{\xi \in \mathcal{X}_3} P_\xi \{ \mathbf{1}\{\hat{\Omega}(\Lambda)^c\} \} \leq C_{13} e^{-C_{14}\epsilon\sqrt{\Lambda}}, \end{aligned} \quad (\text{A54})$$

where  $C'^\Lambda$  is the right hand side in equation (A53) and the last inequality follows from the definition of  $\hat{\Omega}(\Lambda)$ , Proposition 7 and the properties of Brownian Motion enlisted in §C.2. Since under the steady state distribution  $\tilde{Z}^\Lambda(t^*)$  has the same distribution as  $\tilde{Z}^\Lambda(\infty)$ , we can use Proposition 7 and the bounds from §C.2 to establish equation (A49). The transient bound in equation (A50) follows a similar argument in which one replaces in equation (A54) the initial distribution  $\pi^\Lambda$  with an arbitrary initial condition within the set  $\mathcal{X}_4$ . ■

We now turn to the analysis of the number of agents busy cross-selling to type- $i$  customers. We first focus on types  $i \leq \bar{k} - 1$ . Type  $\bar{k}$  requires a separate analysis which is given in Proposition 9. We define

$$\mathcal{X}_5 := \left\{ \xi \in \mathcal{X}_4 : \left| \tilde{z}_{i,2}(\xi) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right| \leq \epsilon\Lambda, \quad i = 1, \dots, \bar{k} - 1 \right\} \quad (\text{A55})$$

and  $\mathcal{X}_4$  is as defined in (A48).

**Proposition 8** *Fix an integer  $q \geq 2$  and  $\epsilon > 0$ . Then,*

$$P \left\{ \left| Z_{i,2}^\Lambda(\infty) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right| \geq \epsilon\Lambda \right\} \leq c_{15} \frac{1}{(\sqrt{\Lambda})^{(q-1)}}, \quad i = 1, \dots, \bar{k} - 1, \quad (\text{A56})$$

and

$$E \left[ \left| Z_{i,2}^\Lambda(\infty) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right| \right] \leq \epsilon\Lambda, \quad (\text{A57})$$

for all  $\Lambda$  large enough and for some constant  $c_{15}$ . Moreover,

$$\sup_{\xi \in \mathcal{X}_5^s} P_\xi \left\{ \left\| Z_{i,2}^\Lambda(\cdot) - \frac{\lambda_i q_i(0)}{\mu_i^{cs}} \right\|_T \geq 2\epsilon\Lambda \right\} \leq c_{16} e^{-c_{17}\epsilon\sqrt{\Lambda}}, \quad i = 1, \dots, \bar{k} - 1, \quad (\text{A58})$$

for all  $\Lambda$  large enough and for some constant  $c_{16}$  and  $c_{17}$ .

**Remark 4 (Proof of Theorem 1 in [3])** Since  $\epsilon$  is arbitrary, Proposition 8 implies that for  $i \leq \bar{k} - 1$ ,  $E[Z_{i,2}^\Lambda(\infty)] = \lambda_i q_i / \mu_i^{cs} + o(\Lambda)$ . In particular, since  $E[Z_1^\Lambda(\infty)] = R + o(\Lambda)$  and  $N = R + \sum_{i=1}^{\bar{k}} \lambda_i q_i \mu_i^{cs}$ , it suffices to show that  $E[I^\Lambda(\infty)] = o(\Lambda)$ , where  $I^\Lambda(t)$  is the number of idle agents at time  $t$ , to conclude that also for  $\bar{k}$ :

$$E[Z_{\bar{k},2}^\Lambda(\infty)] = \frac{\lambda_{\bar{k}} q_{\bar{k}}}{\mu_{\bar{k}}^{cs}} + o(\Lambda).$$

The required result for  $I^\Lambda(\infty)$  is established in Proposition 9, which together with Proposition 7 establish (22) and complete the **proof of Theorem 1** in the main paper [3].

### Proposition 9

$$E[I^\Lambda(\infty)] = o(\Lambda), \quad \text{and} \quad E[Z_{i,2}^\Lambda(\infty)] = \frac{\lambda_i q_i}{\mu_i^{cs}} + o(\Lambda), \quad \forall i \leq \bar{k}. \quad (\text{A59})$$

The proof of Proposition 9 is postponed until after the proof of Proposition 8.

**Proof of Proposition 8:** The argument is very similar in nature to the one used in the proof of Lemma 7. Define first the set

$$\begin{aligned} \hat{\Omega}(\Lambda) := & \Omega^*(\Lambda, T, x\sqrt{\Lambda}) \cap \left\{ \omega \in \Omega : \|W^\Lambda(\cdot)\| \leq \frac{M_2 + 2\epsilon\sqrt{\Lambda}}{\sqrt{\Lambda}} \right\} \cap \left\{ w \in \Omega : \|\tilde{Z}^\Lambda(\cdot)\|_T \leq (\epsilon + x)\Lambda \right\} \\ & \cap \left\{ w \in \Omega : \|Z_1^\Lambda(\cdot) - R\|_T \leq 2\epsilon\Lambda \right\} \cap \left\{ w \in \Omega : \|Q^\Lambda(\cdot)\|_T \leq \eta_{\bar{k}} + \epsilon\sqrt{\Lambda} \right\}, \end{aligned} \quad (\text{A60})$$

where we choose  $M$  so that  $M < (\hat{\eta}_{\bar{k}-1} - \hat{\eta}_{\bar{k}})/2$ . As  $Z_{i,2}^\Lambda(t) = Z_{i,2}^\Lambda(0) - D_{i,2}^\Lambda(t) + \tilde{D}_i^\Lambda(t)$ , we have on  $\hat{\Omega}(\Lambda)$  that

$$Z_{i,2}^\Lambda(t) \geq Z_{i,2}^\Lambda(0) - \mu_i^{cs} \int_0^t Z_{i,2}^\Lambda(u) du + \mu^s \frac{\lambda_i}{\Lambda} q_i \left( 2M_2 / \sqrt{\Lambda} \right) \int_0^t Z^\Lambda(u) - \tilde{Z}^\Lambda(u) du - \|B(m\Lambda \cdot)\|_T + O(\log(m\Lambda T \vee 2)). \quad (\text{A61})$$

Here we used the definition of  $\tilde{Z}^\Lambda$  as well as the fact that while  $Q^\Lambda(t) \leq \eta_{\bar{k}-1}$  all service completions

with type- $i$  customers are followed by a cross-selling offer if the customer agrees to listen. Also,

$$Z_{i,2}^\Lambda(t) \leq Z_{i,2}^\Lambda(0) - \mu_i^{cs} \int_0^t Z_{i,2}^\Lambda(u) du + \mu^s \frac{\lambda_i}{\Lambda} q_i(0) \int_0^t Z_1^\Lambda(u) - \bar{Z}^\Lambda(u) du + \|B(m\Lambda \cdot)\|_T + O(\log(m\Lambda T \vee 2)). \quad (\text{A62})$$

Consider now the differential equation (initialized at  $Z_{i,2}^\Lambda(0)$ ):

$$\bar{Z}_{i,2}^\Lambda(t) = Z_{i,2}^\Lambda(0) + \mu^s \frac{\lambda_i}{\Lambda} q_i(0) R t - \mu_i^{cs} \int_0^t \bar{Z}_{i,2}^\Lambda(u) du. \quad (\text{A63})$$

Then, subtracting  $\bar{Z}_{i,2}^\Lambda(t)$  from  $Z_{i,2}^\Lambda(t)$ , using the inequalities (A61) and (A62) as well as equation (A63) and applying Gronwall's inequality we have that

$$\begin{aligned} \|Z_{i,2}^\Lambda(\cdot) - \bar{Z}_{i,2}^\Lambda(\cdot)\|_T &\leq C e^{CT} [q_i(\|W^\Lambda(\cdot)\|_T) \|Z_1^\Lambda(\cdot) - R\|_T \vee \|\bar{Z}^\Lambda(\cdot)\|_T \\ &\vee (\|B(m\Lambda \cdot)\|_T + O(\log(m\Lambda T \vee 2)))] \end{aligned} \quad (\text{A64})$$

The argument now follows as in the proof of Lemma 7. Re-define the function  $\Phi(x) = |x - \lambda_i q_i(0)/\mu_i^{cs}|$ . We first consider the differential equation (A63). Observe that since  $Z_{i,2}^\Lambda(0) \leq N^\Lambda$ , there exists  $t^*$  so that

$$\Phi(\bar{Z}_{i,2}^\Lambda(t^*)) \leq \epsilon \Lambda.$$

Hence,

$$\begin{aligned} P_{\pi^\Lambda} \{\Phi(Z_{i,2}^\Lambda(t^*)) > 2\epsilon \Lambda\} &\leq P_{\pi^\Lambda} \{\Xi^\Lambda(0) \notin \mathcal{X}_4\} + \sup_{\xi \in \mathcal{X}_4} P_\xi \{C''^\Lambda > 2\epsilon \Lambda, \hat{\Omega}(\Lambda)\} \\ &+ \sup_{\xi \in \mathcal{X}_4} P_\xi \{\hat{\Omega}(\Lambda)^c\} \leq \frac{C}{(\sqrt{\Lambda})^{\frac{q-1}{2}}}, \end{aligned} \quad (\text{A65})$$

where  $C''^\Lambda$  is the right hand side in (A64) and  $\mathcal{X}_4$  is as defined in (A48). Since  $\Phi(Z_{i,2}^\Lambda(t^*))$  has the same distribution as  $\Phi(Z_{i,2}^\Lambda(\infty))$  when starting with the steady distribution we have established equation (A56). Equation (A57) follows a similar argument noting that

$$E_{\pi^\Lambda}[\Phi(Z_{i,2}^\Lambda(t^*))] \leq N^\Lambda P_{\pi^\Lambda} \{\Xi^\Lambda(0) \notin \mathcal{X}_3\} + \sup_{x \in \mathcal{X}_3} E_\xi[\Phi(Z_{i,2}^\Lambda(t^*))]$$

and applying the bounds we have from the previous propositions to show that

$$N^\Lambda P_{\pi^\Lambda} \{\Xi^\Lambda(0) \notin \mathcal{X}_4\} \rightarrow 0, \text{ as } \Lambda \rightarrow \infty$$

as well as

$$\sup_{\xi \in \mathcal{X}_4} E_\xi[\Phi(Z_{i,2}^\Lambda(t^*))] \leq N^\Lambda \sup_{\xi \in \mathcal{X}_4} P_\xi\{\hat{\Omega}^c\} + 2\epsilon\Lambda.$$

Replacing  $\epsilon$  with  $\epsilon/2$  throughout one has the result. The transient bound is easily obtained using similar arguments.  $\blacksquare$

**Proof of Proposition 9:** Re-define the set

$$\begin{aligned} \hat{\Omega}(\Lambda) &= \Omega^{**}(\Lambda, T, \delta\Lambda) \cap \left\{ \omega \in \Omega : \|Z_1^\Lambda(\cdot) - R\|_T \leq 2\epsilon\Lambda \right. \\ &\quad \left\| Z_{i,2}^\Lambda(\cdot) - \frac{\lambda_i q_i}{\mu_i^{cs}} \right\|_T \leq 2\epsilon\Lambda, \forall i = 1, \dots, \bar{k} - 1; \\ &\quad \left. \|\tilde{Z}^\Lambda(\cdot)\|_T \leq 2\epsilon\Lambda; \|W^\Lambda(\cdot)\|_T \leq (M_2 + \epsilon)/\sqrt{\Lambda} \right\}, \end{aligned} \quad (\text{A66})$$

where  $\Omega^{**}$  is defined in (A45). Assume that  $I^\Lambda(0) \geq 2\epsilon\Lambda$  and set  $\tau^\Lambda = \inf\{t \geq 0 : I^\Lambda(t) \leq I^\Lambda(0) - \epsilon\Lambda\}$ . Then, on  $\hat{\Omega}(\Lambda)$ ,

$$\begin{aligned} I^\Lambda(t \wedge \tau^\Lambda) &\leq I^\Lambda(0) - \Lambda(t \wedge \tau^\Lambda) + \sum_{i=1}^{\bar{k}-1} \mu_i^{cs} \int_0^{t \wedge \tau^\Lambda} Z_{i,2}^\Lambda(u) du \\ &\quad + \mu^s \left(1 - q \left(2M_2/\sqrt{\Lambda}\right)\right) \int_0^{t \wedge \tau^\Lambda} Z^\Lambda(u) - \tilde{Z}^\Lambda(u) du \\ &\quad + \mu_i^{cs} \int_0^{t \wedge \tau^\Lambda} \left( N - I^\Lambda(u) - Z^\Lambda(u) - \sum_{i=1}^{\bar{k}-1} Z_{i,2}^\Lambda(u) \right) du + \delta\Lambda t + \epsilon\sqrt{\Lambda}. \end{aligned} \quad (\text{A67})$$

Some algebra yields that

$$I^\Lambda(t \wedge \tau^\Lambda) \leq I^\Lambda(0) + \epsilon\sqrt{\Lambda} - C\Lambda t, \quad (\text{A68})$$

for some constant  $C > 0$ . Starting at  $I^\Lambda(0) > 2\epsilon\Lambda$ , then, there exists  $t^*$  at which  $I^\Lambda(t^*) \leq \epsilon\Lambda$ . If, on the other hand,  $I^\Lambda(0) \leq 2\epsilon\Lambda$ , then we claim that  $I^\Lambda(t) \leq 3\epsilon\Lambda$  for all  $t \leq T$ . Indeed let  $\tau'^\Lambda := \inf\{t \geq t^* : I^\Lambda(t) \geq \epsilon\Lambda\}$  and  $\tau''^\Lambda := \inf\{t \geq \tau'^\Lambda : I^\Lambda(t) \geq 2\epsilon\Lambda\}$ . Then, using (A68) we have that that on  $\hat{\Omega}(\Lambda)$ ,  $\tau''^\Lambda > T$ . Consequently, on  $\hat{\Omega}(\Lambda)$ , there exists  $t^*$  with  $I^\Lambda(t^*) \leq 3\epsilon\Lambda$  regardless of the initial condition. We then have that

$$E_{\pi^\Lambda}[I^\Lambda(t^*)] \leq N^\Lambda P\{\Xi^\Lambda(0) \notin \mathcal{X}_4\} + 3\epsilon\Lambda + \sup_{\xi \in \mathcal{X}_5} P\{\hat{\Omega}(\Lambda)^c\},$$

and the proof is completed by noting that

$$N^\Lambda P\{\Xi^\Lambda(0) \notin \mathcal{X}_5\} + \sup_{\xi \in \mathcal{X}_5} P\{\hat{\Omega}(\Lambda)^c\} \rightarrow 0, \text{ as } \Lambda \rightarrow \infty.$$

■

As explained in Remark 4, with Proposition 9 we conclude the proof of Theorem 1 in the main paper [3]. We conclude this appendix with Lemma 8 that was used in the proof of Proposition 7.

## E Auxiliary Results

In this section we prove Proposition 5 as well as one simple general result that we used in the proof of Proposition 7. We begin with the latter.

The following Lemma is an adaptation of a result that and appears in [6] and is due to Gamarnik and Zeevi. We state and prove it here for completeness. Fix  $\Lambda$  and consider the process  $\Xi(t)$  with the domain  $\mathcal{X}$  and Let  $\Phi(x)$  be a function  $\Phi(x) : \mathcal{X} \mapsto \mathbb{R}_+$ . We fix a subset  $\tilde{\mathcal{X}} \subset \mathcal{X}$ . We let

$$L(t) = \sup_{\xi \in \tilde{\mathcal{X}}} \Phi^{-(q-2)}(\xi) E_\xi [\Phi(\Xi(t)) - \Phi(\xi)]^2 (\Phi(\xi) + |\Phi(\Xi(t)) - \Phi(\xi)|)^{q-2}. \quad (\text{A69})$$

**Lemma 8** *Fix an integer  $q \geq 2$ . Assume that there exists  $K > 0$ ,  $\gamma > 0$  and  $t^* > 0$ , so that*

$$\sup_{\xi \in \tilde{\mathcal{X}}: \Phi(\xi) > K} \{E_\xi[\Phi(\Xi(t^*))] - \Phi(\xi)\} \leq -\gamma, \quad (\text{A70})$$

and that  $L(t^*)$  is finite. Then,

$$\sup_{\xi \in \tilde{\mathcal{X}}: \Phi(\xi) > K'} \{E_\xi[\Phi(\Xi(t^*))^q] - \Phi(\xi)^q\} \leq -\frac{\gamma q}{2} \Phi(\xi)^{q-1}, \quad (\text{A71})$$

with  $K' = \max\{K, L(t^*)(q-1)/\gamma\}$ .

**Proof:** Using second order Taylor's expansion of the function  $x^p$  around  $\Phi(\xi)$  we obtain for every

$\xi \in \tilde{\mathcal{X}}$  such that  $\Phi(\xi) > k$ ,

$$\begin{aligned}
E_\xi[\Phi^q(\Xi(t^*))] - \Phi^q(\xi) &= q\Phi^{q-1}E_\xi E_\xi[\Phi(\Xi(t^*)) - \Phi(\xi)] \\
&+ \frac{q(q-1)}{2}E_\xi[(\Phi(\xi) + Z(\Phi(\Xi(t^*)) - \Phi(\xi)))^{q-2}(\Phi(\Xi(t^*)) - \Phi(\xi))^2] \\
&\leq -\gamma q\Phi^{q-1}(\xi) + \frac{q(q-1)}{2}E_\xi[(\Phi(\xi) + |\Phi(\Xi(t^*)) - \Phi(\xi)|)^{q-2}(\Phi(\Xi(t^*)) - \Phi(\xi))^2] \\
&\leq -\gamma q\Phi^{q-1}(\xi) + \frac{q(q-1)}{2}L(t^*)\Phi^{q-2}(\xi) \tag{A72}
\end{aligned}$$

where  $Z$  is a random variable with support in  $[0, 1]$ . When  $\Phi(\xi) > L(t^*)(q-1)/\gamma$ , we obtain that

$$E_\xi[\Phi^q(\Xi(t^*))] - \Phi^q(\xi) \leq -\frac{\gamma q}{2}\Phi^{q-1}(\xi).$$

■

**Proof of Proposition 5:** We now prove the estimates given in Proposition 5 for the number of agents busy giving service and the queue length. Some of the steps are very similar to those in [1]. The similar parts will be abbreviated and the reader will be referred to the technical appendix of [1] for the details. The first step is the following Lemma, the first part of which is an analogue of Lemma B.1 in [1].

**Lemma 9** *Fix  $\epsilon > 0$ . Then, there exists  $t^0(\epsilon)$  (independent of the initial conditions), such that*

$$P \left\{ \sup_{t^0(\epsilon) \leq t \leq T} (Z_1^\Lambda(t) - R)^- > \epsilon\Lambda \right\} \leq \tilde{c}_9^{-\tilde{c}_{10}\sqrt{\Lambda}}, \tag{A73}$$

for all  $\Lambda$  large enough and for two positive constants  $\tilde{c}_9$  and  $\tilde{c}_{10}$ . Consequently,

$$P \{ (Z_1^\Lambda(\infty) - R)^- > \epsilon\Lambda \} \leq c_{18}e^{-c_{19}\sqrt{\Lambda}},$$

and

$$\sup_{\xi \in \mathcal{X}_1} P_\xi \{ \| (Z_1^\Lambda(\cdot) - R)^- \|_T \geq 2\epsilon\Lambda \} \leq c_9 e^{-c_{10}\sqrt{\Lambda}}. \tag{A74}$$

for all  $\Lambda$  large enough and for some strictly positive constants  $c_9$  and  $c_{10}$ .

**Proof:** The first part of the Lemma is proved just as in [1] and its proof is omitted. Since the bound is independent of the initial state, we can in particular initialize the system with its steady

state distribution in which case we get the bound on this steady state distribution.  $\blacksquare$

The next step establishes a crude bound for the queue-length process. It shows that the queue length is essentially of order  $\Lambda$ . This step is required to obtain later the more refined bound.

**Lemma 10** *Fix  $\epsilon > 0$ . Then,*

$$\limsup_{\Lambda \rightarrow \infty} E \left[ \exp \left( \frac{Q^\Lambda(\infty)}{\Lambda} \right) \right] \leq C_Q$$

for some constant  $C_Q$ .

**Proof:** We use a Lyapunov type of argument along the lines of Gamarnik and Zeevi [5]. First, assume that  $Q^\Lambda(0) > 2M\Lambda$  for some constant  $M > 0$  and define the random time

$$\tau^\Lambda = \inf\{t \geq 0 : Q^\Lambda(t) \leq Q^\Lambda(0) - M\Lambda\}.$$

Since the largest threshold,  $\eta_1^\Lambda$  satisfies  $\eta_1^\Lambda = \hat{\eta}_1 \sqrt{\Lambda}$ , there exists  $\Lambda$  large enough such that  $M\Lambda > \eta_1^\Lambda$ . Consequently, on  $[0, \tau^\Lambda]$ , all service or cross-selling completions will be followed by an admission of a customer from the head of the queue. The queue length process satisfies then the equation

$$Q^\Lambda(t \wedge \tau^\Lambda) = Q^\Lambda(0) + A^\Lambda(t) - D_1^\Lambda(t) - \sum_{i=1}^{\bar{k}} D_{i,2}^\Lambda(t). \quad (\text{A75})$$

Using the strong approximation we have that

$$Q^\Lambda(t \wedge \tau^\Lambda) \leq Q^\Lambda(0) + \Lambda t - \mu^s \int_0^t Z_1^\Lambda(s) ds - \sum_{i=1}^{\bar{k}} \mu_i^{cs} \int_0^t Z_{i,2}^\Lambda(s) ds + \|B(m\Lambda \cdot)\|_T + O(\log(m\Lambda T \vee 2)). \quad (\text{A76})$$

Let

$$\hat{\Omega}(\Lambda) = \Omega^*(\Lambda, T, \epsilon\Lambda) \cap \{\omega \in \Omega : \|Z_1^\Lambda(\cdot) - R\|_T \leq \epsilon\Lambda\},$$

with  $\Omega^*(\Lambda, T, \epsilon\Lambda)$  as defined in (A21). We then have after some algebraic manipulation that on  $\hat{\Omega}(\Lambda)$ ,

$$Q^\Lambda(t \wedge \tau^\Lambda) \leq Q^\Lambda(0) + \epsilon\Lambda - C\Lambda(t \wedge \tau^\Lambda). \quad (\text{A77})$$

We claim now that with  $t^* = 3M/(2C)$  we have that

$$E_{\xi \in \mathcal{X}: q(\xi) > \epsilon \Lambda} \frac{E_{\xi} [\exp(Q^{\Lambda}(t^*)/\Lambda)]}{\exp(q(\xi)/\Lambda)} \leq \exp(-M/4). \quad (\text{A78})$$

We omit the simple argument. We now use equation (A78) to establish a bounds for the steady state queue length using a result from [5]. Towards that end, note first that since  $Q^{\Lambda}(t) \leq Q^{\Lambda}(t) + A^{\Lambda}(t)$ , we have that

$$\sup_{\xi \in \mathcal{X}} \frac{E_{\xi} \left[ e^{Q^{\Lambda}(t^*)/\Lambda} \right]}{e^{q(\xi)/\Lambda}} \leq C_1, \quad (\text{A79})$$

for some constant  $C_1 > 0$  and for all  $\Lambda$ . We can now apply Theorem 5 of [5] to obtain that

$$E \left[ \exp \left( \frac{Q^{\Lambda}(\infty)}{\Lambda} \right) \right] \leq C_Q, \quad (\text{A80})$$

for some constant  $C_Q$  and all  $\Lambda$  large enough. ■

In the next Lemma we show that the queue hardly exceeds the greatest threshold,  $\eta_1^{\Lambda}$ , as stated in Proposition 5.

**Lemma 11**  $\epsilon > 0$  and  $q \geq 2$ . Then,

$$\limsup_{\Lambda \rightarrow \infty} E \left[ \left( (Q^{\Lambda}(\infty) - \eta_1^{\Lambda})^+ \right)^{q-1} \right] \leq C_Q \quad (\text{A81})$$

for some strictly positive constant  $C_Q$ . Moreover,

$$\sup_{\xi \in \mathcal{X}_1} P_{\xi} \left\{ \left\| (Q^{\Lambda}(\cdot) - \eta_1^{\Lambda})^+ \right\|_T \geq 2\epsilon\sqrt{\Lambda} \right\} \leq c_9 e^{-c_{10}\epsilon\sqrt{\Lambda}} \quad (\text{A82})$$

for all  $\Lambda$  large enough and for some strictly positive constants  $c_9$  and  $c_{10}$ .

**Proof:** The proof follows almost exactly the proof of the Lemma 10 up to equation (A77). The main difference is that here, like in the proof of Proposition 7, one replaces the time interval  $[0, T]$  with the time interval  $[0, T/\Lambda]$ . Once the dynamics of the system above the level of  $\eta_1^{\Lambda}$  are established, the proof follows almost exactly as the proof of Proposition 7 with two exceptions: First, we replace  $\mathcal{X}_1$  there with

$$\mathcal{X}'_1 := \{ \xi \in \mathcal{X} : (z_1(\xi) - R)^- \leq \epsilon \Lambda \}.$$

Then, the modification of equations (A43) and (A44) is established here using Lemma 10 (rather than 5). Specifically, by the Cauchy-Schwarz inequality that

$$E_{\pi^\Lambda} [\Phi(Q^\Lambda(0))^{q-1} (\mathbf{1}\{\Xi^\Lambda(0) \notin \mathcal{X}'_1\})] \leq \sqrt{E_{\pi^\Lambda}[(\Phi(Q^\Lambda(\infty)))^2(q-1)]} \sqrt{P_{\pi^\Lambda}\{\Xi^\Lambda(0) \notin \mathcal{X}'_1\}}. \quad (\text{A83})$$

By Lemma 10,

$$\limsup_{\Lambda \rightarrow \infty} E[(\Phi(Q^\Lambda(\infty)))^{2(q-1)}] \leq C\Lambda^{q-1},$$

for some constant  $C > 0$ . By Lemma 9

$$P_{\pi^\Lambda}\{\Xi^\Lambda(0) \notin \mathcal{X}'_1\} \leq c_{18}e^{-c_{19}\sqrt{\Lambda}}.$$

Plugging back into (A83) we have that

$$E_{\pi^\Lambda} [\Phi(Q^\Lambda(0))^{q-1} (\mathbf{1}\{\Xi^\Lambda(0) \notin \mathcal{X}'_1\})] \rightarrow 0, \text{ as } \Lambda \rightarrow \infty.$$

The modification of equation (A44) follows similarly. The remainder of the proof follows almost exactly the remainder of proof of Proposition 7 with the obvious modifications required by the replacement of the smallest threshold  $\eta_k^\Lambda$  with the largest threshold  $\eta_1^\Lambda$ .  $\blacksquare$

The last component required to complete the proof of Proposition 5 is to establish a two-sided bound for the difference  $Z_1^\Lambda(\cdot) - R$ , rather than the one sided bound from Lemma 9. The result is given in the following Lemma.

**Lemma 12** *Fix  $\epsilon > 0$  and  $q \geq 2$ . Then,*

$$P\{|Z_1^\Lambda(\infty) - R| > \epsilon\Lambda\} \leq \frac{c_7}{(\sqrt{\Lambda})^{q-1}}, \quad (\text{A84})$$

and

$$\sup_{\xi \in \mathcal{X}'_1} P_\xi \{\|Z_1^\Lambda(\cdot) - R\|_T \geq 2\epsilon\Lambda\} \leq c_9e^{-c_{10}\sqrt{\Lambda}}, \quad (\text{A85})$$

for all  $\Lambda$  large enough and for some strictly positive constants  $c_9$  and  $c_{10}$ .

**Proof:** The proof is very similar in nature to the proof of Proposition 8 and is actually simpler. First, note that  $Q^\Lambda(t)$  and  $Z_1^\Lambda(t)$  satisfy the equation

$$Q^\Lambda(t) + Z^\Lambda(t) = Q^\Lambda(0) + Z^\Lambda(0) + A^\Lambda(t) - D_1^\Lambda(t),$$

or using the strong approximation decomposition,

$$Q^\Lambda(t) + Z^\Lambda(t) = Q^\Lambda(0) + Z^\Lambda(0) - \Lambda t - \mu^s \int_0^t Z_1^\Lambda(s) ds + B \left( \Lambda t + \mu^s \int_0^t Z_1^\Lambda(s) ds \right) + O(\log(m\lambda t \vee 2)).$$

Consider the differential equation (initialized at  $Z_1^\Lambda(0)$ )

$$\bar{Z}_1^\Lambda(0) = Z_1^\Lambda(0) + \Lambda t - \mu^s \int_0^t \bar{Z}_1^\Lambda(s) ds.$$

Then, subtracting  $\bar{Z}_1^\Lambda(t) - Z_1^\Lambda(t)$  and using Gronwall's inequality we have that

$$\|\bar{Z}_1^\Lambda(\cdot) - Z_1^\Lambda(\cdot)\|_T \leq C e^{CT} [\|Q^\Lambda(\cdot)\|_T + \|B(m\Lambda \cdot)\|_T + O(\log(m\Lambda T \vee 2))].$$

The proof now proceeds as in the proof of Proposition 8 using the bounds for the Brownian motion from §C.2 and the bound for  $Q^\Lambda(\cdot)$  from Lemma 11 above. ■

## References

- [1] M. Armony and I. Gurvich. When promotions meet operations: Cross-selling and its effect on call-center performance. Working Paper. New York University and Columbia University, New York. 2006.
- [2] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Prob.*, 5:49–77, 1995.
- [3] I. Gurvich, M. Armony, and C. Maglaras, Cross-Selling in a Call Center with a Heterogeneous Customer Population. Preprint 2006.
- [4] Whitt W. 2006. Martingale proofs of many-server heavy-traffic limits for Markovian queues. Working Paper. Columbia University, New York.
- [5] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Prob.*, 16:56–90, 2006.

- [6] I. Gurvich and A. Zeevi. Validity of heavy-traffic steady-state approximations in open queueing networks: sufficient conditions involving state-space collapse. Working Paper. Columbia University, New York.
- [7] Karatzas I., S.E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York.
- [8] A. Mandelbaum, Massey W. and Reiman M. 1998. Strong approximations for Markovian service networks, *Queueing Systems* **30** 149-201.
- [9] Puhalskii A. 1994. On the Invariance Principle For the First Passage Time, *Math. Oper. Res.* **19**(4) 946 - 954.