Call center staffing:
Service-level constraints and index priorities

Seung Bum Soh
Kellogg School of Management, Northwestern University, seung-soh@kellogg.northwestern.edu

Itai Gurvich
Kellogg School of Management, Northwestern University, i-gurvich@kellogg.northwestern.edu

Call centers attribute different values to different customer segments. These values are embedded into the call center’s staffing and prioritization optimization via differentiated quality-of-service targets. The Target Service Factor (TSF) formulation requires, for example, that 80% of VIP customers wait less than 20 seconds while setting the target to 30 seconds for non-VIP customers. The call center must determine the staffing level together with a prioritization rule that meets these targets at minimal cost. Thus far, this problem was not solved theoretically. In practice, due to the underlying complexity of these systems, the prioritization rule is often selected in a heuristic manner rather than being systematically optimized.

When considering the universe of prioritization policies, index rules provide a customizable and easy to define heuristic and for this reason are implemented in various call-center software packages. The family of index rules is vast and one may a priori expect that parameters can be adjusted to construct optimal solutions for the call center optimization problem and, in particular, for the TSF formulation.

Our results show that this, in general, is not the case. We first solve (asymptotically) the, yet unsolved, TSF staffing problem. Studying the family of all solutions to this problem, we prove that no instance of an index rule is optimal. This, we prove, hinges on a general property of index rules: these are characterized by a rather strong notion of priority differentiation: restricting attention to index rules (as is heuristically done in practice) is equivalent to requiring that a VIP customer always waits less than a regular (non-VIP) customer who arrives at the same time. This, in particular, implies that the use of index rules in practice can be rationalized through strong service-level differentiation requirements.

1. Introduction

Waiting times are regarded as central to customers’ service experience and are used as a key performance indicator for call centers. In the spirit of measuring what is managed, call centers use metrics that aggregate waiting-time related information. These are collectively referred to as QoS (Quality-of-Service) metrics. A notable example is the Target-Service-Factor (TSF, the fraction of customers who exceed a target waiting time) ; see Chapter 2 of Koole [2013] for further discussion of call-center performance metrics. TSF is one of the key performance indicators that are traced and targetted in call centers.

When the call center serves multiple customer classes, their relative importance is typically reflected in the TSF targets requiring, for example, that 80% of class-1 customers wait less than 20
seconds but 80% of class 2 customers wait less than 30 seconds. Solving the call-center’s staffing problem in a multi-class environment requires determining the capacity level and the prioritization rule that, combined, guarantee that the service-level targets are met at a minimal cost.

In the two-class model that we consider here, the staffing problem is given by

\[
\begin{align*}
\text{(TSF)} & \\
\min & \quad N \\
\text{s.t.} & \quad \mathbb{P}\{W_{1}^{N,\pi} > w_1\} \leq \alpha_1, \\
\text{s.t.} & \quad \mathbb{P}\{W_{2}^{N,\pi} > w_2\} \leq \alpha_2, \\
& \quad N \in \mathbb{Z}^+, \pi \in \Pi
\end{align*}
\]

where, informally at this stage, \(W_{i}^{N,\pi}\) represents the steady-state waiting time of class \(i\) when the number of servers is \(N\) and the prioritization policy is \(\pi\). We seek to find the minimal number of servers \(N\) together with a supporting prioritization policy \(\pi\) so that the TSF constraints are met. Here, \(\Pi\) denotes the family of “admissible” prioritization policies; see Definition 1. To the extent that the TSF constraint reflects relative importance, having \(\alpha_1 \leq \alpha_2\) together with \(w_1 \leq w_2\) corresponds to class 1 being the VIP class.

Truly solving this problem to optimality by jointly determining the staffing level and the priority rule is difficult. This paper is the first to offer a solution for a relatively simple network. In practice, call centers typically restrict attention to parametric families of prioritization rules and subsequently adjust the parameters and the staffing level to meet the QoS targets. This heuristic approach is understandable given the complexity of the problem. Degrees of freedom in the prioritization policy are built into the software that manages the real time call selection and routing of calls. The software provides sufficiently many adjustable parameters while maintaining simplicity.

A family of prioritization heuristics is that of index priorities; see e.g. Chan et al. [2014] or Koole and Pot [2006] where these are referred to as Time Function Scheduling (TFS). This family of policies is also well understood analytically in the context of convex holding cost minimization; see Van Mieghem [1995] and Mandelbaum and Stolyar [2004].

A customer is assigned an index upon arrival. As the customer waits, his index increases periodically. The system manager chooses the initial indices for each customer type as well as the update intervals and update-to levels (these are easily defined in a spreadsheet). This leads to an index function as the one in Figure 1 which is used in an Israeli Bank that serves three categories of retail customers. In real time, a server who becomes available serves the customer with the highest index.

Somewhat more formally, customer type \(i\) is associated with a piecewise-constant index \(g_i(\cdot)\) so that a server that becomes available chooses for service the customer at the head of queue \(i^* \in \arg\max_i g_i(w_i(t))\) where \(w_i(t)\) is the accumulated waiting time of the customers at the head
of the class-$i$ queue. This family of policies is intuitive to understand and easy to implement. Moreover, as both the updated periods and levels can be defined by the user it is expected to be rather flexible. Analytically understanding the role of index policies in constrained staffing optimization is practically important, due to their use in practice, and intellectually important, due to the attention they have received in the literature on the optimization of queues.

We first solve the TSF formulation which, despite its prevalence in practice, has not been solved even in the simplest form of multi-class queues. We provide a simple staffing expression based on the $M/M/N$ queue that explicitly captures the parameters of the problem. We prove that this staffing solution, together with a corresponding priority policy, is asymptotically optimal when the volume of arrivals is high.

Since we are interested in better understanding the role of index rules we proceed to show that all asymptotically optimal solutions to the TSF formulation are, in particular, non-index: all optimal prioritization policies share some non-conventional properties:

(i) Non-monotonicity of waiting time in congestion: For at least one customer class (be it VIP or
Regular), the waiting time has the structure depicted in Figure 2 – it is greater when the congestion (specifically the total queue) in the system is moderate relative to when the congestion is high.

(ii) Discontinuity of waiting time in congestion: Customers arriving within a short time interval may experience starkly different waiting times (this corresponds to the discontinuity point $\star$ in Figure 2).

Since, as we also prove, index rules generate waiting time profiles that are continuous and monotone in the total congestion, we conclude that if one heuristically restrict attention to index rules – even if one choose the “best” index functions – staffing costs are non-negligibly greater than the minimum necessary.

Thus, while the family of index rules is seemingly rich, there is no instance of an index rule that is asymptotically optimal (together with an appropriately chosen staffing level) for the TSF problem. To better understand this gap, it is useful to point to an interesting phenomenon one finds when considering the use of index rules in practice. Figure 3 displays the waiting-time processes\(^1\) on a single day (May 10th, 2007) in a three-class Bank call center that uses the index rule in Figure 1\(^2\).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.png}
\caption{Waiting-time processes in a large Israeli bank}
\end{figure}

A striking fact in this graph is that, sample path by sample path (we see a similar pattern on other days), a perfect real-time ranking between the classes is preserved, that is, “Private” customers wait less than their “Star” counterparts, whereas “Star” customers wait less than their

---

\(^1\)Specifically, for each 2.5 minutes interval the graph depicts the waiting time averaged over customers that arrived over that specific time interval

\(^2\)We are thankful to the Technion Service Enterprise Engineering Laboratory (SEE Lab) for the data
“Rainbow” counterparts. This reflects a rather strong notion of what it means to be a VIP customer – VIP customers wait less at all times (rather than, say, on average over a day or hour). Since in this Bank the targets \( w_i \) are identical, this is “translated” in the means that

\[
P \left\{ \frac{W_1^{N, \pi}}{w_1} \leq \frac{W_2^{N, \pi}}{w_2} \right\} \approx 1.
\]

This is a perfect service-level differentiation. One can view service-level differentiation (SLD) on a continuum parameterized by a constant \( \beta \in [0, 1] \). The case of \( \beta = 0 \) corresponds to requiring that differentiation holds only in aggregate by TSF formulation as in: on average at least 80% of VIP customers wait less than 20 seconds and at least 80% of Regular customers wait less than 30 seconds – whereas \( \beta = 1 \) reflects a perfect real time differentiation as observed in Figure 3.

The observation that perfect SLD is preserved in this data is generalizable. Perfect service-level differentiation is a characterizing property of index policies. That is, requiring perfect service-level differentiation but giving full freedom in the choice of the prioritization policy is equivalent to restricting attention to index rules in the TSF problem. More formally, the following two formulations are equivalent:

\[
\begin{align*}
\min N \\
s.t. & P \left\{ \frac{W_1^{N, \pi}}{w_1} > w_1 \right\} \leq \alpha_1, \\
& P \left\{ \frac{W_2^{N, \pi}}{w_2} > w_2 \right\} \leq \alpha_2, \\
& \frac{W_1^{N, \pi}}{w_1} \leq \frac{W_2^{N, \pi}}{w_2} = 1, \\
& N \in \mathbb{Z}_+, \pi \in \Pi.
\end{align*}
\] (TSF + SLD(1))

\[
\begin{align*}
\min N \\
s.t. & P \left\{ W_1^{N, \pi} > w_1 \right\} \leq \alpha_1, \\
& P \left\{ W_2^{N, \pi} > w_2 \right\} \leq \alpha_2, \\
& N \in \mathbb{Z}_+, \pi \in \Pi(Index).
\end{align*}
\] (TSF + Index)

In words, restricting attention to index rules (as is often heuristically done) is equivalent to requiring perfect SLD (i.e, \( \beta = 1 \)). Thus, whereas the apriori restriction of some call centers to index rules may be driven by the software packages that they use, their choice is consistent with optimality, provided that perfect SLD is desirable to them.

We close this introduction with a few important pointers to index rules in the literature. The use of index-rule based heuristics in the context of staffing subject to QoS targets is not grounded in theoretical foundation. However, it is the opposite case when considering holding cost (or waiting-time cost) minimization for given staffing levels. The best known results in this context pertain to the optimality of the Generalized \( c\mu \) (G\( c\mu \)) rule for the minimization of convex holding costs as pioneered by Van Mieghem [1995]. Waiting-cost minimization is also the subject of subsequent extensions to Van Mieghem’s original work; see e.g. Mandelbaum and Stolyar [2004] and many others that followed. Such rules were also shown to be optimal for some constrained staffing problems. A formulation based on Average Speed of Answer (ASA) constraints, for example, gives rise (as one possible solution) to a Fixed-Queue-Ratio rule – a special case of an index rule; see Gurvich and Whitt [2007].
Thus, in certain contexts, these rules are very well understood. We expand this body of work with some new insights about the role of index rules in constrained staffing problems. Constrained staffing problems in multi-class queues have been also studied and we refer the reader to Aksin et al. [2007] for some references. The typical paper in this literature seeks to identify one optimal (or nearly optimal) solution. To that literature we add our solution to the TSF formulation. However, questions as we raise here, that pertain to structural properties of the family of all asymptotically optimal solutions, do not arise naturally in that line of work. They arise naturally here when we seek to understand the role of index rules in staffing problems.

2. Model and analysis framework

We consider the two-class V model. Arrivals of class-$i$ customers follow a Poisson process $A_i = (A_i(t), t \geq 0)$ with rate $\lambda_i > 0$; $i = 1, 2$. Let $\lambda = \lambda_1 + \lambda_2$. The two Poisson processes are independent. Service times are exponential with rate $\mu$ and are independent across customers and independent of the arrival processes. Various metrics of interest will be superscripted by the staffing level $N$ and the prioritization policy $\pi$ to denote their dependence on these decision variables.

The TSF and SLD constraints were defined in the introduction using a virtual waiting time $W^{N,\pi}_N$. The Poisson Arrivals See Time Averages (PASTA) property guarantees that the fraction of customers who wait more than a target $w$ equals the fraction of time that the virtual waiting time exceeds this value.

We substitute the virtual waiting time with a proxy - the local average waiting time, which is based on the actual waiting times. Let $w^{N,\pi}_{i,k}$ be the waiting time of the $k^{th}$ class-$i$ customer to arrive. The local average waiting time of customers arriving over an interval $(t, t + \delta]$ is defined to be,

$$W^{N,\pi}_{\delta, i}(t) := \frac{1}{A_i(t + \delta) - A_i(t)} \sum_{k = A_i(t) + 1}^{A_i(t + \delta)} w^{N,\pi}_{i,k},$$

where $0/0$ is defined to be $0$.

Local average is the measure used in Figure 3 where $\delta$ is 2.5 minute. If $\delta$ is kept sufficiently small, the local average captures the dynamics and variation of waiting time. The use of local average (rather than virtual waiting time) helps us overcome a technical difficulty (see Lemma 5.1).

We assume throughout that

$$\alpha_1, \alpha_2 > 0, \ \alpha_1 \leq \alpha_2 \text{ and } \alpha_1 + \alpha_2 < 1.$$  

(2)

If $\alpha_1 + \alpha_2 \geq 1$, the staffing problem trivializes and, even for small values of $w_1, w_2 > 0$, any staffing level $N > \lambda / \mu$ is feasible; see the discussion on page page 285 of Gurvich et al. [2008],

We allow the controller to use information about the length of the queues, the elapsed service times of customers that are in service and the accumulated waiting times of the all customers in the
queues and we let $X(t)$ be the state descriptor that captures all this information; see the appendix for the formal definition.

**Definition 1 (admissible policies):** A prioritization policy $\pi$ is admissible if:

1. It does not use admission control – all customers are served.
2. It serves customers in a first-come first-served (FCFS) fashion within each class.
3. It is work conserving and stationary with respect to $X$.

Let $\Pi$ be the family of admissible policies.

Non-preemption and FCFS within each class are both natural assumptions for our application. The literature does have cases in which non-work-conserving policies are shown to be optimal or asymptotically optimal (see Gurvich et al. [2008]). The results in Gurvich et al. [2008] are driven, however, by order-of-magnitude differences between the probability targets of the various classes (say $\alpha_1 = 0.01$ but $\alpha_2 = 0.2$) that lead inherently to solutions that induce perfect SLD rendering our main research questions irrelevant. Work conservation has the following immediate consequence which we state formally and repeatedly refer to.

**Lemma 2.1** Fix the number of servers $N$, the service rate $\mu$ and the arrival rates $\lambda_1, \lambda_2$. Then, under any admissible policy $\pi$, the total number of customers in the system $X_\Sigma(t)$ (respectively the total queue $Q_\Sigma(t)$) has the law of the total number of customers (respectively the queue length) in an $M/M/N$ queue with arrival rate $\lambda = \lambda_1 + \lambda_2$, service rate $\mu$ and $N$ servers.

**Many-server analysis:** The problem TSF+SLD($\beta$) is difficult to solve exactly. Instead, we resort to identifying solutions that are asymptotically optimal as the system size grows. A series of solutions is asymptotically optimal if the induced optimality gap is negligible in a central-limit-theorem (CLT) scaling. This is consistent with many studies in the context of call-center optimization; see e.g. Armony [2005], Armony and Mandelbaum [2011], Borst et al. [2004]. We consider a sequence of $V$ models with the total arrival rate $\lambda = \lambda_1 + \lambda_2$ increasing along the sequence but keeping the service rate $\mu$ fixed. We assume that there exists $a_i > 0$, $i = 1, 2$ so that $a_1 + a_2 = 1$ and $\lambda_i = a_i \lambda$ for $i = 1, 2$ and all $\lambda$.

All relevant quantities are superscripted by $\lambda$ to make the dependence on $\lambda$ explicit. As is standard in the literature, the targets $w_1$ and $w_2$ scale with $\lambda$: $w_1^\lambda = \bar{w}_i/\sqrt{\lambda}$, $i = 1, 2$ where $\bar{w}_i$’s are fixed strictly positive constants. This guarantees that the system is in the Halfin-Whitt regime (see Lemma 5.5 below). This is mostly a mathematical artifact, one that facilitates a CLT type of analysis in a relatively “clean” way.
To simplify the notation, we denote a staffing-policy pair \((N, \pi)\) by \(\xi\) and let \(N(\xi)\) and \(\pi(\xi)\) be, respectively, the staffing and prioritization components of \(\xi\). The problem \(\text{TSF}+\text{SLD}(\beta)\) is now re-written as:

\[
\begin{align*}
\min_{\xi^\lambda} & \quad N(\xi^\lambda) \\
\text{s.t.} & \quad P\left\{ W_{\delta,1}^{\xi^\lambda,\lambda}(\infty) > w_1^\lambda \right\} \leq \alpha_1, \\
& \quad P\left\{ W_{\delta,2}^{\xi^\lambda,\lambda}(\infty) > w_2^\lambda \right\} \leq \alpha_2, \\
& \quad P\left\{ W_{\delta,1}^{\xi^\lambda,\lambda}(\infty)/w_1^\lambda \leq W_{\delta,2}^{\xi^\lambda,\lambda}(\infty)/w_2^\lambda \right\} \geq \beta, \\
& \quad \xi^\lambda \in \mathbb{Z}_+ \times \Pi.
\end{align*}
\]

\(W_{\delta,i}^{\xi^\lambda,\lambda}(\infty)\) should be read as the \(\delta\) average (as in (1)) when the system is initialized (at \(t=0\)) with its steady-state distribution, the customer incoming rate is \(\lambda\) and the staffing-policy pair \(\xi^\lambda\) is used.

**Definition 2 (asymptotic feasibility):** A sequence of staffing-policy pairs \(\{\xi^\lambda\}\) is asymptotically feasible, if \(\xi^\lambda \in \mathbb{Z}_+ \times \Pi\) for all \(\lambda\) and for each \(\epsilon > 0\),

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} P\left\{ W_{\delta,i}^{\xi^\lambda,\lambda}(\infty) > w_1^\lambda (1 - \epsilon) \right\} \leq \alpha_i(1 + \epsilon), i = 1, 2,
\]

and

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} P\left\{ W_{\delta,1}^{\xi^\lambda,\lambda}(\infty)/w_1^\lambda \leq W_{\delta,2}^{\xi^\lambda,\lambda}(\infty)/w_2^\lambda + \epsilon \right\} \geq \beta(1 - \epsilon).
\]

We say that a sequence \(\{N^\lambda\}\) is an asymptotically feasible sequence of staffing levels for (3) if \(N^\lambda = N(\xi^\lambda)\) for a sequence \(\{\xi^\lambda\}\) of asymptotically feasible staffing-policy pairs.

**Definition 3 (asymptotic optimality):** A sequence of staffing-policy pairs \(\{\xi^\lambda\}\) is asymptotically optimal if it is asymptotically feasible and

\[
\left[ N(\xi^\lambda) - N(\tilde{\xi}^\lambda) \right]^+ = o\left( \sqrt{\lambda} \right) \text{ as } \lambda \to \infty,
\]

for any other sequence \(\{\tilde{\xi}^\lambda\}\) of asymptotically feasible staffing-policy pairs.

We say that a sequence \(\{N^\lambda\}\) is an asymptotically optimal sequence of staffing levels for (3) if \(N^\lambda = N(\xi^\lambda)\) for an asymptotically optimal sequence \(\{\xi^\lambda\}\) of staffing policy pairs.

### 3. The proposed solution to TSF+SLD and its implications

This section contains our main results. We identify an asymptotically optimal staffing and prioritization solution. The two components of our solution are inter-dependent. The proposed staffing level is feasible if one uses it together with our proposed prioritization policy. For simplicity of exposition, we discuss them separately (each in a dedicated subsection), starting with the staffing component. Theorems stated in this section are proved in §5 and in the technical supplement Soh and Gurvich [2015].
3.1. Optimal staffing and the cost of SLD

To specify our staffing recommendation, let $Q^{N,\lambda}(\infty)$ be a random variable whose distribution is that of the steady-state queue length in an $M/M/N$ queue with arrival rate $\lambda$, service rate $\mu$ and $N$ servers (as $\mu$ is fixed throughout, we omit it from the superscript).

For each $\lambda$, let $N^\lambda_\star$ be the solution of the following $M/M/N$ staffing problem

$$N^\lambda_\star = \min \left\{ N \in \mathbb{Z}^+ : \mathbb{P}\left\{ Q^{N,\lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2 \right\}.$$  

(4)

Notice that, by (2), $\min \{ \alpha_1, 1 - \beta \} + \alpha_2 < 1$ holds

**Theorem 1** *(asymptotically optimal staffing)* The sequence $\{N^\lambda_\star\}$ is an asymptotically optimal sequence of staffing levels for (3).

The prioritization component $\pi(\xi^\lambda)$ of $\xi^\lambda$, constructed in the next subsection, guarantees that all the inequalities $W_{1,1}^{\xi,\lambda}(\infty) \leq w_1^\lambda$, $W_{2,2}^{\xi,\lambda}(\infty) \leq w_2^\lambda$ and $W_{1,2}^{\xi,\lambda}(\infty)/w_1 \leq W_{2,1}^{\xi,\lambda}(\infty)/w_2$ are (asymptotically) satisfied when the total queue length is smaller than $\lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda$. Thus, the choice of the staffing level (4) guarantees that the total violation is bounded by $\min \{ \alpha_1, 1 - \beta \} + \alpha_2$. The challenge in designing the priorities is to make sure that the this “violation budget” is distributed correctly between the three constraints; for example, the TSF constraint for class 2 absorbs at most $\alpha_2$ of the total violation budget.

**Remark 1** *(aggregate-based staffing)* Theorem 1 maps TSF+SLD($\beta$) into a staffing problem for a single class queue. The fact that this solution works for the multiclass queue will be supported by our careful choice of the prioritization rule. Such a reduction to a one dimensional model is a recurring theme in recent literature on staffing; see e.g. Armony and Mandelbaum [2011], Gurvich et al. [2008], Gurvich and Whitt [2007]. In the latter two papers, a global ASA constraint allows for a relatively simple mapping of the parameters into the single-class staffing problem. This mapping is more subtle in our setting as is reflected in the non-trivial way in which the constants $\alpha_1, \alpha_2$ and $\beta$ appear on the RHS of (4).

**Theorem 2** The sequence $\{N^\lambda_\star\}$ satisfies

$$N^\lambda_\star = \frac{\lambda}{\mu} + \eta^* \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}),$$

where $\eta^*$ is convex increasing in the SLD degree $\beta$. It is strictly increasing for $\beta \geq 1 - \alpha_1$.

Increasing the degree of SLD beyond a certain level (in particular, setting $\beta = 1$) is thus costly: the optimal staffing level for the TSF formulation is too small for TSF+SLD(1). Conversely, relaxing the SLD requirement reduces the staffing costs.
3.2. Prioritization and the optimality of index policies

The asymptotically optimal staffing \( \{N^*_\lambda\} \) is coupled with a carefully chosen sequence of policies \( \{\pi^*_\lambda\} \) such that the sequence of staffing-policy pairs \( \{\xi^\lambda = (N^*_\lambda, \pi^*_\lambda)\} \) is asymptotically optimal in the sense of Definition 3. The policy that we construct in this section is an instance of so-called tracking policies.

Tracking policies defined below are also admissible policies (customers are served in a FCFS manner within the same class). The queue length of class \( i \) at time \( t \) for \( i = 1, 2 \) is denoted by \( Q_i(t) \) and \( Q_\Sigma(t) \) is the total queue length, i.e., \( Q_\Sigma(t) = Q_1(t) + Q_2(t) \). From these, normalized queue lengths are defined as follows:

\[
\hat{Q}_i^\lambda(t) := \frac{Q_i^\lambda(t)}{\sqrt{\lambda}}, \quad \hat{Q}_\Sigma^\lambda(t) := \hat{Q}_1^\lambda(t) + \hat{Q}_2^\lambda(t).
\]

**Definition 4 (queue tracking policies)** Let \( f = (f_1, f_2) : \mathbb{R}_+ \to \mathbb{R}_+^2 \) be a non-negative function such that \( f_1(x) + f_2(x) = x \) for all \( x \geq 0 \). A server that becomes available at time \( t \) chooses the queue to serve according to the following criterion:

1. If a single queue is non-empty, admit to service the customer at the head of this non-empty queue.

2. If both queues are non-empty and \( \hat{Q}_1(t) - f_1(\hat{Q}_\Sigma(t)) \neq \hat{Q}_2(t) - f_2(\hat{Q}_\Sigma(t)) \), choose the class \( i \) with the positive value of \( \hat{Q}_i(t) - f_i(\hat{Q}_\Sigma(t)) \).

3. If both queues are non-empty and \( \hat{Q}_1(t) - f_1(\hat{Q}_\Sigma(t)) = \hat{Q}_2(t) - f_2(\hat{Q}_\Sigma(t)) \), choose class 1.

An arriving customer is immediately assigned to an available server if there are such servers. If all servers are busy upon this customer’s arrival, the customer is placed in the queue.

Figure 5(LHS) illustrates the tracking function mechanism. It plots the targeted value for \( Q_1(f(\hat{Q}_\Sigma)) \) vs. the value of \( \hat{Q}_\Sigma \). If at time \( t \), \( \hat{Q}_1(t) > f_1(\hat{Q}_\Sigma(t)) \) (as in point A in the figure) the rule
Figure 5  Tracking: (LHS) General description (RHS) Proposed tracking function under sub-perfect SLD ($\beta < 1$) prioritizes class 1 so as to pull it back toward $f_1(\hat{Q}_S(t))$. If $\hat{Q}_1(t) < f_1(\hat{Q}_S(t))$ as in point B in the graph, the rule prioritizes class 2 over class 1 so as to push queue 1 towards $f_1(\hat{Q}_S(t))$. The tracking policy targets $\hat{Q}_i(t) \approx f_i(\hat{Q}_{\Sigma}(t))$. In heavy-traffic, a sample path version of Little’s law holds:

$$Q_i(t) \approx \lambda_i W_i(t).$$  

(5)

The family of tracking functions is immense. The challenge is to carefully choose the tracking function $f$ so that together with our proposed staffing the solution is asymptotically optimal (in particular, asymptotically feasible).

**Optimal tracking functions:** We use the tracking function $f^*_1$ defined below, as plotted also in Figure 5(RHS):

$$f^*_1(x) = \begin{cases} 
0, & 0 \leq x < \frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 + a_2 \bar{w}_2} \kappa, \\
\frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 + a_2 \bar{w}_2} x - \frac{a_2 \bar{w}_2}{2}, & \frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 + a_2 \bar{w}_2} \kappa \leq x < a_1 \bar{w}_1 + a_2 \bar{w}_2 - 2 \left(1 + \frac{a_1 \bar{w}_1}{a_2 \bar{w}_2}\right) \kappa, \\
\frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 + a_2 \bar{w}_2} x - a_2 \bar{w}_2 + \kappa, & a_1 \bar{w}_1 + a_2 \bar{w}_2 - 2 \left(1 + \frac{a_1 \bar{w}_1}{a_2 \bar{w}_2}\right) \kappa \leq x < a_1 \bar{w}_1 + a_2 \bar{w}_2, \\
\frac{a_1 \bar{w}_1 - \kappa}{a_1 \bar{w}_1 + a_2 \bar{w}_2}, & a_1 \bar{w}_1 + a_2 \bar{w}_2 \leq x < \bar{q}, \\
\frac{a_1 \bar{w}_1 - \kappa}{a_1 \bar{w}_1 + a_2 \bar{w}_2} x - a_2 \bar{w}_2 + \kappa, & \bar{q} \leq x.
\end{cases}$$

(6)

and $f^*_2(x) = x - f^*_1(x)$.

The threshold $\bar{q}$ is given by

$$\bar{q} = a_1 \bar{w}_1 + a_2 \bar{w}_2 + \log \left\{ \frac{a_2 + \min\{\alpha_1,1-\beta\}}{\alpha_2} \right\} \eta^*(\beta)$$

(7)

where $\eta^*(\beta)$ is the unique solution (keeping other parameters constant) to

$$\left(1 + \frac{\eta^*(\eta)}{\Phi(\eta)}\right)^{-1} \exp(-\eta(a_1 \bar{w}_1 + a_2 \bar{w}_2)) = \min\{\alpha_1,1-\beta\} + \alpha_2.$$

(8)
Here, $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the standard normal density and distribution functions. The constant $\kappa$ can be any
\[
0 < \kappa \leq \frac{a_1 \bar{w}_1 a_2 \bar{w}_2}{2a_1 \bar{w}_1 + a_2 \bar{w}_2}.
\]

Figure 5 (RHS) considers the case of sub-perfect SLD ($\beta < 1$). The dashed line is the function $f^*_1(x_{\Sigma})$ which is the “target” value for $x_1$. The thick solid line is where $x_1 = a_1 \bar{w}_1 x_{\Sigma} / (a_1 \bar{w}_1 + a_2 \bar{w}_2)$. When $(\bar{Q}_1(t), \bar{Q}_2(t))$ is on this line, we have by (5) that $W_{\delta,1}(t)/w_1^\lambda \approx W_{\delta,2}(t)/w_2^\lambda$. As long as queues are below this line we expect to have $W_{\delta,1}(t)/w_1^\lambda < W_{\delta,2}(t)/w_2^\lambda$. It makes sense, then, to refer to this line as the “SLD line”. The key in constructing the function $f^*$ is figuring out how to divide the real line $[0, \infty)$ to segments that differ with respect to their position relative to SLD line (below or above) and relative to the values $w_1^\lambda$ and $w_2^\lambda$. Suppose $\bar{Q}_1(t) \approx f^*_1(\bar{Q}_2(t))$, i.e., that the tracking works. The function $f^*_1$ is constructed so that if $\bar{Q}_2(t) \leq a_1 \bar{w}_1 + a_2 \bar{w}_2$ holds, $\bar{Q}_1(t) \leq a_1 \bar{w}_1$, $\bar{Q}_2(t) \leq a_2 \bar{w}_2$ and $\bar{Q}_1(t)/a_1 \bar{w}_1 \leq \bar{Q}_2(t)/a_2 \bar{w}_2$ hold. By (5), $W_{\delta,1}(t) \leq w_1^\lambda$, $W_{\delta,2}(t) \leq w_2^\lambda$ and $W_{\delta,1}(t)/w_1^\lambda \leq W_{\delta,2}(t)/w_2^\lambda$ also hold. But if $\bar{Q}_2(t) > a_1 \bar{w}_1 + a_2 \bar{w}_2$, it is impossible to satisfy all the inequalities and one has to choose which constraints to sacrifice while others are met. Our choice of the function is such that, when $\bar{Q}_2(t) \in [a_1 \bar{w}_1 + a_2 \bar{w}_2, \bar{q}]$, the inequalities $W_{\delta,1}(t) \leq w_1^\lambda$ and $W_{\delta,1}(t)/w_1^\lambda \leq W_{\delta,2}(t)/w_2^\lambda$ are violated. If $t$ is such that $\bar{Q}_2(t) \in [\bar{q}, \infty)$, then the inequality $W_{\delta,2}(t) \leq w_2$ is violated.

The choice of the parameter $\bar{q}$, combined with the optimal staffing in Theorem 1, guarantees that $\mathbb{P}\left\{ \bar{Q}_2(t) \in [a_1 \bar{w}_1 + a_2 \bar{w}_2, \bar{q}] \right\} \approx \min\{\alpha_1, 1 - \beta\}$ and $\mathbb{P}\left\{ \bar{Q}_2(t) \in [\bar{q}, \infty) \right\} \approx \alpha_2$. Then, $\mathbb{P}\{W_{\delta,1}(t) > w_1\} \approx \mathbb{P}\{W_{\delta,1}(t)/w_1^\lambda > W_{\delta,2}(t)/w_2^\lambda\} \approx \min\{\alpha_1, 1 - \beta\}$ and $\mathbb{P}\{W_{\delta,2}(t) > w_2\} \approx \alpha_2$ hold and the solution is feasible. In the following theorem, $\pi^*_\lambda$ denotes the tracking policy which applies the tracking function $f^*$.

**Theorem 3 (asymptotic optimality)** Let $N^\lambda$ be as in (4). Then, the sequence of staffing-policy pairs $\{\xi^\lambda = (N^\lambda, \pi^*_\lambda)\}$ is asymptotically optimal.

**Remark 2 (on the interplay of cost reduction and prioritization)** There are degrees of freedom in choosing the tracking function – there are various choices that generate the same optimality result. We chose $f^*$ so that SLD is violated only when the total queue is small (specifically, less than $T^\lambda$). With this choice, $f^*$ has the appealing property that class-1 (the VIP) customers get superior service when the total congestion in the system is large. Thus, reducing $\beta$ (from $\beta < 1$) allows for a cost reduction in terms of staffing (recall Theorem 2) while making sure that the VIP customers get superior service when it really matters.

One thing that may strike the reader as non-standard in Figure 5(RHS) is the discontinuity and non-monotonicity of the tracking function $f^*$. This is not a consequence of the specific way in which
we choose the functions – as the next theorem shows it cannot be avoided without compromise to optimality.

For the formal statement, we say that a tracking policy is non-monotone if at least one of the functions $f_1(\cdot)$ and $f_2(\cdot)$ is non-monotone. A function $f$ is defined to be piecewise Lipschitz if it has a finite number of discontinuity points $\{x_1, \ldots, x_m, \}$ and is Lipschitz continuous on each interval (including $[0, x_1)$ and $[x_m, \infty)$).

**Theorem 4 (non-monotonicity and discontinuity)** Assume $\beta < 1$ (sub-perfect SLD) and let $f$ be a piecewise Lipschitz tracking function such that the sequence $\{(N_\star^\lambda, \pi_\star^\lambda)\}$ is an asymptotically optimal sequence of staffing-policy pairs. Then, $f$ is discontinuous and non-monotone.

The optimal staffing allows for a limited “budget” that must be carefully allocated to the three different constraints (the two TSF constraints and the SLD constraints). For different values of $\hat{Q}_\Sigma$ one must choose which constraints are “sacrificed” in favor of meeting the others. What we prove is that the only way to do that while minimizing the total budget is to alternate the priorities and this, in particular, introduces the non-monotonicity and the discontinuity that are stated in the theorem.

**On the equivalence of index policies and perfect SLD:** Theorem 4 proves that, asymptotically, optimal tracking functions must be discontinuous and non-monotone for $\beta < 1$. That is, imposing monotonicity (which is a property of index policies; see below) is related to requiring perfect SLD ($\beta = 1$). We next show that requiring perfect SLD and restricting the tracking functions to be monotone is indeed equivalent in asymptotic sense. More importantly, proving that monotone tracking rules are equivalent to index policies, allows us to conclude the equivalence between imposing a perfect SLD constraint and restricting the optimization to use only index policies.

To formalize this result, recall that an index policy is one that, upon service completion at time $t$, admits to a customer from class

$$i^* = \arg\max_i g_i(Q_i(t)),$$

where $g_1$ and $g_2$ are non-negative increasing functions. We denote by $\Pi(index) \subset \Pi$ the family of index policies.

**Theorem 5 (equivalence of index policies and SLD(1))** The following formulations have asymptotically optimal solutions in common:

$$\min_{\xi} N(\xi) \quad \text{s.t.} \quad \mathbb{P}\left\{ W_{\delta,1}^\xi(\infty) > w_1 \right\} \leq \alpha_1,$$

$$\mathbb{P}\left\{ W_{\delta,2}^\xi(\infty) > w_2 \right\} \leq \alpha_2,$$

$$\mathbb{P}\left\{ W_{\delta,1}^\xi(\infty) / w_1 \leq W_{\delta,2}^\xi(\infty) / w_2 \right\} = 1,$$

$$\xi \in \mathbb{Z}_+ \times \Pi.$$

$$\min_{\xi} N(\xi) \quad \text{s.t.} \quad \mathbb{P}\left\{ W_{\delta,1}^\xi(\infty) > w_1 \right\} \leq \alpha_1,$$

$$\mathbb{P}\left\{ W_{\delta,2}^\xi(\infty) > w_2 \right\} \leq \alpha_2,$$

$$\mathbb{P}\left\{ W_{\delta,1}^\xi(\infty) / w_1 \leq W_{\delta,2}^\xi(\infty) / w_2 \right\} = 1,$$

$$\xi \in \mathbb{Z}_+ \times \Pi(index).$$
Theorem 5 is argued in three steps stated in Propositions 6-8. The first, Proposition 6, shows that Gcµ policies and monotone tracking policies are separate mathematical representations of the same policy.

**Proposition 6 (equivalence of index policies and monotone tracking policies)** Given a monotone tracking function \( f \), there exists an index function \( g \) such that the monotone tracking policy with \( f \) is equivalent to the index policy with index \( g \). Similarly, given an index function \( g \), there exists a monotone tracking \( f \) such that the index policy with \( g \) is equivalent to the tracking policy with \( f \). The equivalence is in the sense that at any given state of \((Q_1(t), Q_2(t))\), both policies take the same action.

Given this equivalence, to prove Theorem 5, it remains to show that TSF+SLD(1) is equivalent to the following

\[
\min_{\xi} N(\xi) \\
\text{(TSF + MT) s.t. } \mathbb{P}\{W_{\delta,1}^\xi(\infty) > w_1\} \leq \alpha_1, \\
\mathbb{P}\{W_{\delta,2}^\xi(\infty) > w_2\} \leq \alpha_2, \\
\xi \in \mathbb{Z}_+ \times \Pi(MT),
\]

where \( \Pi(MT) \subset \Pi \) is the family of monotone tracking policies. Note that a monotone tracking function must also be continuous by \( f(x) = f_1(x) + f_2(x) \).

This equivalence is proved in the next two propositions. The first, Proposition 7, proves that there exists a monotone tracking policy which, with the staffing level from Theorem 1, satisfies the constraints in TSF+SLD(1).

**Proposition 7 (nearly optimal monotone solutions for TSF+SLD(1))** For any \( \vartheta > 0 \), there exists a monotone tracking function \( f^\vartheta \) and a sequence \( \{N_\lambda^\vartheta\} \) such that \( \{(N_\lambda^\vartheta, \pi_\lambda^\vartheta)\} \) is asymptotically feasible for TSF+SLD(1) and \( |N_\lambda^\vartheta - N_*^\lambda| \leq \vartheta \sqrt{\lambda} \).

The nearly optimal tracking function, whose existence is established in Proposition 7, is depicted in Figure 6 and explicitly specified as follows:

\[
f_1^\vartheta(x) = \begin{cases} 
0, & 0 \leq x < \frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1^\lambda}, \\
\frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1^\lambda} x - \kappa^\vartheta, & \frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1^\lambda} \leq x < \frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 + a_2 \bar{w}_2}, \\
\frac{a_1 \bar{w}_1 + a_2 \bar{w}_2}{a_1 \bar{w}_1 - \kappa^\vartheta}, & a_1 \bar{w}_1 + a_2 \bar{w}_2 \leq x.
\end{cases}
\]

The inequalities \( W_{\delta,1}(t) \leq w_1^\lambda, W_{\delta,2}(t) \leq w_2^\lambda \) and \( W_{\delta,1}(t)/w_1^\lambda \leq W_{\delta,2}(t)/w_2^\lambda \) hold at time in which \( Q_\Sigma(t) \leq q_0^\vartheta \) and only the constraint \( W_{\delta,2}(t) \leq w_2^\lambda \) is violated when \( Q_\Sigma(t) > q_0^\vartheta \).
Proposition 8 closes the equivalence argument in showing that, in terms of staffing, requiring monotonicity is as demanding as requiring SLD(1). As we identified, in Proposition 7, a monotone tracking solution that is asymptotically optimal for TSF+SLD(1) we conclude that (asymptotically) TSF+MT and TSF+SLD(1) are equivalent and, by Proposition 6, so are TSF+index and TSF+SLD(1).

**Proposition 8 (equivalence of monotone tracking and SLD(1) under TSF constraints)**

Let \( \{ \xi^1_\lambda \} \) be a series of asymptotic solutions for TSF+SLD(1) and \( \{ \xi^2_\lambda \} \) be one for TSF+MT. Then,

\[
\liminf_{\lambda \to \infty} \frac{N(\xi^2_\lambda) - N(\xi^1_\lambda)}{\sqrt{\lambda}} \geq 0.
\]

**Remark 3 (the cost of index policies)** Figure 7 displays the minimal staffing requirement to satisfy the TSF constraints, \( \mathbb{P}\{W_1 \leq 10 \text{ sec.}\} \leq 0.2 \), \( \mathbb{P}\{W_2 \leq 10 \text{ sec.}\} \leq 0.2 \), when the arrival rate of each class is 50 customers per minute and the average service time is 5 minutes. The optimal staffing is derived using Theorem 1. When adding the SLD constraint, the cost changes with the value of \( \beta \) in the SLD constraint as \( \beta \) varies between 0.65 and 1. This is is captured by the solid line. When, instead of having the SLD constraint, one restricts attention to index policies the staffing level is always 517 (captured by the dashed line). The gap between the two lines is the cost of the the restriction to index policies. For \( \beta = 1 \), consistent with our equivalence result, there is no gap. For small value of \( \beta \) the gap is 7 servers, corresponding to a non-negligible gap relative to the square root of the system size.
4. Numerical experiments

The purpose of this section is to illustrate how our solutions, derived via asymptotic analysis, perform for a given system. We consider 10 parameter sets that correspond to the original TSF formulation (with no SLD constraint). Across parameter combinations we vary the arrival rate, the target parameters \( w_1, w_2 \) and \( \alpha_1, \alpha_2 \), for the TSF constraints \( \mathbb{P}\{W_i > w_i\} \leq \alpha_i, \ i = 1, 2; \) see Table 1.

The mean service time is set to 1 throughput. For staffing we use the recommendation in equation (4) which with \( \beta = 0 \), reduces to

\[
N^\lambda_*(\vec{\lambda}, \vec{w}, \vec{\alpha}) = \min \left\{ N \in \mathbb{Z}_+: \mathbb{P}\{Q_{N,\lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda\} \leq \alpha_1 + \alpha_2 \right\}.
\]

This formula is given explicitly in terms of the parameters and hence requires no asymptotics.

Given \( N^\lambda_* \) (different for each parameter combination), we compute \( \eta = (N^\lambda_* - (\lambda_1 + \lambda_2)/\mu)/\sqrt{\lambda} \) and use it to design the optimal tracking functions as in (6). This gives us the tracking policy—denote this by \( \pi \).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_1 )</td>
<td>125</td>
<td>100</td>
<td>250</td>
<td>200</td>
<td>500</td>
<td>400</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>125</td>
<td>150</td>
<td>250</td>
<td>300</td>
<td>500</td>
<td>600</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>( w_1 ) (min)</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.025</td>
<td>0.025</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>( w_2 ) (min)</td>
<td>0.3</td>
<td>0.3</td>
<td>0.15</td>
<td>0.15</td>
<td>0.07</td>
<td>0.07</td>
<td>0.03</td>
<td>0.03</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.1</td>
<td>0.15</td>
<td>0.1</td>
</tr>
<tr>
<td>( \alpha_2 )</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.2</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 1  Parameter sets for testing the proposed solution for the TSF formulation
For each parameter set in Table 1, we simulate the two-class queue with the proposed solution \((N^\lambda, \pi)(\overline{\lambda}, \overline{w}, \overline{\alpha})\) for 40000 time units—corresponding to an order of 8 million arrivals for the first arrival-rate combination and 40 million for the last one. At the end of the replication we record \[\frac{1}{A_i(T)} \sum_{k=1}^{A_i(T)} 1 \{w_{i,k} > w_i\} \] with \(A_i(T)\) being the number of class \(i\) customers that arrived over the simulation horizon and \(w_{i,k}\) is the actual waiting time of the \(k^{th}\) class \(i\) customer processed. Since almost surely,
\[
\lim_{T \to \infty} \frac{1}{A_i(T)} \sum_{k=1}^{A_i(T)} 1 \{w_{i,k} > w_i(1 - \epsilon)\} \to \mathbb{P}\{W_i(\infty) > w_i(1 + \epsilon)\}, \text{ as } T \to \infty.
\]
This statistic provides a proxy for the performance of the policy relative to the target \(\alpha_i\) for class \(i\). The “test” for our proposed solution is that
\[
\mathbb{P}\{W_i(\infty) > w_i(1 + \epsilon)\} \leq \alpha_i(1 + \epsilon).
\]

Figure 8 displays the results for \(\epsilon = 0.1\). For each of the parameter sets we plot two columns corresponding to \(\mathbb{P}\{W_1 > w_1\}\) and \(\mathbb{P}\{W_2 > w_2\}\). The labels report the values. We color in red the labels that violate \(\alpha_i(1 + \epsilon)\) target. The performance is rather impressive with the exception of parameters sets 3, 5 and 7, 8. Next, we add a single server to each of these violating cases. We find that merely adding one server corrects for all cases except for 7, 8 that still violate their target \(\alpha = (0.15, 0.2)\) and \(\alpha = (0.1, 0.15)\) respectively. The reason for this violation seems to be the very strict time targets \(w_1 = 0.025\) and \(w_2 = 0.03\). Because for these the staffing has to be so high that

\(^3\) The simulation is coded in ARENA
the system is in fact pushed out of heavy-traffic and so that the approximation do not work as well.

Finally, we show that these red values actually become more precise as we increase the system size. Specifically, for those combinations we double the arrival rate of each class and re-compute the required staffing and re-simulate.

5. Proofs

This section is dedicated to the proofs of our main results. §5.1 proves Theorem 3 which is on the asymptotic optimality of our proposed staffing and policy. §5.2 contains proofs of the theorems that deal with the equivalence of index policies and perfect SLD. The proofs of other theorems and auxiliary lemmas are in Soh and Gurvich [2015].

5.1. Proof of Theorem 3:

We divide the lengthy proof into sub-modules as follows: we first prove an equivalence between the waiting time formulation (3) and a queue-based formulation. This may be somewhat expected given (5). The challenge, however, is to establish such a “Little’s law” result at the formulation level (rather than for a given policy as is typically the case). It is in this step where considering local averages is instrumental. Building on this asymptotic equivalence, a lower bound for the local-average-queue-based formulation provides an asymptotic lower bound for the waiting-time formulation. We identify the staffing recommendation $N_{\lambda}^*$ in (4) as such a bound; see §5.1.1.

In §5.1.2 we proceed to show that the tracking policy with our proposed function $f^*$ indeed has the desired asymptotic properties in that, when used, it results in $\hat{Q}_{\lambda i}(t) \approx f^*(\hat{Q}_{\lambda i}(t))$, where

$$\hat{Q}_{\lambda i}(t) := \frac{Q_{\lambda i}(t)}{\sqrt{\lambda}}, \quad \hat{Q}_{\lambda \Sigma}(t) := \hat{Q}_{\lambda 1}(t) + \hat{Q}_{\lambda 2}(t).$$

Whereas this type of so-called state-space collapse is, by now, a standard result, a challenge arises here from the possible discontinuity of the tracking function $f^*$.

We combine the pieces in §5.1.3 to have the proof of Theorem 3. We prove there that our proposed solution is asymptotically feasible. Since its staffing component (and, in particular, its cost) is a lower bound on any such solutions, we conclude that our solution is asymptotically optimal.

5.1.1. From waiting times to queues

Before directly showing the relationship of (5), we introduce local average queue lengths which connect local average waiting times and queue lengths:

$$Q_{\delta, t}^{\lambda, \lambda} (t) = \frac{1}{\delta} \int_{t}^{t+\delta} Q_{s}^{\lambda, \lambda} (s) ds.$$

(10)

The following lemma shows that if a sequence $\{N_{\lambda}^*\}$ is asymptotically feasible for (3) it must (asymptotically) satisfy the appropriate constraints in terms of queues and vice versa. We write $q_{\lambda i} := \lambda_i w_{\lambda i}$ for $i = 1, 2$ and let $Q_{\delta, t}^{\lambda, \lambda} (\infty)$ be the $\delta$ average when the queue length is initialized (at time $t = 0$) with its steady-state distribution.
Lemma 5.1 Let \( \{\xi^\lambda\} \) be a sequence of admissible staffing-policy pairs. Then the followings are equivalent.

1. For each \( \epsilon_1 > 0 \), the following holds.

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ W_{\delta, 1}^{\xi^\lambda} (\infty) > w_1^\lambda (1 - \epsilon_1) \right\} \leq \alpha_1 + \epsilon_1,
\]

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ W_{\delta, 2}^{\xi^\lambda} (\infty) > w_2^\lambda (1 - \epsilon_1) \right\} \leq \alpha_2 + \epsilon_1,
\]

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ W_{\delta, 1}^{\xi^\lambda} (\infty) / w_1^\lambda \leq W_{\delta, 2}^{\xi^\lambda} (\infty) / w_2^\lambda - \epsilon_1 \right\} \geq \beta - \epsilon_1.
\]

2. For each \( \epsilon_2 > 0 \), the following holds.

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ Q_{\delta, 1}^{\xi^\lambda} (\infty) > q_1^\lambda (1 - \epsilon_2) \right\} \leq \alpha_1 + \epsilon_2,
\]

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ Q_{\delta, 2}^{\xi^\lambda} (\infty) > q_2^\lambda (1 - \epsilon_2) \right\} \leq \alpha_2 + \epsilon_2,
\]

\[
\limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \mathbb{P} \left\{ Q_{\delta, 1}^{\xi^\lambda} (\infty) / q_1^\lambda \leq Q_{\delta, 2}^{\xi^\lambda} (\infty) / q_2^\lambda - \epsilon_2 \right\} \geq \beta - \epsilon_2.
\]

Lemma 5.1 proves that replacing the waiting-time constraints with queue-based constraints should yield (asymptotically) identical results in terms of optimality. Thus, we turn to study a queue-based formulation starting with identifying a lower bound. To that end, given \( \epsilon \), consider the problem

\[
\min_{\xi^\lambda} N (\xi^\lambda)
\]

s.t. \( \mathbb{P} \left\{ Q_{\delta, 1}^{\xi^\lambda} (\infty) > q_1^\lambda (1 - \epsilon) \right\} \leq \alpha_1 + \epsilon, \)

\( \mathbb{P} \left\{ Q_{\delta, 2}^{\xi^\lambda} (\infty) > q_2^\lambda (1 - \epsilon) \right\} \leq \alpha_2 + \epsilon, \)

\( \mathbb{P} \left\{ Q_{\delta, 1}^{\xi^\lambda} (\infty) / q_1^\lambda \leq Q_{\delta, 2}^{\xi^\lambda} (\infty) / q_2^\lambda - \epsilon \right\} \geq \beta - \epsilon. \) \tag{11}

We next state our lower bound result. Recall that \( Q^{\xi^\lambda} (\infty) \) is a random variable with the stationary distribution of the queue length in an \( M/M/N \) queue with arrival rate \( \lambda \), service rate \( \mu \) and \( N > \lambda / \mu \) servers. As before, the variable \( Q^{\xi^\lambda} (\infty) \) is the local average on \([0, \delta]\) when the \( M/M/N \) queue is initialized with its steady-state distribution.

Proposition 9 Fix \( \lambda, \delta, \epsilon > 0 \) and let \( \xi^\lambda_{\delta, \epsilon} \) be a feasible solution for (11).

Define,

\[
N^\lambda (\delta, \epsilon) = \min \{ N \in \mathbb{Z}_+ : \mathbb{P} \left\{ Q^{\xi^\lambda_{\delta, \epsilon}}(\infty) \geq q_1^\lambda + q_2^\lambda \right\} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2 + 2\epsilon \}. \tag{12}
\]

Then, \( N^\lambda (\delta, \epsilon) \leq N (\xi^\lambda_{\delta, \epsilon}) \), where \( N (\xi^\lambda_{\delta, \epsilon}) \) is the staffing component of \( \xi^\lambda_{\delta, \epsilon} \).
Note that $N^\lambda_*(\delta, \epsilon)$ is closely related to our proposed staffing component $N^\lambda_*$ in (4). Indeed, $q_1^\lambda + q_2^\lambda = \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda$ and (12) becomes

$$N^\lambda_*(\delta, \epsilon) = \min \{ N \in \mathbb{Z}_+ : \mathbb{P} \left\{ Q_\delta^N,\lambda(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2 + 2\epsilon \}.$$  \hspace{1cm} (13)

The following lemma guarantees that the staffing level $N^\lambda_*$, combined with a feasible policy, is asymptotically optimal.

**Lemma 5.2**

$$\lim_{\epsilon \to 0} \lim_{\delta \to 0} \limsup_{\lambda \to \infty} \frac{N^\lambda_*(\delta, \epsilon) - N^\lambda_*}{\sqrt{\lambda}} \geq 0.$$

### 5.1.2. Asymptotic tracking

We show that the local average queue lengths actually “tracks” the tracking function given by (6). The following lemma is a corollary of a result in Gurvich and Whitt [2009a].

**Lemma 5.3** Let $f$ be a Lipschitz continuous function and let $\{\xi_i^\lambda\}$ be a sequence of staffing-policy pairs where the staffing $N(\xi_i^\lambda)$ follows square-root scaling, i.e., $N(\xi_i^\lambda) = \lambda / \mu + \eta \sqrt{\lambda / \mu} + o(\sqrt{\lambda})$ for some $\eta > 0$ and a tracking policy constructed by $f$ is used. For each $\epsilon > 0$, there exists $\delta' > 0$ such that for all $\delta$ with $0 < \delta \leq \delta'$ the following holds.

$$\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \hat{Q}_i^{\delta,\lambda}(\infty) - \hat{Q}_i^{\xi_i^\lambda}(\infty) \right| > \epsilon \right\} < \epsilon.$$

Notably, Lemma 5.3 requires the smoothness of the tracking functions which is clearly violated by our tracking function $f^*$. Instead, given a piecewise Lipschitz function we will construct two other tracking policies that provide lower and upper bounds for our tracking policy and do satisfy the smoothness requirements.

For the following, fix a piecewise Lipschitz tracking function $f$. Let $\mathcal{D}_f := \{d_1, \ldots, d_n\}$ be the set of discontinuity points of $f$ (recall that this is a finite set). Also, let

$$\mathcal{D}^+_f = \{d_i \in \mathcal{D}_f : f_1(d_i+) - f_1(d_i-) \geq 0\} \quad \text{and} \quad \mathcal{D}^-_f = \{d_i \in \mathcal{D}_f : f_1(d_i+) - f_1(d_i-) < 0\}$$

Given $\zeta < \min \{d_1, d_2 - d_1, \ldots, d_n - d_{n-1}\} / 2$ ($\zeta$ will be adjusted within our proofs), define the following tracking functions $\bar{f}^\zeta$ and $\hat{f}^\zeta$ so that $\bar{f}^\zeta_1(x) \geq f^\zeta_1(x)$ and $\hat{f}^\zeta_1(x) \leq f^\zeta_1(x)$ for all $x \geq 0$:

$$\bar{f}^\zeta_1(x) = \begin{cases} f(d_i - \zeta) \left( \frac{d_i - d_i + \zeta}{\zeta} \right) + f(d_i+) \left( \frac{d_i - x}{\zeta} \right), & d_i - \zeta \leq x \leq d_i, d_i \in \mathcal{D}^+_f, \\ f(d_i - \zeta) \left( \frac{d_i - x}{\zeta} \right) + f(d_i + \zeta) \left( \frac{d_i + \zeta - x}{\zeta} \right), & d_i \leq x \leq d_i + \zeta, d_i \in \mathcal{D}^-_f, \\ f(x), & \text{otherwise}, \end{cases}$$

$$\bar{f}^\zeta_2(x) = \begin{cases} f(d_i - \zeta) \left( \frac{d_i - d_i + \zeta}{\zeta} \right) + f(d_i+) \left( \frac{d_i - x}{\zeta} \right), & d_i - \zeta \leq x \leq d_i, d_i \in \mathcal{D}^+_f, \\ f(d_i - \zeta) \left( \frac{d_i - x}{\zeta} \right) + f(d_i + \zeta) \left( \frac{d_i + \zeta - x}{\zeta} \right), & d_i \leq x \leq d_i + \zeta, d_i \in \mathcal{D}^-_f, \\ f(x), & \text{otherwise}. \end{cases}$$
\[ f^c_\lambda(x) = \begin{cases} 
 f(d_i -) \left( \frac{x - d_i}{\varsigma} \right) + f(d_i +) \left( \frac{d_i + \varsigma - x}{\varsigma} \right), & d_i - \varsigma \leq x \leq d_i, \ d_i \in \mathcal{D}_1, \\
 f(d_i -) \left( \frac{x - d_i}{\varsigma} \right) + f(d_i +) \left( \frac{d_i - x}{\varsigma} \right), & d_i \leq x \leq d_i + \varsigma, \ d_i \in \mathcal{D}_1, \\
 f(x), & \text{otherwise}. 
\]

Note that \( f(d_i -) \) and \( f(d_i +) \) are the left and right limit of \( f \) at \( d_i \) which are well-defined by the fact that there are finite number of discontinuous points. \( \bar{f}^c_\lambda(x) \) and \( \underline{f}^c_\lambda(x) \) are defined by \( x - \bar{f}^c_\lambda(x) \) and \( x - \underline{f}^c_\lambda(x) \). Having defined these functions, we can now prove the following result. Here we use the scaled processes \( \hat{Q}^{\lambda,\lambda}_{\delta,i}(t) = \bar{Q}^{\lambda,\lambda}_{\delta,i}(t)/\sqrt{i}, \ i = 1, 2 \). Below, as before, \( \pi^\ast_\lambda \) is the tracking policy defined with respect to the function \( f^* \) in the \( \lambda \)th system with \( f^* \).

Proposition 10 uses these smoothed functions to show the tracking function works. Lemma 5.4, along with Lemma 5.3, is used in the proof of Proposition 10. Given a tracking function \( f \), we denote by \( \pi_f \) the tracking policy defined with respect to \( f \). Let \( Q^f(t) = (Q^f_1(t), Q^f_2(t)) \) be the queue lengths at time \( t \) if the policy \( \pi_f \) is used.

**Lemma 5.4** Fix \( \lambda, \mu, \) and \( N \). Fix tracking functions \( g = (g_1, g_2) \) and \( h = (h_1, h_2) \) such that \( g_1(x) \leq h_1(x) \) and (consequently) \( g_2(x) \geq h_2(x) \) for all \( x \geq 0 \). Suppose that \( Q^g(0) = Q^h(0) \). There then exists a construction of the sample paths so that the following inequalities hold almost surely,

\[ Q^g_1(t) \leq Q^h_1(t) \quad \text{and} \quad Q^g_2(t) \geq Q^h_2(t), \quad t \geq 0. \tag{14} \]

**Proposition 10** For each \( \lambda \), let \( \xi^\lambda = (N^\lambda, \pi^\ast_\lambda) \) where the sequence \( \{N^\lambda\} \) satisfies \( N^\lambda = \lambda/\mu + \eta\sqrt{\lambda/\mu} + o(\sqrt{\lambda}) \). Then, for any \( \epsilon > 0 \), there exists \( \delta' > 0 \) such that for all \( \delta \in (0, \delta'] \)

\[ \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \hat{Q}^{\lambda,\lambda}_{\delta,i}(\infty) - f^*_i \left( \hat{Q}^{\lambda}_{\delta,i}(\infty) \right) \right| > \epsilon \right\} < \epsilon. \]

**Proof:** From the staffing \( N^\lambda \) and the tracking functions \( \bar{f}^\lambda(x) \) and \( \underline{f}^\lambda(x) \), define the staffing and policy pairs \( \xi^\lambda = (N^\lambda, \pi^\lambda_f) \) and \( \xi^\lambda = (N^\lambda, \pi^\lambda_{\underline{f}}) \). By construction, the tracking function \( \bar{f} \) is Lipschitz continuous. Hence, we can apply Theorem 3.1 of Gurvich and Whitt [2009a] to obtain,

\[ \left( \hat{Q}^{\lambda,\lambda}_{1}(\infty) - \bar{f}^\lambda_1 \left( \hat{Q}^{\lambda}_{1}(\infty) \right) , \hat{Q}^{\lambda,\lambda}_{2}(\infty) - \bar{f}^\lambda_2 \left( \hat{Q}^{\lambda}_{2}(\infty) \right) \right) \Rightarrow (0,0). \]

We then have for each \( \epsilon > 0 \) that

\[ \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \hat{Q}^{\lambda,\lambda}_{1}(\infty) - \bar{f}^\lambda_1 \left( \hat{Q}^{\lambda}_{1}(\infty) \right) \right| > \frac{\epsilon}{3} \right\} < \frac{\epsilon}{6}, \quad \text{for } i = 1, 2. \tag{15} \]
By Lemma 5.3, there exists $\delta' > 0$ such that for all $\delta \in (0, \delta')$,
\[
\limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - \hat{Q}_{i}^{\xi, \lambda} (\infty) \right| > \frac{\epsilon}{3} \right\} < \frac{\epsilon}{6}. \tag{16}
\]

Define the set
\[
F = \left\{ x \geq 0 : \left| \overline{f}_i (x) - f_i^* (x) \right| > \frac{\epsilon}{3} \right\}.
\]

By the construction of $\overline{f}_i$, for each $\epsilon > 0$, there exists $\varsigma > 0$ such that,
\[
P \left\{ \hat{Q}_\varsigma (\infty) \in F \right\} < \frac{\epsilon}{6}.
\]

Moreover, since $\hat{Q}_\varsigma (\infty)$, the limiting process of $\hat{Q}_\varsigma^{\lambda} (\infty)$, has a density (see Halfin and Whitt [1981]), $P \{ \hat{Q}_\varsigma (\infty) \in \partial F \} = 0$ so that the following holds by Portmanteau Theorem (see Billingsley [1999]),
\[
\lim_{\lambda \to \infty} P \left\{ \hat{Q}_\varsigma^{\lambda} (\infty) \in F \right\} = P \left\{ \hat{Q}_\varsigma (\infty) \in F \right\}.
\]

Then, for each $\epsilon > 0$, there exists $\varsigma > 0$ that satisfies
\[
\limsup_{\lambda \to \infty} P \left\{ \left| \overline{f}_i \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) - f_i^* \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) \right| > \frac{\epsilon}{3} \right\} < \frac{\epsilon}{6}. \tag{17}
\]

Combining (15)-(17), we conclude that for each $\epsilon > 0$, there exists $\varsigma > 0$ and $\delta' > 0$ so that for all $\delta \in (0, \delta')$,
\[
\limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) \right| > \epsilon \right\} < \limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - \hat{Q}_{i}^{\xi, \lambda} (\infty) \right| > \frac{\epsilon}{3} \right\} \\
+ \limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{i}^{\xi, \lambda} (\infty) - \overline{f}_i \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) \right| > \frac{\epsilon}{3} \right\} \\
+ \limsup_{\lambda \to \infty} P \left\{ \left| \overline{f}_i \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) - f_i^* \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) \right| > \frac{\epsilon}{3} \right\} \\
< \frac{\epsilon}{6} + \frac{\epsilon}{6} + \frac{\epsilon}{6} = \frac{\epsilon}{2}.
\]

Similarly for $\xi_i^{\varsigma}$,
\[
\limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_\varsigma^{\lambda} (\infty) \right) \right| > \epsilon \right\} < \epsilon / 2. \tag{18}
\]

Finally, note that by construction $f_i^* (x) \leq f^* (x) \leq \overline{f}_i (x)$ for all $x \geq 0$ so that, by Lemma 5.4
\[
\hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) \leq \hat{Q}_0^{\lambda} (\infty) \leq \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty)
\]
and, in turn,
\[
\limsup_{\lambda \to \infty} P \left\{ \left| \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_0^{\lambda} (\infty) \right) \right| > \epsilon \right\} \leq \limsup_{\lambda \to \infty} P \left\{ \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_0^{\lambda} (\infty) \right) > \epsilon \right\} + \limsup_{\lambda \to \infty} P \left\{ \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_0^{\lambda} (\infty) \right) < - \epsilon \right\} \\
\leq \limsup_{\lambda \to \infty} P \left\{ \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_0^{\lambda} (\infty) \right) > \epsilon \right\} + \limsup_{\lambda \to \infty} P \left\{ \hat{Q}_{\delta, i}^{\xi, \lambda} (\infty) - f_i^* \left( \hat{Q}_0^{\lambda} (\infty) \right) < - \epsilon \right\} \\
\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$
The result for \( i = 2 \) follows from the fact that \( f_2^* = x - f_1^* \). This completes the proof. \( \square \)

Proposition 10 requires that the staffing have the form \( N = \lambda/\mu + \eta \sqrt{\lambda/\mu} + o(\sqrt{\lambda}) \). A priori, it is not guaranteed that optimal solutions to \( \text{TSF} + \text{SLD}(\beta) \) will satisfy this requirement. The following lemma closes this gap.

**Lemma 5.5** Let \( \{ \xi^\lambda \} \) be a sequence of feasible staffing-policy pairs for (3). Then,

\[
\liminf_{\lambda \to \infty} \frac{N(\xi^\lambda) - \lambda/\mu}{\sqrt{\lambda}} > 0, \tag{19}
\]

a steady-state exists for \( X \) and, furthermore,

\[
\limsup_{\lambda \to \infty} \frac{1}{\sqrt{\lambda}} \mathbb{E}[Q_1^{\xi^\lambda}(\infty) + Q_2^{\xi^\lambda}(\infty)] < \infty. \tag{20}
\]

### 5.1.3. Asymptotic feasibility and optimality: Proof of Theorem 3

We start with the asymptotic feasibility. Specifically, we argue that with \( \xi^\lambda \) being our proposed staffing-policy pair, it holds that for all \( \epsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{Q_{\delta,1}^{\xi^\lambda}(\infty)}{\bar{q}_1} > \bar{q}_i (1 - \epsilon) \right\} \leq \alpha_i + \epsilon, \ i = 1, 2,
\]

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{Q_{\delta,2}^{\xi^\lambda}(\infty)}{\bar{q}_2} \leq \frac{Q_{\delta,2}^{\xi^\lambda}(\infty)}{\bar{q}_2} - \epsilon \right\} \geq \beta - \epsilon.
\]

Asymptotic feasibility for the waiting time formulation then follows from Lemma 5.1.

Note the relation that \( \bar{q}_i = a_i w_i \) for \( i = 1, 2 \). The following inclusions follow directly from the structure of \( f^* \):

\[
\left\{ f_1^* \left( \hat{Q}_\Sigma(\infty) \right) - \bar{q}_1 \geq 0 \right\} \subseteq \left\{ \hat{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\}, \quad \left\{ f_2^* \left( \hat{Q}_\Sigma(\infty) \right) - \bar{q}_2 \geq 0 \right\} \subseteq \left\{ \hat{Q}_\Sigma(\infty) \in [T, \infty) \right\},
\]

and

\[
\left\{ \frac{f_1^* \left( \hat{Q}_\Sigma(\infty) \right)}{\bar{q}_1} - \frac{f_2^* \left( \hat{Q}_\Sigma(\infty) \right)}{\bar{q}_2} \geq 0 \right\} \subseteq \left\{ \hat{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\}.
\]

Consequently,

\[
\mathbb{P} \left\{ f_1^* \left( \hat{Q}_\Sigma(\infty) \right) - \bar{q}_1 \geq 0 \right\} \leq \mathbb{P} \left\{ \hat{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\},
\]

\[
\mathbb{P} \left\{ f_2^* \left( \hat{Q}_\Sigma(\infty) \right) - \bar{q}_2 \geq 0 \right\} \leq \mathbb{P} \left\{ \hat{Q}_\Sigma(\infty) \in [T, \infty) \right\},
\]

\[
\mathbb{P} \left\{ \frac{f_1^* \left( \hat{Q}_\Sigma(\infty) \right)}{\bar{q}_1} - \frac{f_2^* \left( \hat{Q}_\Sigma(\infty) \right)}{\bar{q}_2} \geq 0 \right\} \leq \mathbb{P} \left\{ \hat{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\}.
\]

The constant, \( T \), satisfies that

\[
\mathbb{P} \left\{ \hat{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\} \leq \alpha_1,
\]
By Portmanteau Theorem (and using the 0 measure of the discontinuity points), we have
\[
\mathbb{P}\left\{ \bar{Q}_\Sigma(\infty) \in [T, \infty) \right\} \leq \alpha_2, \\
\mathbb{P}\left\{ \bar{Q}_\Sigma(\infty) \in [\bar{q}_1 + \bar{q}_2, T) \right\} \leq 1 - \beta.
\] (21)

Let \( \mathcal{D}_{f^*} \) be the set of discontinuity points of \( f^* \). Since \( \bar{Q}_\Sigma(\infty) \) has a density, \( \mathbb{P}\left\{ \bar{Q}_\Sigma(\infty) \in \mathcal{D}_{f^*} \right\} = 0 \), so that, by continuous mapping theorem (See Theorem 3.4.3. of Whitt [2002]),
\[
\left( \bar{Q}_\Sigma(\infty), f_1^* \left( \bar{Q}_\Sigma(\infty) \right), f_2^* \left( \bar{Q}_\Sigma(\infty) \right) \right) \Rightarrow \left( \bar{Q}_\Sigma(\infty), f_1^* \left( \bar{Q}_\Sigma(\infty) \right), f_2^* \left( \bar{Q}_\Sigma(\infty) \right) \right),
\]

By Portmanteau Theorem (and using the 0 measure of the discontinuity points), we have
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ f_1^* \left( \bar{Q}_\Sigma^\delta(\infty) \right) - \hat{q}_1 \geq 0 \right\} = \mathbb{P}\left\{ f_1^* \left( \bar{Q}_\Sigma(\infty) \right) - \hat{q}_1 \geq 0 \right\} \leq \alpha_i \text{ for } i = 1, 2,
\]
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ \frac{f_1^* \left( \bar{Q}_\Sigma(\infty) \right)}{\bar{q}_1} - \frac{f_2^* \left( \bar{Q}_\Sigma(\infty) \right)}{\bar{q}_2} \geq 0 \right\} = \mathbb{P}\left\{ \frac{f_1^* \left( \bar{Q}_\Sigma(\infty) \right)}{\bar{q}_1} - \frac{f_2^* \left( \bar{Q}_\Sigma(\infty) \right)}{\bar{q}_2} \geq 0 \right\} \leq 1 - \beta.
\]

By Proposition 10, for each \( \epsilon > 0 \) there exist \( \delta' > 0 \) such that for all \( \delta \in (0, \delta'] \) and \( i = 1, 2, \)
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ \left| \bar{Q}_{\delta,i}^{\lambda}(\infty) - f_1^* \left( \bar{Q}_\Sigma(\infty) \right) \right| > \epsilon \min \left\{ \frac{\hat{q}_1}{2}, \frac{\hat{q}_2}{2} \right\} \right\} < \frac{\epsilon}{2},
\]
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ \bar{Q}_{\delta,i}^{\lambda}(\infty) \geq \bar{q}_i (1 - \epsilon) \right\} \leq \limsup_{\lambda \to \infty} \mathbb{P}\left\{ \left| \bar{Q}_{\delta,i}^{\lambda}(\infty) - f_1^* \left( \bar{Q}_\Sigma(\infty) \right) \right| > \epsilon \min \left\{ \frac{\hat{q}_1}{2}, \frac{\hat{q}_2}{2} \right\} \right\} + \limsup_{\lambda \to \infty} \mathbb{P}\left\{ f_1^* \left( \bar{Q}_\Sigma(\infty) \right) \geq \bar{q}_i \right\} \leq \frac{\epsilon}{2} + \alpha_i \leq \alpha_i + \epsilon,
\]
and
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ \frac{\bar{Q}_{\delta,1}^{\lambda}(\infty)}{\bar{q}_1} > \frac{\bar{Q}_{\delta,2}^{\lambda}(\infty)}{\bar{q}_2} - \epsilon \right\} \leq \limsup_{\lambda \to \infty} \mathbb{P}\left\{ \left| \bar{Q}_{\delta,1}^{\lambda}(\infty) - f_1^* \left( \bar{Q}_\Sigma(\infty) \right) \right| > \epsilon \min \left\{ \frac{\hat{q}_1}{2}, \frac{\hat{q}_2}{2} \right\} \right\} \leq \frac{\epsilon}{2} + 1 - \beta + \frac{\epsilon}{2} \leq 1 - \beta + \epsilon.
\]

This concludes the feasibility argument and we turn to optimality. Let \( \hat{\xi}_\lambda \) be another asymptotically feasible sequence. By the definition of asymptotic feasibility, for any \( \epsilon > 0 \) there exists a \( \delta_1 > 0 \) such that for all \( \delta \in (0, \delta_1] \) the following holds
\[
\limsup_{\lambda \to \infty} \mathbb{P}\left\{ \bar{Q}_{\delta,i}^{\lambda}(\infty) \geq \bar{q}_i (1 - \epsilon) \right\} \leq \alpha_i + \epsilon, \ i = 1, 2,
\]
\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{\hat{Q}_{3,1}^{\lambda}(\infty)}{\overline{q}_1} \leq \frac{\hat{Q}_{3,2}^{\lambda}(\infty)}{\overline{q}_2} - \epsilon \right\} \geq \beta - \epsilon.
\]

Thus, given \( \delta \) and \( \epsilon \) we have by Proposition 9 that \( N(\xi^{\lambda}) \geq N_{\epsilon}(\delta, 2\epsilon) \). In particular, given \( \delta, \epsilon > 0 \)

\[
\limsup_{\lambda \to \infty} \frac{N^{\lambda} - N(\xi^{\lambda})}{\sqrt{\lambda}} \leq \limsup_{\lambda \to \infty} \frac{N^{\lambda} - N_{\epsilon}(\delta, 2\epsilon)}{\sqrt{\lambda}} + \limsup_{\lambda \to \infty} \frac{N_{\epsilon}(\delta, 2\epsilon) - N(\xi^{\lambda})}{\sqrt{\lambda}} \leq \limsup_{\lambda \to \infty} \frac{N^{\lambda} - N_{\epsilon}(\delta, 2\epsilon)}{\sqrt{\lambda}}.
\]

Finally, since \( \epsilon, \delta \) are arbitrary, we have by Lemma 5.2 that

\[
\limsup_{\epsilon \to 0} \limsup_{\delta \to 0} \limsup_{\lambda \to \infty} [N^{\lambda} - N(\xi^{\lambda})]^{+} \leq 0,
\]

which, together with the formerly established asymptotic feasibility, establishes the asymptotic optimality of \( \xi^{\lambda} = (N^{\lambda}_*, \pi^{\lambda}_*) \).

\[\square\]

### 5.2. Proofs on the equivalence between perfect SLD and index policies

#### 5.2.1. Discontinuity and non-monotonicity: Proof of Theorem 4

**Proof on non-monotonicity:** Define the following three sets:

\[
A_1 = \left\{ x : f_1(x) \leq \frac{\tilde{q}_1}{\overline{q}_1 + \overline{q}_2} x, x > \tilde{q}_1 + \overline{q}_2 \right\}, \quad A_2 = \left\{ x : \frac{\tilde{q}_1}{\overline{q}_1 + \overline{q}_2} x < f_1(x) < x - \overline{q}_2, x > \tilde{q}_1 + \overline{q}_2 \right\},
\]

\[
A_3 = \left\{ x : f_1(x) \geq x - \overline{q}_2, x > \tilde{q}_1 + \overline{q}_2 \right\}.
\]

Note that these sets constitute a partition of \((\tilde{q}_1 + \overline{q}_2, \infty)\).

The sequence of staffing levels given in (4) is asymptotically optimal. Changing the parameters, we have

\[
N^{\lambda}_* = \min \left\{ N \in \mathbb{Z}_+ : \mathbb{P} \left\{ \hat{Q}^{N^{\lambda}_*}(\infty) \geq \tilde{q}_1 + \overline{q}_2 \right\} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2 \right\}
\]

If this sequence of staffing solutions is used, the following should hold.

\[
\mathbb{P} \left\{ \hat{Q}_2(\infty) \in A_1 \right\} + \mathbb{P} \left\{ \hat{Q}_2(\infty) \in A_2 \right\} + \mathbb{P} \left\{ \hat{Q}_2(\infty) \in A_3 \right\} = \mathbb{P} \left\{ \hat{Q}_2(\infty) \in A_1 \cup A_2 \cup A_3 \right\}
\]

\[
= \mathbb{P} \left\{ \hat{Q}_2(\infty) \in (\tilde{q}_1 + \overline{q}_2, \infty) \right\} = \min \{ \alpha_1, 1 - \beta \} + \alpha_2.
\]

There may exist multiple asymptotically optimal solutions other than the one in (4) but by Definition 3, the difference between the sequence of staffing levels in (4) and other asymptotically optimal one is \( o \left( \sqrt{\lambda} \right) \) as \( \lambda \to \infty \). Then the above equality should hold for all asymptotically optimal solutions.

If \( x \in A_2 \)

\[
f_1(x) > \frac{\tilde{q}_1}{\overline{q}_1 + \overline{q}_2} x > \frac{\tilde{q}_1}{\overline{q}_1 + \overline{q}_2} \left( 1 + \frac{\tilde{q}_1}{\overline{q}_2} \right) \overline{q}_2 = \tilde{q}_1, \quad x - f_1(x) > \overline{q}_2, \quad f_1(x) > \frac{\tilde{q}_1}{\overline{q}_1 + \overline{q}_2} x > \frac{\tilde{q}_1}{\overline{q}_2} (x - f_1(x)).
\]
As in the proof of Proposition 10, if \( f \) is piecewise Lipschitz continuous, \( \tilde{Q}_1(\infty) \) and \( f_1(\tilde{Q}_\Sigma(\infty)) \) follow the same distribution and also \( \tilde{Q}_2(\infty) \) and \( \tilde{Q}_\Sigma(\infty) - f_1(\tilde{Q}_\Sigma(\infty)) \) do so. Hence, none of the inequalities
\[
\tilde{Q}_1(\infty) \leq \tilde{q}_1, \tilde{Q}_2(\infty) \leq \tilde{q}_2, \tilde{Q}_1(\infty) / \tilde{q}_1 \leq \tilde{Q}_2(\infty) / \tilde{q}_2,
\]
hold if \( \tilde{Q}_\Sigma(\infty) \in A_2 \). Likewise, \( \tilde{Q}_2(\infty) > \tilde{q}_2 \) if \( \tilde{Q}_\Sigma(\infty) \in A_1 \). Hence, by the assumed feasibility
\[
P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \cup A_2 \} = P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \} + P \{ \tilde{Q}_\Sigma(\infty) \in A_2 \} \leq \alpha_2.
\]
By a similar argument, \( \tilde{Q}_1(\infty) > \tilde{q}_1 \) and \( \tilde{Q}_1(\infty) / \tilde{q}_1 > \tilde{Q}_2(\infty) / \tilde{q}_2 \) if \( \tilde{Q}_\Sigma(\infty) \in A_3 \), so that
\[
P \{ \tilde{Q}_\Sigma(\infty) \in A_2 \cup A_3 \} = P \{ \tilde{Q}_\Sigma(\infty) \in A_2 \} + P \{ \tilde{Q}_\Sigma(\infty) \in A_3 \} \leq \min \{ \alpha_1, 1 - \beta \}
\]
Combining these, we get
\[
P \{ \tilde{Q}_\Sigma(\infty) \in A_3 \} + P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \} + 2P \{ \tilde{Q}_\Sigma(\infty) \in A_2 \} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2.
\]
But since \( P(A_1) + P(A_2) + P(A_3) = \min \{ \alpha_1, 1 - \beta \} + \alpha_2, \) \( P \{ \tilde{Q}_\Sigma(\infty) \in A_2 \} = 0. \) Since the distribution of \( \tilde{Q}_\Sigma(\infty) \) is continuous on \([0, \infty)\), it must be that the Lebesgue measure of \( A_2 \) is 0.

If \( A_3 = \emptyset \), then \( P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \} = \min \{ \alpha_1, 1 - \beta \} + \alpha_2. \) Since \( \beta < 1 \) and \( \alpha_1 > 0 \) by the assumption, \( P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \} > \alpha_2 \). But \( \tilde{Q}_2(\infty) > \tilde{q}_2 \) at \( A_1 \) and hence \( P \{ \tilde{Q}_\Sigma(\infty) \in A_1 \} \leq \alpha_2 \), which contradicts the previous sentence. If \( A_1 = \emptyset \), then \( P \{ \tilde{Q}_\Sigma(\infty) \in A_3 \} = \min \{ \alpha_1, 1 - \beta \} + \alpha_2. \) Since \( \alpha_2 > 0 \), \( P \{ \tilde{Q}_\Sigma(\infty) \in A_3 \} > \min \{ \alpha_1, 1 - \beta \} \). \( \tilde{Q}_1(\infty) > \tilde{q}_1 \) and \( \tilde{Q}_1(\infty) / \tilde{q}_1 > \tilde{Q}_2(\infty) / \tilde{q}_2 \) at \( A_3 \) and hence \( P \{ \tilde{Q}_\Sigma(\infty) \in A_3 \} \leq \min \{ \alpha_1, 1 - \beta \} \), which also contradicts the previous sentence. In conclusion, none of \( A_1 \) or \( A_3 \) is empty.

Suppose, towards contradiction, that the function \( f \) is monotone. To make \( f_1 \) monotone as the assumption, the elements of \( A_1 \) should precede those of \( A_3 \). But in that case \( f_2 \) should be non-monotone and in any case and hence \( f \) should be non-monotone. Rigorous proof of this fact now follows.

We can pick \( x_3 \in A_3 \) and \( x_1 \in A_1 \). First, suppose that \( x_3 < x_1 \) as in Figure 9(LHS). With this, we will show that we can choose \( x'' \in A_1 \) and \( x'' \in A_3 \) such that \( x'' < x'' \) and \( f_1(x'') > f_1(x'') \). Let \( x' := \sup \{ x : x_3 \leq x \leq x_1, x \in A_3 \} \). Let \( \epsilon \) satisfy \( 0 < \epsilon < (x_3 - \tilde{q}_1 - \tilde{q}_2) \tilde{q}_2/2\tilde{q}_1 \) and \( x'' \) be a point which satisfies both \( \max \{ x' - \epsilon, x_3 \} < x'' \leq x' \) and \( x'' \in A_3 \). The existence of such a point is guaranteed by the definition of \( x' \).

Then we can choose \( x''' \) such that \( x''' \in \min \{ x'' + 2\epsilon, x_1 \} - \epsilon/2, \min \{ x'' + 2\epsilon, x_1 \} \} \) and \( x''' \in A_1 \). If \( x'' + 2\epsilon \geq x_1 \), letting \( x_1 \) be such a choice since \( x_1 \in A_1 \). On the other hand, suppose \( x'' + 2\epsilon < x_1 \). If \( (x'' + 2\epsilon - \epsilon/2, x'' + 2\epsilon) \cap A_3 \neq \emptyset \), this violates the definition of \( x' \) since for \( x''' \in (x'' + 2\epsilon - \epsilon/2, x'' + 2\epsilon) \cap A_3 \), \( x_3 \leq x''' \leq x_1 \) while \( x''' > x' \). Hence, \( (x'' + 2\epsilon - \epsilon/2, x'' + 2\epsilon) \cap A_3 = \emptyset \).
Then, \((x'' + 2\epsilon - \epsilon/2, x'' + 2\epsilon) \subset A_1 \cup A_2\). But we already argued that \(A_2\) has a measure zero, so the open set \((x'' + 2\epsilon - \epsilon/2, x'' + 2\epsilon)\) must include an element of \(A_1\).

Since \(x'' \in A_3\), \(f_1(x'') \geq x'' - \bar{q}_2\), and since \(x'' \in A_1\), we have

\[
\begin{align*}
f_1(x'') &\leq \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x'' < \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} \left(x'' + 2\epsilon\right) \\
&< \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x'' + \frac{x_3 - \bar{q}_1 - \bar{q}_2}{1 + \bar{q}_1/\bar{q}_2} < \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x'' + \frac{x'' - \bar{q}_1 - \bar{q}_2}{1 + \bar{q}_1/\bar{q}_2} \\
&\leq x'' - \bar{q}_2 \leq f_1(x'').
\end{align*}
\]

Hence \(f_1\) must be non-monotone since \(x'' < x'''\) and \(f_1(x'') > f_1(x''')\).

Next, suppose \(x_1 < x_3\), then \(f_2\) has to be non-monotone since

\[
f_2(x_1) = x_1 - f_1(x_1) \geq x_1 - \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x_1 = \frac{\bar{q}_2}{\bar{q}_1 + \bar{q}_2} x_1 \geq \bar{q}_2 \geq x_3 - f_1(x_3) = f_2(x_3).
\]

Hence for \(f_1\) and \(f_2\) to be monotone, either \(A_1 = \emptyset\) or \(A_3 = \emptyset\) should hold.

Proof on discontinuity: To prove the discontinuity of \(f_1\), we define sets \(A'_1\), \(A'_2\) and \(A'_3\).

\[
A'_1 = \left\{(x, y) \in \mathbb{R}^2 : y \leq \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x, x > \bar{q}_1 + \bar{q}_2\right\},
\]

\[
A'_2 = \left\{(x, y) \in \mathbb{R}^2 : \frac{\bar{q}_1}{\bar{q}_1 + \bar{q}_2} x < y < x - \bar{q}_2, x > \bar{q}_1 + \bar{q}_2\right\},
\]

\[
A'_3 = \left\{(x, y) \in \mathbb{R}^2 : y \geq x - \bar{q}_2, x > \bar{q}_1 + \bar{q}_2\right\}
\]

Note that \(A'_1\) and \(A'_3\) are not connected to each other. Define a new function \(g : (\bar{q}_1 + \bar{q}_2, \infty) \to \mathbb{R}^2\) as

\[
g(x) = (x, f_1(x)).
\]
We again define \( h(x) = (x, \bar{f}_1(x)) : (\bar{q}_1 + \bar{q}_2, \infty) \to \mathbb{R}^2 \) which is obtained from \( g \) as in Figure 9(RHS). If \( g(x) = (x, f_1(x)) \in A'_2 \), the following holds.

\[
\bar{q}_1 x/ (\bar{q}_1 + \bar{q}_2) < f_1(x) < x - q_2
\]

For these \( x \)'s, newly define \( \bar{f}_1(x) \) so that \( (x, \bar{f}_1(x)) \notin A'_2 \). If

\[
|f_1(x) - \bar{q}_1 x/ (\bar{q}_1 + \bar{q}_2)| < |f_1(x) - x + \bar{q}_2|,
\]

as \( x' \) in Figure 9(RHS), let \( \bar{f}_1(x) = q_1 x/ (\bar{q}_1 + \bar{q}_2) \). If

\[
|f_1(x) - \bar{q}_1 x/ (\bar{q}_1 + \bar{q}_2)| > |f_1(x) - x + \bar{q}_2|,
\]

as \( x'' \), let \( \bar{f}_1(x) = x - \bar{q}_2 \). Lastly, if

\[
|f_1(x) - \bar{q}_1 x/ (\bar{q}_1 + \bar{q}_2)| = |f_1(x) - x + \bar{q}_2|,
\]

as \( x''' \), also let \( \bar{f}_1(x) = x - \bar{q}_2 \). By these changes, we exclude the case where \( h(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_2 \neq \emptyset \), so

\[
h(\bar{q}_1 + \bar{q}_2, \infty) \subset A'_1 \cup A'_3.
\]

By the previous discussion, \( A_1 \neq \emptyset \) and \( A_3 \neq \emptyset \). If \( x \in A_1, f_1(x) \leq \bar{q}_1 x/ (\bar{q}_1 + \bar{q}_2) \) and \( (x, f_1(x)) \in A'_1 \). Hence \( g(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_1 \neq \emptyset \). Likewise, \( g(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_3 \neq \emptyset \).

\( h \) was obtained by only changing points \( (x, f_1(x)) \in A'_2 \) to \( (x, \bar{f}_1(x)) \notin A'_2 \) and points in \( A'_1 \) and \( A'_3 \) remain the same. Hence

\[
h(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_1 \neq \emptyset, \text{ and } h(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_3 \neq \emptyset.
\]

The sets \( A'_1 \) and \( A'_3 \) are disconnected to each other, so \( h(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_1 \) and \( h(\bar{q}_1 + \bar{q}_2, \infty) \cap A'_3 \) are also disconnected to each other. Since \( (\bar{q}_1 + \bar{q}_2, \infty) \) is a connected set, \( h \), whose domain is \( (\bar{q}_1 + \bar{q}_2, \infty) \) and the image is a disconnected set, cannot be a continuous function by Theorem 4.37 of Apostol and Makai [1974], and there exists a discontinuous point, say \( x' \).

Next step is to show that this \( x' \), which is discontinuity of \( h \), is also a discontinuous point for \( g \). For contrapositive, let’s assume that \( x' \) is a continuous point for \( g \).

First, suppose \( h(x') \neq g(x') \). The only possible case is \( g(x') \in A'_2 \) while \( h(x') \notin A'_2 \). If \( x' \) was a continuous point of \( g \), then for every \( \epsilon > 0 \), there exists \( \delta > 0 \) such that for all \( x'' \in (x' - \delta, x' + \delta) \),

\[
d(g(x'') - g(x')) = \left( (x'' - x')^2 + (g(x'') - g(x'))^2 \right)^{1/2} < \epsilon.
\]

Define \( \epsilon \) to be \( 0 < \epsilon < \min \{x - \bar{q}_2 - f(x'), f(x') - \bar{q}_1 x'/ (\bar{q}_1 + \bar{q}_2) \} / 2 \) and let \( \delta \) satisfies the above inequality for this \( \epsilon \) and also satisfies \( 0 < \delta < \epsilon \). Then, it is easy to see that

\[
g(x' - \delta, x' + \delta) \subset (x' - \epsilon, x' + \epsilon) \times (f_1(x') - \epsilon, f_1(x') + \epsilon) \in A'_2,
\]
which is equivalent to \((x' - \delta, x' + \delta) \in A_2\). But this is impossible since Lebesgue measure of \(A_2\) should be 0. Hence \(x'\) is also a discontinuous point of \(g\).

Next suppose \(h(x') = g(x')\). If \(x'\) is a continuous point of \(g\), then for all \(\epsilon > 0\), there exists \(\delta > 0\) such that for all \(x'' \in (x' - \delta, x' + \delta)\), the following holds.

\[
\delta(g(x'') - g(x')) < \epsilon.
\]

Define \(\epsilon\) to satisfy \(0 < \epsilon < (\bar{q}_2x'/(\bar{q}_1 + \bar{q}_2) - \bar{q}_2)/4\) and let \(\delta\) satisfy the above condition with this \(\epsilon\) and \(0 < \delta < \epsilon\).

If \(g(x') \in A_1\), then for \(x'' \in (x' - \delta, x' + \delta)\), either \(g(x'') \in A_1\) or \(g(x'') \in A_2\). If \(g(x'') \in A_2\), \(f_1(x'')\) is closer to \(\bar{q}_1 x'/ (\bar{q}_1 + \bar{q}_2)\) than \(x' - \bar{q}_2\). On the other hand, if \(g(x') \in A_1\), then for \(x'' \in (x' - \delta, x' + \delta)\), either \(g(x'') \in A_1\) or \(g(x'') \in A_2\). If \(g(x'') \in A_2\), then \(f_1(x'')\) is closer to \(x' - \bar{q}_2\) than \(\bar{q}_1 x'/ (\bar{q}_1 + \bar{q}_2)\).

Hence, even with the change from \(g\) to \(h\), the points which were originally at the neighborhood \(N(g(x'), \epsilon)\) is still in it. Therefore, if \(g\) was continuous at \(x'\), \(h\) is also continuous at \(x'\). By the contrapositive, we can conclude that if \(h\) is discontinuous at \(x'\), \(g\) is also discontinuous at \(x'\).

The remaining thing is to show that \(f_1\) is also discontinuous on \(x'\). If \(f_1\) is continuous on \(x'\), for \(\epsilon > 0\), we can find \(0 < \delta < \epsilon/\sqrt{2}\) such that \(x'' \in (x - \delta, x + \delta)\) implies \(|f_1(x'') - f_1(x')| < \epsilon/\sqrt{2}\). Then \(d(g(x), g(x'')) < \epsilon\) and \(g\) should also be continuous on \(x\). By contrapositive, if \(g\) is discontinuous, \(f_1\) is discontinuous. By \(f_2(x) = x - f_1(x)\), \(f_2\) is also discontinuous.

\[
\text{5.2.2. Index rule and tracking policy: Proof of Proposition 6}
\]

For a given monotone tracking function \(f = (f_1, f_2)\), define the generalized inverses of \(f_1\) and \(f_2\) as follows.

\[
g_1(y) := \sup \{ x \geq 0 : f_1(x) \leq y \}, \quad g_2(y) := \inf \{ x \geq 0 : f_2(x) \geq y \}.\]

We will show that the tracking policy by \(f = (f_1, f_2)\) and the index policy by \(g_1\) and \(g_2\) are equivalent. For the index policy, assume that the server chooses class 1 to server if \(g_1(Q_1(t)) \geq g_2(Q_2(t))\) and it chooses class 2, if \(g_1(Q_1(t)) < g_2(Q_2(t))\).

Suppose class 1 is to be served at time \(t\) under the tracking policy \(f\). Then

\[
Q_2(t) \leq f_2(Q_{\Sigma}(t)) \quad \text{and} \quad f_1(Q_{\Sigma}(t)) \leq Q_1(t).
\]

\(g_1\) and \(g_2\) are increasing. Also \(g_i(f_i(x)) = x\) for \(i = 1, 2\). Hence applying \(g_2\) to the first inequality and \(g_1\) to the second one result in

\[
g_2(Q_2(t)) \leq Q_{\Sigma}(t) \quad \text{and} \quad Q_{\Sigma}(t) \leq g_1(Q_1(t)).
\]

Combining those two, we can see class is also served under the index policy by \(g_1\) and \(g_2\). The case where class 2 is to be served can be proved similarly and hence the first statement of the theorem is proved.
Now we will prove the second statement of the theorem. For simplicity, we assume that $g_1$ is right continuous with left limits and $g_2$ is left continuous with right limits.

For a given index functions $g_1$ and $g_2$, define $f_1$, which is a function of total queue length $Q$, as follows.

$$f_1(Q):=\begin{cases} 0, & g_2(Q) < g_1(0), \\ Q, & g_1(Q) < g_2(0), \\ \inf \{x: 0 \leq x \leq Q, [g_1(x-), g_1(x)] \cap [g_2(Q-x), g_2(Q-x+)] \neq \emptyset\}, & \text{otherwise}. \end{cases}$$

We will show that $f_1(Q)$ is a well-defined function and the tracking policy by $f = (f_1, f_2)$ is equivalent to index policy with $g_1$ and $g_2$. We complete the proof by showing that $f = (f_1, f_2)$ is an increasing function.

If $g_2(Q) < g_1(0)$ or $g_1(Q) < g_2(0)$, $f_1(Q)$ is evidently well defined. It remains to show there exists $x$ such that $0 \leq x \leq Q$ and $[g_1(x-), g_1(x)] \cap [g_2(Q-x), g_2(Q-x+)] \neq \emptyset$ when $g_2(Q) \geq g_1(0)$ and $g_1(Q) \geq g_2(0)$.

We show this with the assumption that $g_1$ and $g_2$ have finite number of discontinuous points on $[0, Q]$. Define

$$g(x) := g_1(x) - g_2(Q-x).$$

Let $\{x_1, x_2, ..., x_n\}$ be the discontinuous points of $g(x)$ on $[0, Q]$. Note that $g(x)$ is right continuous with left limits. Since $g_1$ and $g_2$ are increasing,

$$g(x-) \leq g(x).$$

We first prove that

$$\bigcup_{x \in [0,Q]} [g(x-), g(x)],$$

is a connected set. Since $g$ is continuous at $[0, x_1)$,

$$A := \bigcup_{x \in [0, x_1)} [g(x-), g(x)] = \bigcup_{x \in [0, x_1]} g(x),$$

is a connected set, i.e., an interval. We will show that $A \cup B$ is also an interval where

$$B := [g(x_1-), g(x_1)].$$

Let $g'$ be a modified function from $g$ such that $g'(x_1) = g(x_1-)$ and $g'(x) = g(x)$ for all $x \in [0, x_1)$. Then $g'(x)$ is continuous at $x \in [0, x_1]$.

Evidently, $g'(x) \in A$ for all $x \in [0, x_1]$ and $g'(x_1) \in B$. Hence a continuous function $g'(x)$ on $[0, x_1]$ is a path from $g'(0) = g(0) \in A$ to $g'(x_1) \in B$. $A$ and $B$ are both path-connected sets and hence $A \cup B$ is a path-connected set, i.e., an interval. Repeating this argument until $x = Q$, we conclude that
\[ \bigcup_{x \in [0, Q]} [g(x^-), g(x)], \]

is a connected set.

By the conditions \( g_2(Q) \geq g_1(0) \) and \( g_1(Q) \geq g_2(0) \), \( g(0) = g_1(0) - g_2(Q) \leq 0 \) and \( g(Q) = g_1(Q) - g_2(0) \geq 0 \). Therefore, the above interval contains 0 and there exists \( x' \) be \( x \) such that

\[ 0 \in [g(x^-), g(x)]. \]

Then \( g_1(x') \geq g_2(Q - x') \) and \( g_1(x') \leq g_2(Q - x'^+) \) hold and hence

\[ [g_1(x'^-), g_1(x')] \cap [g_2(Q - x'), g_2(Q - x'^+)] \neq \emptyset. \]

\( f_1(Q) \) is defined to be the infimum of such \( x' \) and therefore is well-defined.

Now we show that the tracking policy by \( f = (f_1, f_2) \) is equivalent to the index policy by \( g_1 \) and \( g_2 \).

If \( g_2(Q_1(t) + Q_2(t)) < g_1(0) \), \( f_1(Q_1(t) + Q_2(t)) \) is defined to be 0. Since \( g_1 \) is increasing, \( g_2(Q_1(t) + Q_2(t)) < g_1(0) < g_1(Q_1(t) + Q_2(t)) \). Then, class 2 is to be served under both the tracking policy and the index policy.

If \( g_1(Q_1(t) + Q_2(t)) < g_2(0) \), \( f_1(Q_1(t) + Q_2(t)) \) is defined to be \( Q \) and class 1 is to be served under the both policies.

From now on we check the remaining case when \( g_2(Q_1(t) + Q_2(t)) \geq g_1(0) \) and \( g_1(Q_1(t) + Q_2(t)) \geq g_2(0) \). Here \( f_1(Q_1(t) + Q_2(t)) \) is defined to be

\[ \inf \{ x : 0 \leq x \leq Q, [g_1(x^-), g_1(x)] \cap [g_2(Q - x), g_2(Q - x'^+)] \neq \emptyset \}. \]

Suppose class 1 is to be served at time \( t \) with the index policy. Then

\[ g_1(Q_1(t)) \geq g_2(Q_2(t)). \]

If \( [g_1(Q_1(t^-), g_1(Q_1(t))] \cap [g_2(Q_2(t)), g_2(Q_2(t) +)] \neq \emptyset, f_1(Q_1(t) + Q_2(t)) \) is less or equal to \( Q_1(t) \), since it is defined to be the infimum of \( x \) that satisfies

\[ [g_1(x^-), g_1(x)] \cap [g_2(Q - x), g_2(Q - x'^+)] \neq \emptyset. \]

If \( [g_1(Q_1(t^-), g_1(Q_1(t))] \cap [g_2(Q_2(t)), g_2(Q_2(t) +)] = \emptyset, \max [g_2(Q_2(t)), g_2(Q_2(t) +)] < \min [g_1(Q_1(t^-), g_1(Q_1(t))] \) and hence \( f_1(Q_1(t) + Q_2(t)) \) must be less than \( Q_1(t) \), since \( g_1 \) and \( g_2 \) are increasing. Hence

\[ Q_1(t) \geq f_1(Q_1(t) + Q_2(t)), \]

and class 1 is also served at time \( t \) with the tracking policy.
Suppose class 2 is to be served at time $t$ with the index policy. Then
\[ g_1(Q_1(t)) < g_2(Q_2(t)). \]

Evidently, $[g_1(Q_1(t)-), g_1(Q_1(t))] \cap [g_2(Q_2(t)), g_2(Q_2(t)+)] = \emptyset$ and since $g_1$ and $g_2$ are increasing functions, $f_1(Q_1(t) + Q_2(t))$ is larger than $Q_1(t)$, i.e., class 2 is to be served at time $t$ with the tracking policy.

It remains to show that $f = (f_1, f_2)$ is a monotone function. Suppose, on the contrary, that $f$ is a non-monotone function, i.e., there exist $q'$ and $q''$ such that $q' < q''$ and $f_1(q') > f_1(q'')$ for some $i$. Assume without loss of generality that $i = 1$ ($f_1(q') > f_1(q'')$). Choose a $q'''$ such that $f_1(q''') < q''' < f_1(q')$. If $Q_1(t) = q'''$ and $Q_2(t) = q'' - q'''$, $f_1(Q_1(t) + Q_2(t)) = f_1(q''') < q''' = Q_1(t)$ and hence class 1 customer must be served. If $Q_1(t) = q'''$ and $Q_2(t) = q' - q'''$, $f_1(Q_1(t) + Q_2(t)) = f_1(q') > q''' = Q_1(t)$ and hence class 2 customer must be served.

Under the equivalent $c\mu$ policy by $g_1$ and $g_2$, $g_1(q''') \geq g_2(q'' - q''')$ and $g_1(q''') < g_2(q' - q''')$ must hold. But this contradicts the fact that $g_2$ is a monotone function since $q'' - q''' > q' - q'''$. Therefore, $f = (f_1, f_2)$ must be a monotone function.

5.2.3. Near optimality of monotone solutions: Proof of Proposition 7 Having defined $f^\vartheta$, the proof of this theorem is a straightforward adaptation of the proof of Theorem 3 and, in particular, of the arguments in sections 5.1.2 and 5.1.3. Fix $\vartheta > 0$ and let $\hat{Q}_{\Sigma}^\vartheta(\infty)$ defined, as before, to be the limit of the properly scaled $M/M/N$ steady-state queues when the staffing in the $\lambda^i$ queues is $N^i = \lfloor N^i + \vartheta \sqrt{\lambda} \rfloor$. Asymptotic tracking follows as in Proposition 10 with $f^*$ replaced by $f^\vartheta$ there. Note that
\[ \left\{ f_1^\vartheta \left( \hat{Q}_{\Sigma}^\vartheta(\infty) \right) - \bar{q}_1 \geq 0 \right\} = \emptyset, \quad \left\{ f_2^\vartheta \left( \hat{Q}_{\Sigma}^\vartheta(\infty) \right) - \bar{q}_2 \geq 0 \right\} \subseteq \left\{ \hat{Q}_{\Sigma}^\vartheta(\infty) \in [T^\vartheta, \infty) \right\}, \]
and
\[ \left\{ f_1^\vartheta \left( \hat{Q}_{\Sigma}^\vartheta(\infty) \right) \bar{q}_1 - f_2^\vartheta \left( \hat{Q}_{\Sigma}^\vartheta(\infty) \right) \bar{q}_2 \geq 0 \right\} = \emptyset \]

 Proceeding as in §5.1.3, it follows trivially that the SLD constraint (with $\beta = 1$ here) and the constraint for class 1 are (asymptotically) satisfied and we turn to the class-2 constraint. Recall that with the sequence $\{N^i\}$, it holds that $\mathbb{P}\{\hat{Q}_{\Sigma}(\infty) \in [T, \infty)\} \leq \alpha_2$ (see (21)). Since $\vartheta$ is strictly positive, it then follows (by the explicit expressions for $\hat{Q}_{\Sigma}(\infty)$ and $\hat{Q}_{\Sigma}^\vartheta(\infty)$) that there exists $\varpi > 0$ such that $\mathbb{P}\{\hat{Q}_{\Sigma}^\vartheta(\infty) \in [T, \infty)\} \leq \alpha_2 - \varpi$. We can now choose $\kappa^\vartheta$ sufficiently small (and in turn $T^\vartheta$ close to $T$) such that $\mathbb{P}\{\hat{Q}_{\Sigma}^\vartheta(\infty) \in [T^\vartheta, T]\} \leq \varpi/2$, in which case, we have that $\mathbb{P}\{\hat{Q}_{\Sigma}^\vartheta(\infty) \in [\bar{q}_b^\vartheta, \infty)\} \leq \alpha_2$. From here the proof of asymptotic feasibility is concluded as in §5.1.3. \qed
5.2.4. Perfect SLD and monotone tracking policy: Proof of Proposition 8  Suppose, on the contrary, that $N(\xi_1) > N(\xi_2)$. We will show that the TSF constraints cannot be satisfied with the staffing $N(\xi_2)$ and hence $N(\xi_2) \geq N(\xi_1)$ must hold.

If $N(\xi_1) > N(\xi_2)$, the queue length distribution under the staffing $N(\xi_1)$ is first order stochastically dominated by the one under $N(\xi_2)$, i.e., for any $q > 0$, $\mathbb{P}\left\{\hat{Q}_{\xi_1}^2(\infty) \in [q, \infty)\right\} < \mathbb{P}\left\{\hat{Q}_{\xi_2}^2(\infty) \in [q, \infty)\right\}$. Also by the proof of Theorem 3, \[ \mathbb{P}\left\{\hat{Q}_{\xi_1}^2 \geq \bar{q}_1 + \bar{q}_2\right\} = \alpha_2. \] Then there exists $\epsilon > 0$ such that $\mathbb{P}\left\{\hat{Q}_{\xi_2}^2 \geq \bar{q}_1 + \bar{q}_2 + \epsilon\right\} > \alpha_2$.

Let $f$ be the tracking function for $\xi_2$. Then either $f_1(\bar{q}_1 + \bar{q}_2 + \epsilon) > \bar{q}_1$ or $f_2(\bar{q}_1 + \bar{q}_2 + \epsilon) > \bar{q}_2$ or both. Since $f$ is monotone, either
\[ \mathbb{P}\left\{\hat{Q}_{\xi_2}^2(\infty) > \bar{q}_2\right\} \geq \mathbb{P}\left\{\hat{Q}_{\xi_2}^2(\infty) > \bar{q}_1 + \bar{q}_2 + \epsilon\right\} > \alpha_2. \]
or
\[ \mathbb{P}\left\{\hat{Q}_{\xi_2}^1(\infty) > \bar{q}_1\right\} \geq \mathbb{P}\left\{\hat{Q}_{\xi_2}^1(\infty) > \bar{q}_1 + \bar{q}_2 + \epsilon\right\} > \alpha_2 \geq \alpha_1. \]

Hence at least one of the TSF constraints cannot be met when the staff is less than $N(\xi_1)$.

6. Concluding remarks

The index policy, which is prevalent in call center management, is well-known for being optimal in minimizing convex waiting/holding cost for a given staffing level. But it turns out that the family of index policies is not as wide to cover the optimal policy of staffing minimization problem under TSF constraints. Restricting the policy to be index policies requires strictly more staffing than the one needed to just to satisfy TSF constraints and all the optimal policy for TSF constraint formulation have nonintuitive properties – non-monotone and discontinuous – which index policies do not entail.

Nevertheless, index policies are widely used in call centers that also specify TSF as their performance measure. Besides the ease of implementation, we suggest that an intuitive notion of perfect SLD – a VIP customer waits more than a regular customer who arrives at the same time – justifies the use of index policies along with TSF constraints by showing that restricting the policy to be index policies are shown to be equivalent to imposing the perfect SLD constraint.

As far as we know, this paper is the first paper to study the implications of using index policies in staffing problems. Our results are from a simple model with two customer classes and we chose to study the formulation with TSF measures. We expect one can derive richer implications of index policies by studying more generalized model and also formulations with other QoS measures.

A starting point in process management is to identify a sufficient set of so-called Key Performance Indicators (KPIs). Call center must choose QoS metrics that reliably reflect the call center’s objective. This is a question of formulation choice.
Economic and marketing considerations and customer should influence this fundamental design decision. To make an informed formulation choice, one must understand how different formulations translate into outcomes – these would be the optimal solutions of the staffing and prioritization policies.

Given the complexity of these problems, one may mistakenly assume that certain properties do not need “mathematization” into the formulation as they will be obtained for free. Within this context, one can interpret our result as saying Service-Level-Differentiation imposes a real tradeoff. If required it necessarily increases the staffing cost. If it were not to be included, one must incur a “structural” cost in terms of non-conventional characteristics of the resulting policies.

Making informed formulation choices requires expanding our understanding of such tradeoffs.

Acknowledgements

We thank Yong Pin Zhou and Retsef Levi for helpful comments in early stages of this work. We are grateful to three anonymous reviewers and the associate editor for their comments and suggestions that lead to significant improvements to this work.

References


7. Technical supplement to the paper: Call center staffing: Service-level constraints and index priorities

This supplement has three parts. In the first part, we prove auxiliary lemmas in §5.1. In the second part, more general SLD constraint is discussed and Proposition 9 is proved under this general formulation. In the last part, the remaining results that were stated without proofs are given.

7.1. Proofs of auxiliary lemmas in §5.1

We first provide an explicit construction of the process \( X \). Recall that \( Q_i(t) \) is the number of customers in the class-\( i \) queue at time \( t \) and \( Z(t) \) is the number of customers in service at time \( t \). We number the servers and let \( A^s(t) \) be an \( N \)-dimensional process whose \( l \)th element captures the elapsed service time of the customer in service with server \( l \). This coordinate is set to 0 if server \( l \) is idle at time \( t \). We let \( W_i(t) \) be a vector of dimension of \( Q_i(t) \) whose \( j \)th entry is the accumulated waiting time of the \( j \)th customer in the class-\( i \) queue. Customers are listed in their order of arrival so that the first element of \( W_i(t) \) is the accumulated waiting time of the class-\( i \) customer that arrived first amongst those that are still waiting. An arriving customer is added at the end of the vector with an initial value of 0. We remove the item in the head of the vector when the first customer in the queue is admitted to service.

Admissible policies are assumed to be stationary with respect to the process

\[
X(t) = (Z(t), Q_1(t), Q_2(t), A^s(t), W_1(t), W_2(t)),
\]

i.e., an assignment decision at time \( t \) may use only the information captured by \( X(t) \). This guarantees that \( X(t) \) is a Markov process.

Proof of Lemma 5.1: We first show that if part (1.) holds, part (2.) also holds. This will be proved by showing that for a given \( \epsilon_2 \), we can specify \( \bar{\delta} \) such that for all \( \delta \in (0, \bar{\delta}] \), the inequalities of (2.) hold.

We defined \( \epsilon_1 \) as

\[
\epsilon_1 = \epsilon_2 \min \left\{ \frac{\bar{q}_1}{4}, \frac{\bar{q}_2}{4}, \frac{1}{2} \right\}.
\]

By the statement of (1.), we can specify \( \bar{\delta} \) such that for all \( \delta \in (0, \bar{\delta}] \), the inequalities in (1.) hold by newly defined \( \epsilon_1 \). Let \( \delta \) used from now on be \( \delta \in (0, \bar{\delta}] \).

For arbitrary \( T > 0 \), fix any \( a, b \in [0, T] \) with \( 0 \leq a < b \leq T \). Then the following holds.

\[
\lim_{\lambda \to \infty} \left| \frac{1}{\lambda(b-a)} \int_a^b \hat{W}_i^{\xi,\lambda}_s(s) dA^s_i(s) - \frac{1}{b-a} \int_a^b \hat{Q}_i^{\xi,\lambda}_s(s) ds \right| = 0 \text{ a.s.};
\]

see the proof of Proposition 4.2 in Gurvich and Whitt [2009b]. Specifically, we fix \( a = 0 \) and \( b = \delta \) to obtain

\[
\lim_{\lambda \to \infty} \left| a_i \hat{W}_{\delta,i}^{\xi,\lambda}(0) - \hat{Q}_{\delta,i}^{\xi,\lambda}(0) \right| = 0 \text{ a.s.}
\]
Suppose the initial distribution is the steady-state distribution, which exists for sufficiently large \( \lambda \) by Lemma 5.5. Then the system also follows stationary distribution at \( t = 0 \) and hence

\[
\lim_{\lambda \to \infty} \left| a_i \hat{W}^{\xi,i,\lambda} (\infty) - \hat{Q}^{\xi,i,\lambda} (\infty) \right| = 0, \text{ a.s.} \tag{22}
\]

Then the following holds by (22).

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| a_i \hat{W}^{\xi,i,\lambda} (\infty) - \hat{Q}^{\xi,i,\lambda} (\infty) \right| > \epsilon_2 \min \left\{ \frac{\bar{q}_1}{4}, \frac{\bar{q}_2}{4} \right\} \right\} = 0, \quad i = 1, 2.
\]

Since the inequalities in (1.) hold for arbitrary \( \epsilon_1 > 0 \), we put \( \epsilon_1 \) defined above. Then

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \hat{W}^{\xi,1,\lambda} (\infty) > \bar{w}_1 (1 - \frac{3\epsilon_2}{4}) \right\} \leq \alpha_1 + \epsilon_2,
\]

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \hat{W}^{\xi,2,\lambda} (\infty) > \bar{w}_2 (1 - \frac{3\epsilon_2}{4}) \right\} \leq \alpha_2 + \epsilon_2,
\]

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{\hat{W}^{\xi,1,\lambda} (\infty)}{\bar{w}_1} > \frac{\hat{W}^{\xi,2,\lambda} (\infty)}{\bar{w}_2} - \frac{\epsilon_2}{2} \right\} \leq 1 - \beta + \frac{\epsilon_2}{2}.
\]

Using the above inequalities,

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \hat{Q}^{\xi,1,\lambda} (\infty) > \bar{q}_1 (1 - \epsilon_2) \right\} \leq \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \hat{Q}^{\xi,1,\lambda} (\infty) - a_1 \hat{W}^{\xi,1,\lambda} (\infty) \right| > \frac{\bar{q}_1 \epsilon_2}{4} \right\} + \limsup_{\lambda \to \infty} \mathbb{P} \left\{ a_1 \hat{W}^{\xi,1,\lambda} (\infty) > a_1 \bar{w}_1 (1 - \frac{3\epsilon_2}{4}) \right\} \leq 0 + \alpha_1 + \epsilon_2,
\]

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \hat{Q}^{\xi,2,\lambda} (\infty) > \bar{q}_2 (1 - \epsilon_2) \right\} \leq \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \hat{Q}^{\xi,2,\lambda} (\infty) - a_2 \hat{W}^{\xi,2,\lambda} (\infty) \right| > \frac{\bar{q}_2 \epsilon_2}{4} \right\} + \limsup_{\lambda \to \infty} \mathbb{P} \left\{ a_2 \hat{W}^{\xi,2,\lambda} (\infty) > a_2 \bar{w}_2 (1 - \frac{3\epsilon_2}{4}) \right\} \leq 0 + \alpha_2 + \epsilon_2,
\]

\[
\limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{\hat{Q}^{\xi,1,\lambda} (\infty)}{\bar{q}_1} > \frac{\hat{Q}^{\xi,2,\lambda} (\infty)}{\bar{q}_2} - \epsilon_2 \right\} \leq \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \left| \frac{\hat{Q}^{\xi,1,\lambda} (\infty) / \bar{q}_1 - \hat{W}^{\xi,1,\lambda} (\infty) / \bar{w}_1}{\bar{q}_1} \right| > \frac{\bar{q}_1 \epsilon_2}{4 \bar{q}_1} \right\} + \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{\hat{W}^{\xi,1,\lambda} (\infty) / \bar{w}_1}{\bar{q}_1} > \frac{\hat{W}^{\xi,1,\lambda} (\infty) / \bar{w}_2 - \epsilon_2}{2} \right\} + \limsup_{\lambda \to \infty} \mathbb{P} \left\{ \frac{\hat{W}^{\xi,2,\lambda} (\infty) / \bar{w}_1 - \hat{Q}^{\xi,2,\lambda} (\infty) / \bar{q}_2}{\bar{q}_2} > \frac{\bar{q}_2 \epsilon_2}{4 \bar{q}_2} \right\} \leq 0 + 1 - \beta + \epsilon_2.
\]

Note that \( a_1 \bar{w}_1 = \bar{q}_1 \) and \( a_2 \bar{w}_2 = \bar{q}_2 \).

It was shown that for arbitrary \( \epsilon_2 > 0 \), the inequalities of (2.) hold for \( \delta \in (0, \delta] \) where \( \delta \) is specified for (1.). It shows the statement of (2.) also hold and the first direction of the lemma is proved. The other direction is proved similarly.

The following lemma is the restatement of Theorem 1 of Halfin and Whitt [1981].
Lemma 7.1 Define

\[ N^*_\lambda = \min \{ N \in \mathbb{Z}_+ : \mathbb{P}\{ Q^{N,\lambda}(\infty) \geq \sqrt{\lambda \bar{w}} \} \leq \alpha \} . \]  

(23)

Then

\[ N^*_\lambda = \frac{\lambda}{\mu} + \varphi \sqrt{\frac{\lambda}{\mu} + o(\sqrt{\lambda})} , \]

(24)

where \( \varphi \) is the unique solution to

\[ \left( 1 + \frac{\varphi \Phi(\varphi)}{\phi(\varphi)} \right)^{-1} e^{-\varphi \bar{w}} = \alpha \]

(25)

Proof of Lemma 5.2: Define

\[ N^*_\lambda(0, \epsilon) = \min \{ N \in \mathbb{Z}_+ : \mathbb{P}\{ Q^{N,\lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \} \leq \min \{ \alpha_1, 1 - \beta \} + \alpha_2 + 2\epsilon \} . \]

(26)

By Lemma 7.1, it follows that

\[ N^*_\lambda(0, \epsilon) = \frac{\lambda}{\mu} + \varphi(\epsilon) \sqrt{\frac{\lambda}{\mu} + o(\sqrt{\lambda})} , \]

(27)

where \( \varphi = \varphi(\epsilon) \) is the unique solution to

\[ \left( 1 + \frac{\varphi \Phi(\varphi)}{\phi(\varphi)} \right)^{-1} e^{-\varphi(\alpha_1 \bar{w} + a_2 \bar{w}_2)} = \min \{ \alpha_1, 1 - \beta \} + \alpha_2 + 2\epsilon . \]

(28)

Since the LHS of the equation is a strictly increasing and continuous function of \( \varphi \), \( \varphi(\epsilon) \) is strictly increasing in \( \epsilon \geq 0 \) and continuous so that \( \varphi(\epsilon) \to \varphi^* = \eta^*(\beta) \) where that \( N^* = \lambda/\mu + \eta^*(\beta) \sqrt{\lambda/\mu} + o(\sqrt{\lambda}) \); see (45). It follows, then, that there is a function \( \theta(\epsilon) \) with \( \theta(\epsilon) \to 0 \) as \( \epsilon \to 0 \) and such that

\[ \limsup_{\lambda \to \infty} \left| \frac{N^*_\lambda - N^*_\lambda(0, \epsilon)}{\sqrt{\lambda}} \right| \leq \theta(\epsilon) . \]

(29)

To complete the proof of the lemma we will show that, for all sufficiently small \( \epsilon > 0 \),

\[ \limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \frac{N^*_\lambda(0, \epsilon) - N^*_\lambda(\delta, \epsilon)}{\sqrt{\lambda}} \leq 0 . \]

(30)

Towards contradiction, let \( \epsilon > 0 \) be such that (30) does not hold, i.e., such that

\[ \limsup_{\delta \to 0} \limsup_{\lambda \to \infty} \frac{N^*_\lambda(0, \epsilon) - N^*_\lambda(\delta, \epsilon)}{\sqrt{\lambda}} > \chi , \]

(31)

for some \( \chi > 0 \). There must then exists \( \bar{\delta} > 0 \) such that for all \( \delta \in (0, \bar{\delta}] \) and all sufficiently large \( \lambda \),

\[ N^*_\lambda(\delta, \epsilon) \leq \left[ \frac{\lambda}{\mu} + \left( \varphi(\epsilon) - \frac{\chi}{2} \right) \sqrt{\frac{\lambda}{\mu}} \right] = : \tilde{N}^*_\lambda , \]

(32)
where \( \varphi(\epsilon) \) is as in (27). It is a trivial ordering result now that, since \( \hat{N}^\lambda \geq N_i^\lambda(\delta, \epsilon) \), for all \( \delta \) as above and all sufficiently large \( \lambda \),

\[
\mathbb{P}\left\{ Q_\delta^{N_i^\lambda(\delta, \epsilon) \lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\} \geq \mathbb{P}\left\{ \hat{Q}_\delta^{\hat{N}^\lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\}.
\]

(33)

Let \( \alpha(\varphi) = \left(1 + \frac{\varphi(\epsilon)}{\varphi(\epsilon)}\right)^{-1} \). By Lemma 7.1, (28) and using the fact that \( \alpha(\cdot) \) is decreasing in its argument, we have

\[
\lim_{\lambda \to \infty} \mathbb{P}\left\{ \hat{Q}_\delta^{\hat{N}^\lambda}(\infty) > \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\} = \alpha(\varphi(\epsilon) - \chi/2)e^{-(\varphi(\epsilon) - \chi/2)(\alpha_1 w_1 + \alpha_2 w_2)}
\]

\[
> \min\{\alpha_1, 1 - \beta\} + \alpha_2 + 2\epsilon + \hat{\chi},
\]

(34)

for some \( \hat{\chi} > 0 \). Finally, we claim that,

\[
\lim_{\delta \to 0} \lim_{\lambda \to \infty} \mathbb{P}\left\{ \left| \hat{Q}_\delta^{\hat{N}^\lambda}(\infty) - \hat{Q}_{\delta}^{\hat{N}^\lambda}(\infty) \right| > \hat{\chi}/2 \right\} = 0.
\]

(35)

Together with (33) and (34), this proves that

\[
\lim_{\lambda \to \infty} \mathbb{P}\left\{ Q_\delta^{N_i^\lambda(\delta, \epsilon) \lambda}(\infty) \geq \lambda_1 w_1^\lambda + \lambda_2 w_2^\lambda \right\} > \min\{\alpha_1, 1 - \beta\} + \alpha_2 + 2\epsilon,
\]

contradicting the definition of \( N_i^\lambda(\delta, \epsilon) \) and, in turn, contradicting (31). Proving (35) will allow us to conclude that both (29) and (30) hold, proving the result of the lemma.

To prove (35), note that with \( \hat{N}^\lambda \) as in (32), we have by [Halfin and Whitt, 1981, Theorem 2] that the sequence of processes \( \hat{Q}_\delta^{\hat{N}^\lambda} \) converges to a continuous limit and is, in particular, \( C \)-tight. With \( C \)-tightness the argument to establish (35) mimics closely our proof of Lemma 5.3 and we refer the reader to that proof. This completes the proof of this lemma.

**Proof of Lemma 5.3:** Define,

\[ w(x, \delta, T) = \sup\{ |x(t_1) - x(t_2)| : 0 \leq t_1 < t_2 \leq (t_1 + \delta) \wedge T \}. \]

By Theorem 3.2 of Whitt [2007], if \( \{X^\lambda\} \) is \( C \)-tight, for each \( T > 0 \) and \( \epsilon > 0 \), there exists \( \delta' > 0 \) such that for all \( \delta \) with \( 0 < \delta < \delta' \),

\[ \lim_{\lambda \to \infty} \mathbb{P}\left\{ w(X^\lambda, \delta, T) > \epsilon \right\} < \epsilon. \]

By Theorem 3.1 of Halfin and Whitt [1981], the limit process of normalized total queue, \( \hat{Q}_\Sigma \), is a diffusion process. Then, by Theorem 4.1 of Gurvich and Whitt [2009a], if \( f \) is Lipschitz continuous, \( \left\{ \hat{Q}_i^{\xi_i^{\lambda}} \right\} \) converges to \( f_\lambda(\hat{Q}_\Sigma) \), which is a continuous process. Hence \( \left\{ \hat{Q}_i^{\xi_i^{\lambda}} \right\} \) is \( C \)-tight and the above inequality holds if \( X^\lambda \) is substituted by \( \hat{Q}_i^{\xi_i^{\lambda}} \). Fix \( \epsilon > 0 \), \( T > 0 \) and \( \delta < t \leq T \). Then,

\[
\lim_{\lambda \to \infty} \mathbb{P}\left\{ \left| \hat{Q}_{\delta_i}^{\xi_i^{\lambda}}(t) - \hat{Q}_i^{\xi_i^{\lambda}}(t) \right| > \epsilon \right\} = \lim_{\lambda \to \infty} \mathbb{P}\left\{ \left| \frac{1}{\delta} \int_t^{t+\delta} \hat{Q}_i^{\xi_i^{\lambda}}(s) \, ds - \hat{Q}_i^{\xi_i^{\lambda}}(t) \right| > \epsilon \right\}
\]
realizations of the queue lengths under tracking functions $g$ will show that it holds for assigned the service duration $\theta$. Available servers at this arrival time. The $m$-wise. Upon an arrival of a class-$i$ customer, this customer enters service immediately if there are available servers at this arrival time. The $m$th customer to enter service (from either classes) is assigned the service duration $\theta_n$. Let $Q^g(t) = (Q^g_1(t), Q^g_2(t))$ and $Q^h(t) = (Q^h_1(t), Q^h_2(t))$ be the corresponding realizations. $Q^g(t)$ and $Q^h(t)$ are uniquely determined by the policy and the given sequence of interarrival and service times.

Since the tracking policies are work conserving, the sample paths of $Q^g$ and $Q^h$ share the same event times. That is, there is a sequence $\sigma_1, \sigma_2, \ldots$ such that an event (service completion or customer arrival) occurs in either system if and only if $t \in \{\sigma_1, \sigma_2, \ldots\}$. Moreover, arrivals and departures from the system occur at the same time points so that the total number of customers in queue is the same for both, namely,

$$Q^g_1(t) + Q^g_2(t) = Q^h_1(t) + Q^h_2(t), \quad t \geq 0. \quad (36)$$

However queues of individual classes may differ. We now proceed by induction on the event epoch to establish that the ordering $Q^g_1(t) \leq Q^h_1(t)$ is preserved at all time. Clearly, at time 0, the ordering holds by our assumption that both systems are initialized at the same state. Next, assume that the ordering is preserved up to event epoch $\sigma_{n-1}$ for $n \geq 1$, i.e., $Q^g_1(\sigma_{n-1}) \leq Q^h_1(\sigma_{n-1})$, and we will show that it holds for $\sigma_n$, i.e., $Q^g_1(\sigma_n) \leq Q^h_1(\sigma_n)$.
At time \( \sigma_n \), we either have an arrival to one of the customer classes or we have a service completion. Clearly, the ordering is preserved under customer arrivals. Assume that at time \( \sigma_n \) there is a service completion (and, in turn, a departure from the system). If \( Q_i^g (\sigma_n-) < Q_i^h (\sigma_n-) \), regardless of the customer class that both policies choose to serve, \( Q_i^g (\sigma_n) \leq Q_i^h (\sigma_n) \) holds since only one service completion occurs in one epoch. If \( Q_i^g (\sigma_n-) = Q_i^h (\sigma_n-) > 0 \), then \( g_i (x) \leq h_i (x) \) and (36) lead to,

\[
Q_i^g (\sigma_n-) - g_i (Q_i^g (\sigma_n) + Q_2^g (\sigma_n-)) \geq Q_i^h (\sigma_n-) - h_i (Q_i^g (\sigma_n) + Q_2^g (\sigma_n-)).
\]

and since \( g_2 (x) \geq h_2 (x) \),

\[
Q_2^g (\sigma_n-) - g_2 (Q_2^g (\sigma_n-) + Q_2^g (\sigma_n-)) \leq Q_2^h (\sigma_n-) - h_2 (Q_2^g (\sigma_n-)) + Q_2^g (\sigma_n-)).
\]

Suppose that class 1 was not to be served at \( t \) under \( g \). Then it should be the case that

\[
Q_i^g (\sigma_n-) - g_i (Q_2^g (\sigma_n-)) < Q_2 (\sigma_n-) - g_2 (Q_2^g (\sigma_n-)) ,
\]

by the definition of the tracking function. Combining the three inequalities above,

\[
Q_i^h (\sigma_n-) - h_i (Q_2^h (\sigma_n-)) \leq Q_i^g (\sigma_n-) - g_i (Q_2^g (\sigma_n-))
\]

\[
< Q_2^g (\sigma_n-) - g_2 (Q_2^g (\sigma_n-))
\]

\[
\leq Q_2^h (\sigma_n-) - h_2 (Q_2^g (\sigma_n-)) .
\]

Hence, class 1 would not be admitted to service at time \( \sigma_n \) also by the system using tracking function \( h \). Thus, we conclude that in this case \( \parallel x \parallel = Q_i^g (\sigma_n) = Q_i^h (\sigma_n) \) holds too.

If class 1 was to be served by the system using \( g \), by the induction assumption, \( Q_i^g (\sigma_n) \leq Q_i^h (\sigma_n) \) holds regardless of customer class chosen by the system using \( h \).

Since \( Q_i^g (0) = Q_i^h (0), i = 1, 2 \), it follows that \( Q_i^g (t) \leq Q_i^h (t) \) for all \( t \geq 0 \) and, by (36), that \( Q_2^g (t) \geq Q_2^h (t) \) for all \( t \geq 0 \). \( \square \)

Proof of Lemma 5.5: The proof has two parts. In the first part we prove that, if \( N(\xi) \) is a sequence of feasible staffing-policy pairs, then \( N(\xi) > \lambda/\mu \) for all sufficiently large \( \lambda \). This will guarantee the existence of a steady-state for \( X \) that we will subsequently use to establish (19).

First, observe that \( X_{\Sigma} (t) = Z (t) + Q_1 (t) + Q_2 (t) = 0 \) if and only if \( \| X (t) \| = 0 \). Since \( X_{\Sigma} (t) \) has the law of the total number of customers in a related \( M/M/N \) queue (see Lemma 2.1) the state 0 is positive recurrent provided that \( N > \lambda/\mu \). In turn (see e.g. [Asmussen, 2003, Theorem VI.1.2]) a steady-state distribution exists for \( X \) under any admissible policy provided that \( N > \lambda/\mu \).

It remains to establish that, for sufficiently large \( \lambda \) (what we mean by “sufficiently large” will be explicitly defined within the proof) any feasible solution for (3) must satisfy that \( N > \lambda/\mu \). This
will be established by showing that if \( N \leq \lambda/\mu \) at least one of the TSF constraints in (3) must be violated.

To that end, note first that if \( N \leq \lambda/\mu \) then, for any \( K \) and any initial condition \( X_0 \lambda(0) \),
\[
\mathbb{P}\{Q^\lambda_\Sigma(t) > K \} \rightarrow 1 \text{ as } t \rightarrow \infty,
\]
holds for the transient/null recurrent \( M/M/N \) queue (see e.g. [Asmussen, 2003, Corollary II.4.7]) and in turn for \( Q^\lambda_\Sigma(t) \) by Lemma 2.1. We use (37) in what follows.

Let \( D^\lambda_i(t, t + s) \) be the number of class-\( i \) customers admitted to service from queue \( i \) on the interval \( (t, t + s) \). Under our admissibility assumptions one can construct a Poisson process \( N^\lambda_i \) with rate \( N\mu \) such that \( D^\lambda_i(t, t + s) \leq N^\lambda_i(t + s) - N^\lambda_i(t) \) for all \( t \geq s \geq 0 \). We omit the proof of this standard fact.\(^4\) Given \( \epsilon, \lambda, t, u \) define the event
\[
\Omega^\lambda_i(t, u) := \{ \omega \in \Omega : |A^\lambda_i(t + s) - A^\lambda_i(t) - s\lambda_i| \leq \epsilon\lambda, |N^\lambda_i(t + s) - N^\lambda_i(t) - sN\mu| \leq \epsilon\lambda; i = 1, 2; s \leq u \}.
\]
Due to the properties of the Poisson process the probability of this event does not depend on \( t \). From the functional strong law of large numbers for the Poisson process (see e.g. [Chen and Yao, 2001, Lemma 5.8]) we have for all \( \epsilon, u > 0 \) that
\[
\mathbb{P}\{(\Omega^\lambda_i(t^\lambda, u))^c \} \rightarrow 0, \text{ as } \lambda \rightarrow \infty,
\]
for any non-negative sequence \( t^\lambda \).

Fix \( u = 4\delta, \epsilon = (1/4)(1 - (\alpha_1 + \alpha_2)) \) where \( \alpha_1, \alpha_2 \) are as in the formulation (3). Let \( K^\lambda(\delta, \epsilon) = 2(\lambda\delta + N\mu\delta + \epsilon\lambda) \). Then, fix any \( \lambda \) such that \( \mathbb{P}\{(\Omega^\lambda_i(t^\lambda, u))^c \} \leq \epsilon \) and \( K^\lambda(\delta, \epsilon) \geq 2(\lambda u - \epsilon\lambda)(w^\lambda_1 \lor w^\lambda_2) \) where \( w^\lambda_1, w^\lambda_2 \) are as in (3) (we recall that \( w^\lambda = \hat{w}_i/\sqrt{\lambda} \) this is what is meant by “sufficiently large” in the statement of the lemma. Also, let us assume towards contradiction that the staffing component \( N \) of the feasible solution is such that \( N \leq \lambda/\mu \).

By definition \(-(D^\lambda_i(t) - D^\lambda_i(s)) \leq Q^\lambda_i(t) - Q^\lambda_i(s) \leq A^\lambda_i(t) - A^\lambda_i(s) \). In words, the queue cannot increase by more than the arrivals or decrease by more than the admissions to service. As the increments of \( A^\lambda \) and \( D^\lambda \) are non-negative,
\[
\sup_{0 \leq s \leq \delta} |Q^\lambda_i(t + s) - Q^\lambda_i(t)| \leq A^\lambda_i(t + \delta) - A^\lambda_i(t) + D^\lambda_i(t + \delta) - D^\lambda_i(t) \leq A^\lambda_i(t + \delta) - A^\lambda_i(t) + N^\lambda_i(t + \delta) - N^\lambda_i(t),
\]
for all \( t \geq 0 \). In particular, on \( \Omega^\lambda_i(t, u) \):
\[
\sup_{0 \leq s \leq \delta} |Q^\lambda_i(t + s) - Q^\lambda_i(t)| \leq \lambda\delta + N\mu\delta + 2\epsilon\lambda, \ i = 1, 2
\]
\(^4\) Indeed, the sample path of \( D_i \) can be generated through appropriate thinning of a Poisson process with this rate.
and, in turn,

\[
\sup_{0 \leq s \leq \delta} |Q_{\Sigma}^t(t + s) - Q_{\Sigma}^t(t)| \leq 2(\lambda\delta + N\mu\delta + 2\epsilon\lambda).
\]

Recalling that \(K^\lambda(\delta, \epsilon) = 2(\lambda\delta + N\mu\delta + 2\epsilon\lambda)\) and, using (37), let \(t_0^\lambda\) be such that

\[
\mathbb{P}\{Q_{\Sigma}^t(t) > K^\lambda(\delta, \epsilon)\} \geq 1 - \epsilon,
\]

for all \(t \geq t_0^\lambda\). Note that \(t_0^\lambda\) need not depend on the specific value of \(N\) beyond the fact that \(N \leq \lambda/\mu\). Indeed, due to the well-known monotonicity properties of the \(M/M/N\) queue in the number of servers, if (39) holds for \(N = \lfloor \lambda/\mu \rfloor\) it must hold for all integer \(N \leq \lambda/\mu\). Fixing \(\lambda\) sufficiently large and \(t \geq t_0^\lambda\), we consequently have that

\[
\mathbb{P}\left\{ \|Q_{\Sigma}^t\|_{1,t+\delta}^\lambda \leq K^\lambda(\delta, \epsilon)/2 \right\} \
\leq \mathbb{P}\{Q_{\Sigma}^t(t) \leq K^\lambda(\delta, \epsilon)\} + \mathbb{P}\left\{ \sup_{0 \leq s \leq \delta} |Q_{\Sigma}^t(t + s) - Q_{\Sigma}^t(t)| > K^\lambda(\delta, \epsilon)/2 \right\} \leq 2\epsilon,
\]

where for a process \(x\), we use \(\|x\|_{s,t}^i := \inf_{s \leq \ell \leq t} x(\ell)\). The virtual waiting time for class-\(i\) at time \(s\) is defined

\[
W_{\lambda i}^\lambda(s) = \inf \{ \ell \geq 0 : D_i(s, s + \ell) \geq Q_{\lambda i}^\lambda(s) + 1 \}.
\]

Recalling that \(D_i^\lambda(t, t + s) \leq N_i^\lambda(t + s) - N_i^\lambda(t)\), we have on \(\Omega_i^\lambda(t, u)\),

\[
W_{\lambda i}^\lambda(s) \geq \frac{Q_{\lambda i}^\lambda(s)}{N_i^\lambda(t + \delta) - \lambda_i^\lambda(t)},
\]

for all \(s \in [t, t + u - \delta]\). Finally, on \(\Omega_i^\lambda(t, u)\), \(A_i^\lambda(t + \delta) - A_i^\lambda(t) \geq 1\) for \(i = 1, 2\) so that \(W_{\delta i}^\lambda(t)\) is well defined and

\[
W_{\delta i}^\lambda(t) = \frac{1}{A_i^\lambda(t + \delta) - A_i^\lambda(t)} \int_t^{t+\delta} W_{\lambda i}^\lambda(s) dA_i^\lambda(s) \geq \frac{1}{A_i^\lambda(t + \delta) - A_i^\lambda(t)} \int_t^{t+\delta} \frac{Q_{\lambda i}^\lambda(s)}{N_i^\lambda(t + \delta) - \lambda_i^\lambda(t)} dA_i^\lambda(s),
\]

where the first equality is the definition of \(W_{\delta i}^\lambda(t)\). Using (38), we then have that

\[
W_{\delta i}^\lambda(t) \geq \frac{1}{A_i^\lambda(t + \delta) - A_i^\lambda(t)} \int_t^{t+\delta} \frac{Q_{\lambda i}^\lambda(s)}{N_i^\lambda(t + \delta) - \lambda_i^\lambda(t)} dA_i^\lambda(s) \geq \frac{Q_{\lambda i}^\lambda(t)}{N_i^\lambda(t + \delta) - \lambda_i^\lambda(t)} - (\lambda_i^\lambda - N_i^\lambda(t + \delta) + 2\epsilon\lambda),
\]

on \(\Omega_i^\lambda(t, u)\). Since, by assumption, \(N \leq \lambda/\mu\), we have for all \(t \geq t_0^\lambda\)

\[
\mathbb{P}\{W_{\delta 1}^\lambda(t) \vee W_{\delta 2}^\lambda(t) \geq w_{\lambda 1}^\lambda \vee w_{\lambda 2}^\lambda\} \geq \mathbb{P}\{Q_{\Sigma}^t(t) \geq (N\mu u - \epsilon\lambda)(w_{\lambda 1}^\lambda \vee w_{\lambda 2}^\lambda) + K^\lambda(\delta, \epsilon)/2\}
\]

\[
\geq \mathbb{P}\{Q_{\Sigma}^t(t) \geq (\lambda u - \epsilon\lambda)(w_{\lambda 1}^\lambda \vee w_{\lambda 2}^\lambda) + K^\lambda(\delta, \epsilon)/2\}
\]

\[
\geq \mathbb{P}\{Q_{\Sigma}^t(t) \geq K^\lambda(\delta, \epsilon)\} \geq 1 - 2\epsilon,
\]

where we used the fact that \(K^\lambda(\delta, \epsilon) \geq 2(\lambda u - \epsilon\lambda)(w_{\lambda 1}^\lambda \vee w_{\lambda 2}^\lambda)\). Recalling that \(\epsilon = (1/4)(1 - (\alpha_1 + \alpha_2))\), we conclude that

\[
\lim_{t \to \infty} \inf \mathbb{P}\{W_{\delta 1}^\lambda(t) \geq w_{\lambda 1}^\lambda\} + \mathbb{P}\{W_{\delta 2}^\lambda(t) \geq w_{\lambda 2}^\lambda\} \geq \mathbb{P}\{W_{\delta 1}^\lambda(t) \vee W_{\delta 2}^\lambda(t) \geq w_{\lambda 1}^\lambda \vee w_{\lambda 2}^\lambda\} \geq 1 - 2\epsilon > \alpha_1 + \alpha_2.
\]
so that the TSF constraints are violated contradicting the feasibility of $N\lambda/\mu$. We conclude that, for sufficiently large $\lambda$, any feasible solution must have $N > \lambda/\mu$ and, by the argument at the beginning of this proof that $X$ has a well defined steady-state distribution.

We turn next to establish (19). Assume, towards contradiction, that

$$\liminf_{\lambda \to \infty} \frac{N(\xi^\lambda) - \lambda/\mu}{\sqrt{X}} = 0. \tag{42}$$

As $\{\xi^\lambda\}$ is fixed we omit it from the notation for the remainder of this proof. The proof builds on comparisons to the $M/M/N$ queue. To that end, let $\tilde{W}^\lambda$ be the virtual waiting time in an $M/M/N$ queue with arrival rate $\lambda$, service rate $\mu$ and $N(\xi^\lambda)$ servers in the $\lambda$th system. We suppose that the system is initialized with its steady-state distribution (so that by Lemma 2.1, the total number in system has the distribution of $X_{M/M/N}(\infty)$ at time 0).

We will use the following two facts. First, sample paths can be constructed (for our system and the $M/M/N$ queue) so that, for each $\lambda$, almost surely

$$W^\lambda_1(t) \lor W^\lambda_2(t) \geq \tilde{W}^\lambda(t), \quad t \geq 0, \tag{43}$$

where $W^\lambda_i(t), \ i = 1, 2$ is the class-$i$ virtual waiting time. The $M/M/N$ queue is also initialized with its stationary distribution. The next fact is related to the $M/M/N$ queue. Assume that the $M/M/N$ queue is initialized with its stationary distribution at time $t = 0$, then, for any $T > 0$,

$$\liminf_{\lambda \to \infty} P\{\inf_{0 \leq t \leq T} W^\lambda_1(t) \lor W^\lambda_2(t) > \frac{\kappa}{2\sqrt{\lambda}}\} = 1, \text{ for all } \kappa > 0. \tag{44}$$

We provide the detailed arguments for (43) and (44) at the end of this proof.

Next, fixing $\kappa > 0$, and initializing the system with its stationary distribution,

$$\liminf_{\lambda \to \infty} P\left\{\inf_{0 \leq t \leq T} W^\lambda_1(t) \lor W^\lambda_2(t) > \frac{\kappa}{2\sqrt{\lambda}}\right\} = 1,$n

and in turn, for any $T > 0$,

$$\liminf_{\lambda \to \infty} P\left\{\inf_{0 \leq t \leq T} W^\lambda_{\delta,1}(t) \lor W^\lambda_{\delta,2}(t) > \frac{\kappa}{2\sqrt{\lambda}}\right\} = 1.$$

By stationarity (and fixing $\kappa > 4(\bar{w}_1 + \bar{w}_2)$) we have

$$\limsup_{\lambda \to \infty} P\{W^\lambda_{\delta,1}(\infty) \leq w^\lambda_1, W^\lambda_{\delta,2}(\infty) \leq w^\lambda_2\} \leq \limsup_{\lambda \to \infty} P\{W^\lambda_{\delta,1}(\infty) \lor W^\lambda_{\delta,2}(\infty) \leq w^\lambda_1 + w^\lambda_2\} = 0,$$

so that

$$0 = \limsup_{\lambda \to \infty} P\{W^\lambda_{\delta,1}(\infty) \leq w^\lambda_1, W^\lambda_{\delta,2}(\infty) \leq w^\lambda_2\} \geq 1 - \liminf_{\lambda \to \infty} (P\{W^\lambda_{\delta,1}(\infty) > w^\lambda_1\} + P\{W^\lambda_{\delta,2}(\infty) > w^\lambda_2\}).$$
Since $\alpha_1 + \alpha_2 < 1$, there must exist $i \in \{1, 2\}$ such that

$$\liminf_{\lambda \to \infty} \Pr \{ W^\lambda_{\delta,i}(\infty) > w_i^\lambda \} > \alpha_i,$$

which is a contradiction to the asymptotic feasibility of $\xi^\lambda$.

To conclude the proof it remains to establish (43) and (44) starting with the former. For the $M/M/N$ queue

$$\tilde{W}^\lambda(t) = \inf \{ s \geq 0 : \tilde{D}^\lambda(t, t + s) \geq \tilde{Q}^\lambda(t) + 1 \},$$

where $\tilde{Q}^\lambda(t)$ is the queue length at time $t$ and $\tilde{D}^\lambda(t, t + s)$ is the number of customers entering service on $(t, t + s]$. Using the assumed work conservation, it is straightforward to construct out system together with the $M/M/N$ queue on a common sample space so that, almost surely

$$Q^\lambda_1(t) + Q^\lambda_2(t) = \tilde{Q}^\lambda(t), \ t \geq 0 \text{ and } \tilde{D}^\lambda(t, t + s) = D^\lambda_1(t, t + s) + D^\lambda_2(t, t + s), \ s, t \geq 0.$$  

The simple argument is omitted. Since $D^\lambda_i(t, t + s), \ i = 1, 2$ and $\tilde{D}^\lambda(t, t + s)$ are increasing in $s$ for each $t \geq 0$ and since $Q^\lambda_i(t), \ i = 1, 2$ are positive, it follows that for each $t \geq 0$, $W^\lambda_1(t) \lor W^\lambda_2(t) \geq \tilde{W}^\lambda(t)$ as required.

To prove (44), fix $\epsilon > 0$ and let $N^\lambda = \frac{\lambda}{\mu} + \epsilon \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})$. Let $\hat{W}^\lambda_\epsilon(t)$ be the virtual waiting time in this $M/M/N$ queue. Note that for all sufficiently large $\lambda$, $N^\lambda \geq N(\xi^\lambda)$. It is again a simple comparison result that, initializing both $M/M/N$ queues in stationarity (the original one with $N(\xi^\lambda)$ and the one with $N^\lambda$), sample paths can be constructed so that, for all sufficiently large $\lambda$, almost surely

$$\tilde{W}^\lambda(t) \geq \hat{W}^\lambda_\epsilon(t), \ t \geq 0.$$  

For $\hat{W}^\lambda_\epsilon$ it holds that

$$\sqrt{N^\lambda} \hat{W}^\lambda_\epsilon \Rightarrow \frac{1}{\mu} [\hat{X}_\epsilon]^+, \epsilon > 0$$

where $\hat{X}_\epsilon$ is the Halfin-Whitt diffusion and has, at time $t = 0$, the corresponding steady-state distribution; see e.g. [Gurvich et al., 2008, Corollary B.3].

$$\Pr \left\{ \|[\hat{X}_\epsilon]^+\|_T^\dagger \leq \tilde{\kappa} \right\} \leq \Pr \{ [\hat{X}_\epsilon(0)]^+ \leq 2\tilde{\kappa} \} + \Pr \left\{ \|[\hat{X}_\epsilon]^+\|_T^\dagger \leq \tilde{\kappa} \left| [\hat{X}_\epsilon(0)]^+ \right| \geq 2\tilde{\kappa} \right\}. \tag{48}$$

From the explicit expressions for the steady-state of $\hat{X}_\epsilon(0)$ (see [Halfin and Whitt, 1981, Theorem 1 and Corollary 2]) it easily follows that for each $\eta, \check{\kappa} > 0$ there exists $\epsilon$ such that $\Pr \{ [\hat{X}_\epsilon(0)]^+ \leq 2\tilde{\kappa} \} \leq \eta/2$. Simple direct bounds (the detailed argument is omitted) also show that given $\eta, \check{\kappa} > 0$, there exists $\epsilon > 0$ such that $\Pr \{ \|[\hat{X}_\epsilon]^+\|_T^\dagger \leq \check{\kappa} \left| [\hat{X}_\epsilon(0)]^+ \right| \geq 2\check{\kappa} \} \leq \eta/2$. Overall, for each $\eta, \check{\kappa}, T > 0$, there exists $\epsilon$ such that

$$\Pr \{ \|[\hat{X}_\epsilon]^+\|_T^\dagger \leq \check{\kappa} \} \leq \eta.$$
Since the infimum is a continuous function (see [Whitt, 2002, Section 13.4]) we have that 
$$\sqrt{N^{\lambda}}\|W^\lambda\|_T \Rightarrow \|[X_\epsilon]^+\|_T$$ and in turn that 
$$\liminf_{\lambda \to \infty} P\{\sqrt{N^{\lambda}}\|W^\lambda\|_T > \tilde{\kappa}\} \geq 1 - \eta.$$ As 
$$N^{\lambda}/\lambda \to 1/\mu,$$ setting 
$$\kappa = \tilde{\kappa}/2\sqrt{\mu},$$ we have that 
$$\liminf_{\lambda \to \infty} P\{\sqrt{\lambda}\|W^\lambda\|_T > \kappa\} \geq 1 - \eta.$$ Finally, since this applies 
to any \(\eta, \tilde{\kappa}\) (and in turn to any \(\kappa\)) we may conclude that (44) holds which completes the proof. \(\square\)

7.2. Proofs of remaining results

7.2.1. Convexity: Proof of Theorem 2

By Lemma 7.1, the sequence \(\{N^{\lambda}_\star\}\) satisfies

$$N^{\lambda}_\star = \frac{\lambda}{\mu} + \eta^\star(\beta)\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}),$$

(45)

where (keeping all parameters \(\bar{w}_1, \bar{w}_2, \alpha_1, \alpha_2\) constant) \(\eta^\star = \eta^\star(\beta)\) is the unique solution to

$$\alpha(\eta) \exp(-\eta(a_1\bar{w}_1 + a_2\bar{w}_2)) = \min\{\alpha_1, 1 - \beta\} + \alpha_2,$$

(46)

and

$$\alpha(\eta) = \left(1 + \frac{\eta \Phi(\eta)}{\Phi(\eta)}\right)^{-1}.$$

(47)

For \(1 - \beta \geq \alpha_1\) the function \(\eta^\star(\cdot)\) is constant and the function is obviously continuous at \(\alpha_1 = 1 - \beta\).

Thus, to establish convexity it suffices to consider \(\eta(\beta)\) as the unique solution to

$$\alpha(\eta)e^{-\eta(a_1\bar{w}_1 + a_2\bar{w}_2)} = 1 - \beta + \alpha_2,$$

(48)

where \(\alpha(\cdot)\) is as in (47). Let \(\eta(\beta)\) be a solution to this equation for \(\beta\) given. The function \(\alpha(\cdot)\) is convex decreasing; see Lemma B.1 in Borst et al. [2004]. \(e^{-\eta(a_1\bar{w}_1 + a_2\bar{w}_2)}\) is also convex decreasing in \(\eta\) and hence \(\alpha(\eta)e^{-\eta(a_1\bar{w}_1 + a_2\bar{w}_2)}\) is convex decreasing as a product of two convex decreasing functions.

Define \(\varphi(\eta) := \alpha(\eta)e^{-\eta(a_1\bar{w}_1 + a_2\bar{w}_2)}\).

We re-write (48) as

$$\varphi(\eta(\beta)) = 1 - \beta + \alpha_2.$$

Differentiating both sides with respect to \(\beta\) yields

$$\varphi'(\eta(\beta))\eta'(\beta) = -1,$$

(49)

Since \(\varphi'(\cdot) < 0, \eta'(\cdot) > 0\). Differentiating (49) again on both sides we obtain

$$0 = \varphi''(\eta(\beta))(\eta'(\beta))^2 + \varphi'(\eta(\beta))\eta''(\beta).$$

\(\varphi''(\cdot) > 0\) and \(\varphi(\cdot) < 0\) yields \(\eta''(\beta) > 0\). Hence \(\eta(\cdot)\) is convex increasing. \(\square\)
7.2.2. Supplement to §5.2.2 We prove that $f_1$, defined in §5.2.2, is a well-defined function when $g_1$ and $g_2$ have countably infinite number of points of discontinuity:

Since $g_1$ and $g_2$ are increasing, $g_1(Q) \geq g_1(0)$ and $g_2(Q) \geq g_2(0)$. Hence $\max \{g_1(0), g_2(0)\} \leq \min \{g_1(Q), g_2(Q)\}$ holds under the conditions $g_2(Q) \geq g_1(0)$ and $g_1(Q) \geq g_2(0)$. The following set relations evidently hold.

$$[\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}] \subset [g_1(0), g_1(Q)] \subset \bigcup_{0 \leq z \leq Q} \left( \lim_{y \to x^-} g_1(y), g_1(x) \right),$$

$$[\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}] \subset [g_2(0), g_2(Q)] \subset \bigcup_{0 \leq z \leq Q} \left( g_2(x), \lim_{y \to x^+} g_2(y) \right).$$

By the above, for an arbitrary value $z \in [\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}]$, there exists $x$ such that $z \in [\lim_{y \to x^-} g_1(y), g_1(x)]$. Then the following is well defined.

$$X_1(z) := \left\{ x : z \in \left[ \lim_{y \to x^-} g_1(y), g_1(x) \right] \right\}.$$  

Similar definition can be made on $g_2$ as follows

$$X_2(z) := \left\{ x : z \in \left[ g_2(x), \lim_{y \to x^+} g_2(y) \right] \right\}.$$  

By the monotonicity of $g_1$ and $g_2$, $X_1(z)$’s and $X_2(z)$’s are convex sets.

We will show that $X_1(z)$’s and $X_2(z)$’s are closed intervals. Suppose on the contrary that $X_i(z)$ does not include one of its limit points for some $i$ and $z$. Let $x'$ be a boundary point of $X_i(z)$ which is not included in $X_i(z)$. Then there exist $\{x_n\}$ such that $x_n \to x'$ for $n \to \infty$. Since $g_i$ is increasing and $X_i(z)$ is an open interval, it is impossible for an arbitrary element $x''$ of $X_i(z)$ to have $g_i(x) \neq z$. Hence $g_i(x_n) = z$ for all $n$. This argument holds for all the sequences $\{x_n\}$ such that $x_n \to x'$ and therefore $z \in [\lim_{y \to x'^-} g_i(y), \lim_{y \to x'^+} g_i(y)]$, which contradicts the assumption that $X_i(z)$ is open.

Therefore, $X_1(z)$’s and $X_2(z)$’s are closed intervals and the minimum and the maximum of $X_1(\max \{g_1(0), g_2(0)\})$ and $X_1(\min \{g_1(Q), g_2(Q)\})$ are well-defined.

For all $q$ such that, $\min \{X_1(\max \{g_1(0), g_2(0)\})\} \leq q \leq \max \{X_1(\min \{g_1(Q), g_2(Q)\})\},$

$$g_i(q) \in [\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}],$$

by the monotonicity of $g_1$ and hence

$$[\min \{X_1(\max \{g_1(0), g_2(0)\})\}, \max \{X_1(\min \{g_1(Q), g_2(Q)\})\}] \subset \bigcup_{z \in [\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}]} X_1(z).$$

The monotonicity of $g_1$ also guarantees,

$$\bigcup_{z \in [\max \{g_1(0), g_2(0)\}, \min \{g_1(Q), g_2(Q)\}]} X_1(z) \subset [\min \{X_1(\max \{g_1(0), g_2(0)\})\}, \max \{X_1(\min \{g_1(Q), g_2(Q)\})\}].$$
Therefore

\[ \min \{ X_1 (\max \{ g_1 (0), g_2 (0) \}) \} , \max \{ X_1 (\min \{ g_1 (Q), g_2 (Q) \}) \} \} = \bigcup_{z \in [\max \{ g_1 (0), g_2 (0) \}, \min \{ g_1 (Q), g_2 (Q) \}]} X_1 (z). \]

Likewise

\[ \min \{ X_2 (\max \{ g_1 (0), g_2 (0) \}) \} , \max \{ X_2 (\min \{ g_1 (Q), g_2 (Q) \}) \} \} = \bigcup_{z \in [\max \{ g_1 (0), g_2 (0) \}, \min \{ g_1 (Q), g_2 (Q) \}]} X_2 (z). \]

Suppose \( g_2 (0) \leq g_1 (0) \). Then \( X_1 (\max \{ g_1 (0), g_2 (0) \}) = X_1 (g_1 (0)) \) and hence \( 0 \in X_1 (\max \{ g_1 (0), g_2 (0) \}) \). Also \( X_2 (g_1 (0)) \) should include a point smaller or equal to \( Q \) by the condition \( g_2 (Q) \geq g_1 (0) \) and hence,

\[
\min \{ X_1 (\max \{ g_1 (0), g_2 (0) \}) \} + X_2 (\max \{ g_1 (0), g_2 (0) \}) \leq Q. \tag{50}
\]

Likewise, (50) holds in the case \( g_1 (0) < g_2 (0) \).

Suppose \( g_2 (Q) \leq g_1 (Q) \). Then \( X_1 (\min \{ g_1 (Q), g_2 (Q) \}) = X_1 (g_1 (Q)) \) and \( Q \) is included in \( X_1 (\min \{ g_1 (Q), g_2 (Q) \}) \). \( X_2 (g_1 (Q)) \) should include a point larger or equal to \( 0 \) by the condition \( g_1 (Q) \geq g_2 (0) \) and hence

\[
Q \leq \max \{ X_1 (\min \{ g_1 (Q), g_2 (Q) \}) \} + X_2 (\min \{ g_1 (Q), g_2 (Q) \}). \tag{51}
\]

(51) also holds when \( g_2 (Q) > g_1 (Q) \).

Hence, the following two relations hold.

\[
\min \{ X_1 (\max \{ g_1 (0), g_2 (0) \}) \} + X_2 (\max \{ g_1 (0), g_2 (0) \}) \leq Q, \\
\max \{ X_1 (\min \{ g_1 (Q), g_2 (Q) \}) \} + X_2 (\min \{ g_1 (Q), g_2 (Q) \}) \geq Q. \tag{52}
\]

We will show there exists \( z \) such that \( Q \in X_1 (z) + X_2 (z) \). It is evidently true if either of the inequalities in (52) holds. Hence, we assume

\[
\min \{ X_1 (\max \{ g_1 (0), g_2 (0) \}) \} + X_2 (\max \{ g_1 (0), g_2 (0) \}) < Q, \\
\max \{ X_1 (\min \{ g_1 (Q), g_2 (Q) \}) \} + X_2 (\min \{ g_1 (Q), g_2 (Q) \}) > Q. \tag{53}
\]

We suppose there is no \( z \) that satisfies \( Q \in X_1 (z) + X_2 (z) \) and will show there is a contradiction. By (53), there exist \( z_1 \) and \( z_2 \) such that \( \max \{ X_1 (z_1) + X_2 (z_1) \} < Q \) and \( \min \{ X_1 (z_2) + X_2 (z_2) \} > Q \). Then the followings are well defined

\[
Z_1 := \{ z_1 \in [\max \{ g_1 (0), g_2 (0) \} , \min \{ g_1 (Q), g_2 (Q) \}] : \max \{ X_1 (z_1) + X_2 (z_1) \} < Q \} \\
Z_2 := \{ z_2 \in [\max \{ g_1 (0), g_2 (0) \} , \min \{ g_1 (Q), g_2 (Q) \}] : \min \{ X_1 (z_2) + X_2 (z_2) \} > Q \}.
\]
Since we supposed there is no $z$ such that $Q \in X_1 (z) + X_2 (z)$, all the elements of

$$[\max \{g_1 (0), g_2 (0)\}, \min \{g_1 (Q), g_2 (Q)\}],$$

belong to either $Z_1$ or $Z_2$ and hence

$$[\max \{g_1 (0), g_2 (0)\}, \min \{g_1 (Q), g_2 (Q)\}] = Z_1 \cup Z_2. \quad (54)$$

$g_1$ and $g_2$ are increasing functions and hence $\sup Z_1 \leq \inf Z_2$. If $\sup Z_1 \notin Z_1$ and $\inf Z_2 \notin Z_2$, then $Z_1 \cup Z_2$ is not convex and it contradicts (54). Hence either $\sup Z_1 \in Z_1$ or $\inf Z_2 \in Z_2$ holds.

First assume $\sup Z_1 \in Z_1$. Then,

$$\max \{X_1 (\sup Z_1) + X_2 (\sup Z_1)\} < Q < \min \{X_1 (z_2) + X_2 (z_2)\},$$

for all $z_2 \in Z_2$ and hence,

$$\max \{X_1 (\sup Z_1) + X_2 (\sup Z_1)\} < Q < \lim_{z_2 \to \sup Z_1^+} \min \{X_1 (z_2) + X_2 (z_2)\} \quad (55)$$

Since

$$\bigcup_{z \in Z_1 \cup Z_2} X_i (z) = \bigcup_{z \in [\max \{g_1 (0), g_2 (0)\}, \min \{g_1 (Q), g_2 (Q)\}]} X_i (z)$$

$$= [\min \{X_i (\max \{g_1 (0), g_2 (0)\})\}, \max \{X_i (\min \{g_1 (Q), g_2 (Q)\})\}],$$

for $i = 1, 2$, it should be the case that

$$\max \{X_i (\sup Z_1)\} = \lim_{z_2 \to \sup Z_1^+} \min \{X_i (z_2)\},$$

for $i = 1, 2$ to make $\bigcup_{z \in Z_1 \cup Z_2} X_i (z)$ cover the interval. But this contradicts (55). It can also be shown that $\inf Z_2 \in Z_2$ leads to the same contradiction and we can conclude that there exists $z$ such that $Q \in X_1 (z) + X_2 (z)$.

Then there exist $x$ such that $x \in X_1 (z)$ and $Q - x \in X_2 (z)$. By the definition of $X_1 (z)$ and $X_2 (z)$, $z \in [\lim_{y \to x^{-}} g_1 (y), g_1 (x)] \cap [g_2 (Q - x), \lim_{y \to Q - x^{+}} g_2 (y)]$. Hence the existence of $x \in [0, Q]$ such that,

$$\left[\lim_{y \to x^{-}} g_1 (y), g_1 (x)\right] \cap \left[g_2 (Q - x), \lim_{y \to Q - x^{+}} g_2 (y)\right] \neq \emptyset,$$

is proved. $f_1 (Q)$ is defined to be the infimum of such $x$ and therefore is well-defined.