Dimensioning Large Call Centers

Sem Borst[†,‡], Avi Mandelbaum[§,*] Martin I. Reiman[‡]

[†]CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[‡]Bell Labs, Lucent Technologies
Murray Hill, NJ 07974-0636, USA

[§]Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

May 1, 2002

Abstract

We develop a framework for asymptotic optimization of a queueing system. The motivation is the staffing problem of call centers with 100's of agents (or more). Such a call center is modeled as an M/M/N queue, where the number of agents N is large. Within our framework, we determine the asymptotically optimal staffing level N^* that trades off agents' costs with service quality: the higher the latter, the more expensive is the former. As an alternative to this optimization, we also develop a constraint satisfaction approach where one chooses the least N^* that adheres to a given constraint on waiting cost. Either way, the analysis gives rise to three regimes of operation: quality-driven, where the focus is on service quality; efficiency-driven, which emphasizes agents' costs; and a rationalized regime that balances, and in fact unifies, the other two. Numerical experiments reveal remarkable accuracy of our asymptotic approximations: over a wide range of parameters, from the very small to the extremely large, N^* is *exactly* optimal, or it is accurate to within a single agent. We demonstrate the utility of our approach by revisiting the square-root safety staffing principle, which is a long existing rule-of-thumb for staffing the M/M/N queue. In its simplest form, our rule is as follows: if c is the hourly cost of an agent, and a is the hourly cost of customers' delay, then $N^* = R + y^*(\frac{a}{c})\sqrt{R}$, where R is the offered load, and $y^*(\cdot)$ is a function that is easily computable.

[*]The research of A.M. was partially supported by the ISF (Israeli Science Foundation) grant 388/99–2 and by the Technion funds for the promotion of research and sponsored research.

Contents

1 Introduction

Worldwide, telephone-based services have been expanding dramatically in both volume and scope. This has given rise to a huge growth industry – the (telephone) call center industry. Indeed, some assess [5] that 70% of all customer-business interactions in the U.S. occur in call centers, which employ (conservatively speaking) 3% of the U.S. workforce (about 1.5 million agents). Marketing managers refer to call centers as the modern business frontier, being the focus of Customer Relationship Management (CRM); Operations managers are challenged with the fact that personnel costs, specifically staffing, account for over 65% of the cost of running the typical call center. The trade-off between service quality (marketing) and efficiency (operations) thus naturally arises, and a central goal of ours is to contribute to its understanding.

We argue that call centers typify an emerging business environment in which the traditional quality-efficiency trade-off paradigm could collapse: extremely high levels of both service quality and efficiency can coexist. Consider, for example, a best-practice U.S. sales call center that attends to an average of 15,000 phone callers daily; the average duration of a call is 4 minutes and the variability of calls is significant; agents are highly utilized (over 90%), yet customers essentially never encounter a busy signal, hardly anyone abandons while waiting, the average wait for service is a mere few seconds, and over *half* of the customers are served without delay (that is, they find an idle agent who serves them immediately upon calling). Prerequisites for sustaining such performance, to the best of our judgment, are technology-enabled economies-of-scale and scientifically-based managerial principles and laws. In this paper we develop an analytical framework (Sections 4 and 9) that supports such principles. It is based on asymptotic optimization, which yields insight that does not come out of exact analysis. A convincing example is the *square-root safety staffing principle*, described in Section 2 below. It supports simple useful rules-of-thumb for staffing large call centers, rules that so far have been justified only heuristically. Indeed, rigorous asymptotic justifications of such rules are not common in the Operations Research literature. Hence another goal here is to convince the reader of their benefits.

1.1 Costs, optimization and constraint satisfaction

The cost of staffing is the principal component in the operating expenses of a call center. The staffing level is also the dominant factor to determine service level, as measured in terms of delay statistics: poor service levels incur either opportunity losses due to deteriorating goodwill, or more direct revenue losses in case of abandonment and blocking (busy signals). While the need to balance service quality and staffing cost is universal, the weight placed on each may vary dramatically. In some call centers, providing maximal customer care is the primary drive whereas in others, handling a high traffic volume at minimal cost is the overriding goal. The challenge, so we argue, is to translate such strategically-articulated goals into concrete staffing levels: simply put, *how many agents are to be staffed in order to provide acceptable service quality and operational efficiency?* In this paper we answer this question for the M/M/N (Erlang-C) queue, which is the simplest yet most prevalent model that supports call center staffing. (In future research we are hoping to add crucial features of call centers such as abandonment and retrials [16].)

Within the M/M/N model, we postulate a staffing cost function $F(N)$ for employing N agents. Low costs (small N) give rise to long waits, which we quantify in terms of a delay cost function $D(t)$ for a customer being served after waiting t units of time. When F dominates D (or conversely D dominates F), the least costs are achieved in an *efficiency-driven* (or conversely a *quality-driven*) operation. When F and D are comparable, optimization leads to a *rationalized* operation which, as it turns out, is robust enough to encompass most circumstances. Formally, the three regimes emerge from an asymptotic analysis of the M/M/N queue, as the arrival rate λ, and accordingly the optimal staffing level N^*_λ, both scale up to infinity. We refer to such responsive staffing, in response to increased load, as *dimensioning* the call center, which inspired our title. While the staffing levels that we recommend are only asymptotically optimal, they are nevertheless remarkably accurate – to within a *single* agent in the majority of cases. The asymptotics also provide insight, beyond that of exact analysis, about the dependence of the optimal N^*_λ on λ, F and D.

In industry practice, staffing levels are rarely determined through optimization. (One reason is that there is no standard practice for quantifying waiting costs, let alone abandonment, busy-signal and retrial costs; see [1] for some attempts.) Thus, if not by mere experience-based guessing, common practice seeks the least number of agents N^* that satisfies a given constraint on service level. The latter is expressed in terms of some congestion measure, for example the industry-standard Total Service Factor (TSF) given by

$$TSF = \Pr\{\text{Wait} > T\}, \quad \text{for some} \quad T \geq 0,$$

perhaps combined with 1-800 operating costs. We call this practice *constraint satisfaction*. It is to be contrasted with our previous *optimization* practice, where N^* was determined by cost minimization.

1.2 Introduction to our asymptotic framework

As already mentioned, the justification for our proposed optimal staffing level is based on an asymptotic framework, which we formally develop in Section 4. Its basic idea is as follows. The primitives of our call center model are the arrival rate λ, the number of servers N, and the average service time $1/\mu$. The latter will be fixed throughout our analysis, while $\lambda \uparrow \infty$ is our asymptotic regime, and N is the parameter over which we optimize. Specifically, given the staffing cost function $F(N)$ and the customer's cost of delay $D(t)$, we express the overall cost per unit of time $C(N, \lambda)$ in terms of three entities: staffing costs, waiting costs and the probability that an arriving customer is delayed in queue (Erlang-C formula); see (7). Our goal is to solve the discrete optimization problem that seeks N^*_λ which minimizes $C(N, \lambda)$ and, no less importantly, understand the behavior of N^*_λ for large λ. To this end, we translate the discrete optimization problem into a continuous one that is easier to solve, which is carried out by replacing the three entities above with continuous approximations. The optimal solution for the continuous optimization problem provides an approximately optimal solution to our original discrete problem. The approximation is asymptotically optimal in that, as λ increases indefinitely, the ratio of the overall cost at the approximate staffing level to the cost at the true optimal level (both reduced by the cost of staffing at the least level λ/μ needed to assure stability) converges to unity; see Corollary 4.3, and the discussion following it.

Having set up the framework for asymptotic optimality, we then identify continuous approximations to the three cost entities, doing it separately for each of the rationalized, efficiency- and quality-driven regimes described in the previous subsection (see Sections 6-8, respectively). We then derive similar approximations in the context of constraint satisfaction (Section 9).

Of central importance to our approximation is the asymptotic analysis of Halfin & Whitt [9], especially their approximation to the Erlang-C delay function (Lemma 5.1). It gives rise to a square-root safety staffing principle, which reads roughly as follows: for Erlang-C, staffing levels must always exceed the offered load (λ/μ) to ensure stability; this excess is naturally measured in units of the square-root of the offered load, and our optimization problems in fact search for the optimal number y^* of such units. The value of y^* depends on the operational regime under discussion. For example, y^* in the rationalized regime is a function of the cost data, which is independent of λ. (The special case with linear staffing and waiting costs is presented in the next section.) On the other hand, in the efficiency-driven regime where fewer resources suffice, y^* vanishes as $\lambda \uparrow \infty$. The fundamental law behind the square-root scaling is the central limit theorem – see [18] for further insight.

1.3 Structure of the paper

The next section is devoted to an exposition of the square-root safety staffing principle, followed by a review of the related literature.

In Section 3 we set up our M/M/N model and its cost structure. The framework for asymptotic optimality is introduced in Section 4. Its applications require some special functions which are introduced in Section 5, notably the Halfin-Whitt delay function $P(\cdot)$ [9], plotted in Figure 3. It provides an approximation for the delay probability in the M/M/N queue, N large, which operates in the rationalized regime. (Some useful properties of $P(\cdot)$ and other functions are verified in Appendices A-C.) In Sections 6-8 we analyze the rationalized, efficiency- and quality-driven regimes, under the optimization approach. While the analysis is abstract, each of these sections concludes with examples of specific cost structures, for concreteness. In Section 9 we introduce the constraint satisfaction approach, which gives rise to the same three regimes of operation as optimization.

Section 10 describes numerical experiments that test the accuracy of our asymptotically-supported approximations. As already mentioned, the findings are astounding – rarely do we miss by more than a single agent, as far as optimal staffing levels are concerned. In addition, even though the theory is asymptotic, our approximations are accurate with as few as 3 agents. In order to apply our approximations, guidelines are required for fitting a given call center, represented by its parameters and costs, to one of the three operational regimes. This turns out simpler than expected. Indeed, our numerical experiments, backed up by some theory, clearly establish the robustness of the rationalized approximation, as it covers accurately both the efficiency- and quality-driven regimes. Thus, except for extreme settings, the rationalized approximation is the one to use, as we do in the following section.

We conclude in Section 11 with a few worthy directions for future research.

2 The square-root safety staffing principle

To recapitulate, we determine asymptotically optimal staffing levels in accordance with the relative importance of agents' costs and efficiency versus customers' service quality. The very special case of linear staffing and delay costs (this is Example 6.3) already leads to the (re)discovery, as well as a deeper understanding, of a remarkably robust rule-of-thumb, the *square-root safety staffing* rule. It reads as follows: Suppose that the arrival rate is λ customers per hour, and service rate is μ, which implies that the system's *offered load* is given by $R = \frac{\lambda}{\mu}$; if the staffing cost is \$ c per agent per hour, and waiting cost is \$ a per customer per hour, our recommended number of servers N^* is given by

$$N^* = R + y^*(\frac{a}{c})\sqrt{R}, \tag{1}$$

where the function $y^*(r)$, $r \geq 0$, is plotted in Figures 1 and 2; see (23). In simple words, at least R agents ($\lfloor R \rfloor + 1$ to be exact) are required to guarantee stability; however, safety staffing must be added to the minimum as a protection against stochastic variability. This number of additional agents is proportional to $\sqrt{R}$, and the proportionality coefficient $y^*(\frac{a}{c})$ is determined through the optimization (23), by the relative importance of customers' delay (a) to agents' salary (c).

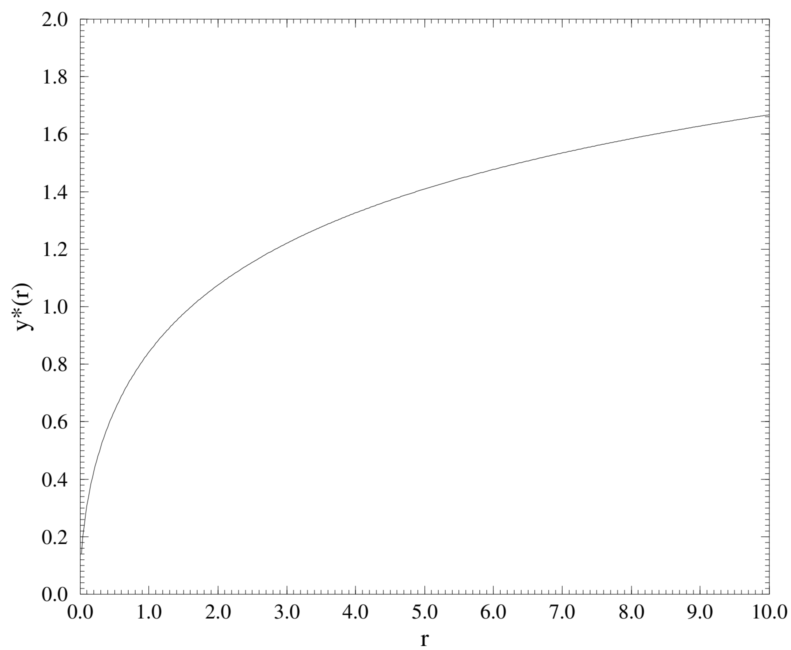

Figure 1: $y^*(r)$ as function of r, $0 \leq r \leq 10$.

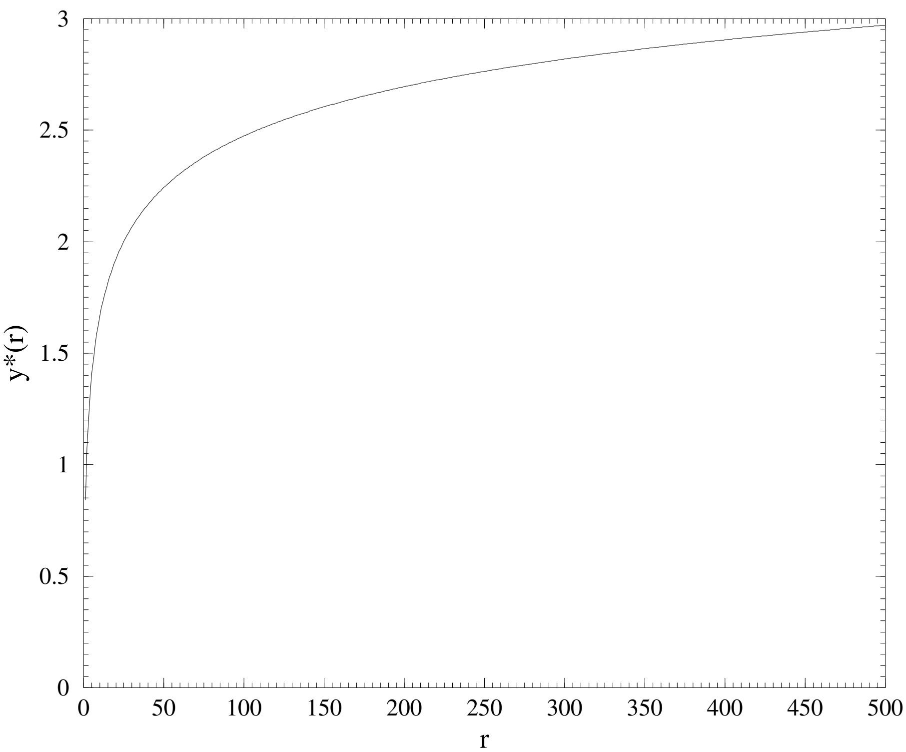

Figure 2: $y^*(r)$ as function of r, $0 \leq r \leq 500$.

Note that the right-hand side of (1) need not be an integer, in which case N^* is obtained by rounding it off. We demonstrate in Section 10, below (40), that this yields the staffing level that minimizes waiting plus staffing costs, exactly in most cases and off by a single agent in the other ones.

The *form* of (1) already carries with it important insight. Let $\Delta = y^*(\frac{a}{c})\sqrt{R}$ denote the *safety staffing* level (the excess number of servers above the minimum $R = \frac{\lambda}{\mu}$). Then, with a and c fixed, an n-fold increase in the offered load R requires that the safety staffing Δ increases by only $\sqrt{n}$-fold, which constitutes significant economies of scale [18].

Now suppose that R is measured in 100's, as it is in large call centers. Then $\sqrt{R}$ is in the low 10's, hence Δ is as well (since y^* grows so slowly: $y^*(100) \approx 2.5$). It follows that the bulk of the agents, namely R, must be present for stability, and only a small fraction $\Delta/R = y^*/\sqrt{R}$ of these must be added as safety against stochastic variability (up to 10%, and in fact significantly less for large call centers, as the practical values of Δ/R indicate). This results in high agent utilization levels R/N^*, around 90% and up. Nevertheless, as shown in Examples 2.1-2.3 below, operational service quality ranges from the acceptable to the extremely high. (Indeed, small changes in Δ, which amounts to small changes in agents' utilization, have noticeable effects on performance.) We are thus operating in a regime where high resource utilization and service level coexist, which is due to economies of scale that dominate stochastic variability. It is important and interesting to note that

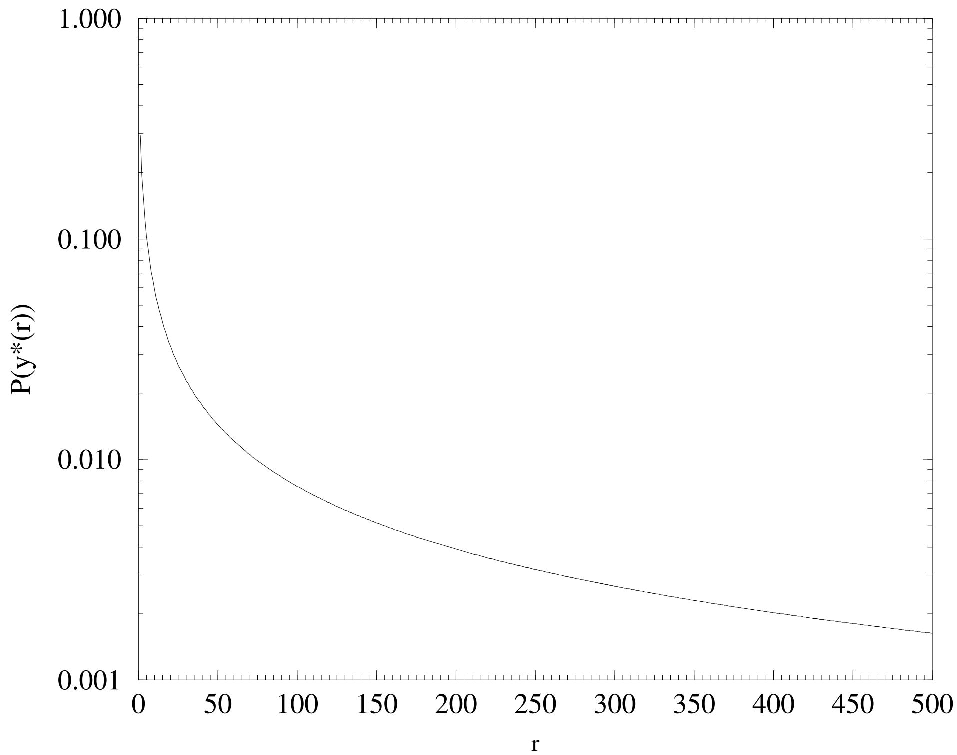

Figure 3: The Halfin-Whitt delay function $P(y)$.

data from large call centers confirms these observations [6].
Small values of r correspond to efficiency-driven staffing. In this range, the function $y^*(\cdot)$ is reasonably approximated by

$$y^*(r) \approx \sqrt{\frac{r}{1 + r(\sqrt{\frac{\pi}{2}} - 1)}}, \quad 0 \leq r < 10.$$

Large values of r correspond to quality-driven staffing. In this range, a close lower bound is $y^*(r) \approx \sqrt{s - \ln s}$, where $s = 2 \ln \frac{r}{\sqrt{2\pi}}$, $r \uparrow \infty$. (See Remark 6.4 for some details on these asymptotic expansions.)
Under our square-root safety staffing, it is anticipated that service level, as expressed by the industry-standard TSF, equals

$$TSF = \Pr\{\text{Wait} > T\} \approx P(y^*) \times \mathrm{e}^{-Ty^*\sqrt{\lambda\mu}}; \quad y^* = y^*(\frac{a}{c});$$

in which

$$P(y^*) = [1 + \frac{y^*\Phi(y^*)}{\phi(y^*)}]^{-1} \approx \Pr\{\text{Wait} > 0\}, \tag{2}$$

is the *Halfin-Whitt delay function* [9] (see Figure 3 and Section 5); $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of the standard normal distribution, respectively. A more management-friendly representation of TSF is

$$TSF = \Pr\{\text{Wait} > T \times \text{E[Service Time]}\} \approx \Pr\{\text{Wait} > 0\} \times \mathrm{e}^{-T\Delta}. \tag{3}$$

Here delay is measured in units of average service time (E[Service Time] $= \frac{1}{\mu}$), and $\Delta = y^*\sqrt{R}$ is the safety staffing level. Another service level standard is the average waiting time, often referred to as Average Speed of Answer (ASA). With N^* as in (1), and again naturally quantified in units of service durations, it is given by

$$\frac{\text{ASA}}{1/\mu} = \frac{\text{E[Wait]}}{\text{E[Service Time]}} \approx \frac{P(y^*)}{\Delta}. \tag{4}$$

The industry standard for measuring operational efficiency is agent utilization, namely $\frac{R}{N}$, which is traded off against service level. Agents are thus idle, or more appropriately described as being available for service, a fraction $\frac{\Delta}{N}$ of their time.

Example 2.1
Consider, for example, the best-practice call center, described in the second paragraph of our Introduction. Assuming 1800 calls per busy hour, the offered load equals $R = 120$. With 90% utilization, one expects that about $N^* \approx 133$ agents share the load ($\Delta = 13$), hence the center operates with $y^* \approx 1.22$. Inverting $y^*(\cdot)$ in Figure 1 shows that, in this call center, an hour wait of customers is valued as 3 times the hourly wage of an agent. With this staffing level, it is expected that about 15% of the customers ($P(1.22) = 0.15$) are delayed; that 5% of the customers are delayed over 20 seconds (using (3) with $T = \frac{1}{12}$); and that, by (4), ASA equals 2.7 seconds (while those who were delayed actually averaged 18 seconds waiting).

□

But the staffing level in the example can be interpreted differently. To this end, recall that the prevalent alternative to the above optimization approach is constraint satisfaction. Specifically, in Example 9.5 it is shown that the least N that guarantees $\Pr\{\text{Wait} > 0\} < \epsilon$ is closely approximated by rounding up

$$N^* = R + P^{-1}(\epsilon)\sqrt{R}, \tag{5}$$

where $P(\cdot)$ is the Halfin-Whitt delay function introduced in (2). Returning to the above best-practice call center, $P^{-1}(\epsilon) = 1.22$ yields, as expected, $\epsilon = 0.15$.

Example 2.2
One should note that a constraint on the fraction of delayed customers is severe, hence it fits call centers that cater to say emergency calls. This can be nicely explained within our framework. For example, requiring that $\epsilon = 0.01$, namely 1 customer out of 100 delayed on average, corresponds to $y^* = P^{-1}(0.01) = 2.38$ (see Figure 3), which *could* be interpreted as saying (via Figure 2) that $\frac{a}{c} = (y^*)^{-1}(2.38) = 75$! An evaluation of customers' time as

TSF with $T = 20$ seconds							
λ (minute^{-1})	$1/\mu$ (minute)	$R = \lambda/\mu$	N^*	y^*	$\lambda/(N^*\mu)$	a/c	TSF
100	4	400	411	0.53	0.97	0.32	0.20
100	4	400	429	1.4	0.93	4.9	0.01
30	4	120	140	1.75	0.86	12.5	0.01
240	0.5	120	126	0.53	0.95	0.32	0.01

Table 1: Example 2.3 - summary of scenarios.

being worth 75-fold of agents' time seems reasonable only under extreme circumstances: for example, if the 'servers' are 'cheap' being say Interactive Voice Response (IVR) units, or customers' time is highly valued as with emergency call centers.

□

Example 2.3
Most call centers define TSF with a positive T, and then requiring $\epsilon = 0.01$ need not be extreme. We now illustrate this by analyzing a prevalent industry standard, which is to aspire that no more than 80% of the callers are delayed over $T = 20$ seconds. Incidentally, we believe that the source of this standard is the familiar 20:80 managerial rule-of-thumb, stating in great generality and vagueness that "only 20% of the reasons already give rise to 80% of the problems". While there is no apparent reason for connecting this rule-of-thumb with any staffing standard, it is nevertheless worthwhile to note that our framework provides some interesting implications for using this rule. This will now be demonstrated via four scenarios which, for convenience, are also summarized in Table 1.
Consider a large call center with $\lambda = 100$ calls per minute, and 4 minutes average call duration. Thus $R = 400$, and adhering to the 20:80 rule implies that $y^* = 0.53$, hence $N^* = 411$. By Figure 1, this translates into $\frac{a}{c} = 0.32$. It follows that, while customers are not highly valued, the 20:80 rule is 'easy' to adhere to because of the call center's size. To wit, increasing N^* to 429 amounts to $y^* = 1.4$, or $\frac{a}{c} = 4.9$, reflecting a significant yet reasonable increase of the relative value of customers' to agents' time. This is accompanied by an increase in server availability (idleness), from 3% to 7%, which enables an order-of-magnitude reduction in TSF, from 0.2 to little less than 0.01: about 1 out of 100 customers is delayed for more than 20 seconds.
To underscore the role of scale in the above scenario, consider a call center with the same offered load parameters as Example 2.1: 30 calls per minute, and again 4 minutes average call duration. Now $R = 120$, but it takes $N^* = 140$ to achieve TSF $= 0.01$, with $T = 20$ seconds. This corresponds to $y^* = 1.75$, or $\frac{a}{c} = 12.5$, a 2.5-fold increase over the large call center. It is interesting to note that with an average call duration of 30 seconds (as in 411 services), with T held at 20 seconds, $N^* = 126$ would suffice, which amounts to $y^* = 0.53$ and $\frac{a}{c} = 0.32$. This is identical to the large call center with the 20:80 rule operation, but the latter accommodates mean service time of 4 minutes, in contrast to the 30 seconds here.

□

The square-root safety staffing principle emerged from the simplest cost structure (linear staffing and waiting costs). While our framework accommodates general costs, the corresponding safety staffing levels are nevertheless *always* proportional to $\sqrt{R}$; it is only the proportionality coefficient that varies with the cost.

2.1 Related literature

The square-root safety staffing principle has been part of the queueing-theory folklore for a long time. This is well documented by Grassmann [7, 8], and recently revisited by Kolesar & Green [14], where both its accuracy and applicability have been convincingly confirmed. The principle was substantiated by Whitt [18], then adapted in Jennings *et.al.* [12] to non-stationary models. All of this work applies infinite-server heuristics, grounded in the fact that the steady-state number of customers in the M/M/∞ queue, say Q^∞, is Poisson distributed with mean $R = \frac{\lambda}{\mu}$. It follows that Q^∞ is approximately normally distributed, with mean R and standard deviation $\sqrt{R}$, when R is not too small. To relate this to staffing in the M/M/N model, one approximates the latter's probability of delay by

$$\Pr\{Q^\infty \geq N\} \approx 1 - \Phi(\frac{N-R}{\sqrt{R}}).$$

Then, the staffing level N^* that guarantees ϵ delay probability is chosen to be

$$N^* = R + \bar{\Phi}^{-1}(\epsilon)\sqrt{R}, \tag{6}$$

where $\bar{\Phi} = 1 - \Phi$.

The square-root principle has two parts to it: first, the conceptual observation that the safety staffing level is proportional to the square-root of the offered load; and second, the explicit calculation of the proportionality coefficient y^*. Our framework accommodates *both* of these two needs, while in all previous works, to the best of our understanding, at least one of them is treated in a heuristic fashion or simply ignored. (We shall be specific shortly.) More important, however, is the fact that our approach and framework allow an arbitrary cost structure, and they have the potential to generalize beyond Erlang-C. For a concrete example, Garnet *et. al.* [6] accommodate impatient customers: in their main result, the square-root rule arises conceptually, but the determination of the value of y^* is left open.

Being specific now, [18] and [12] refer to y^* as a measure of service level, but leave out any explicit calculation of it. Grassmann [8], taking the optimization approach, leads the reader through an instructive progression of increasingly complex staffing models, culminating in his 'equilibrium model' (Erlang-C), for which no 'square-root' justification is provided. (It is justified for his less complex model, under the 'Independence Assumption', but this amounts to using (6).) Some numerical experiments, inspired by [8], are reported at the beginning of Section 10. Finally Kolesar & Green [14] advocate the use of (6), in order to support constraint satisfaction that achieves $\Pr\{\text{Wait} > 0\} \leq \epsilon$. We, on the other hand, recommend the use of (5) for constraint satisfaction, which is proven asymptotically accurate in Example 9.5. The approximations (5) and (6) essentially coincide for small values of ϵ, but (5) is uniformly more accurate. We refer to the beginning of Section 10 for more details.

3 Model description

We consider the classical M/M/N (Erlang-C) model with N servers and infinite-capacity waiting room. Customers arrive as a Poisson process of rate λ, and have independent exponentially distributed service times with mean $1/\mu$. The service rate μ will be arbitrary but fixed, whereas the arrival rate λ will grow large in order to obtain asymptotic scaling results. We assume $\lambda/N\mu < 1$ for stability. Customers are served in order of arrival; then (see for instance [4]) the waiting-time distribution is given by

$$\Pr\{\text{Wait} > t\} = \pi(N, \lambda/\mu)\, \mathrm{e}^{-(N\mu-\lambda)t},$$

where the probability of waiting $\pi(N, \lambda/\mu) = \Pr\{\text{Wait} > 0\}$ is determined by

$$\pi(N,\nu) = \frac{\nu^N}{N!}\left\{(1-\nu/N)\sum_{n=0}^{N-1}\frac{\nu^n}{n!} + \frac{\nu^N}{N!}\right\}^{-1}.$$

We consider the problem of determining the staffing level N that optimally balances staffing cost against quality-of-service. To this end, a staffing cost $F(N)$ per unit of time is associated with staffing N servers. We assume that $F(N)$ is also defined for all non-integer values $N > \lambda/\mu$, and that this extended function $F(\cdot)$ is convex and strictly increasing.
Quality-of-service is quantified in terms of a waiting-cost function $D_\lambda(\cdot)$: a cost $D_\lambda(t)$ is incurred when a customer waits for t time units. (The subscript λ is attached to allow for the possibility that the primitives vary with the arrival intensity.) We assume that $D_\lambda(\cdot)$ is strictly increasing. Without loss of generality, we may take $D_\lambda(0) = 0$. The expected total cost per unit of time is then given by

$$C(N,\lambda) = F(N) + \lambda \mathrm{E}D_\lambda(\text{Wait}) = F(N) + \lambda\pi(N,\lambda/\mu)G(N,\lambda), \tag{7}$$

where

$$G(N,\lambda) = \mathrm{E}D_\lambda(\text{Wait}|\text{Wait} > 0) = (N\mu-\lambda)\int_0^\infty D_\lambda(t)\,\mathrm{e}^{-(N\mu-\lambda)t}dt.$$

Notice that $G(N,\lambda)$ is also defined for all non-integer values $N > \lambda/\mu$. We assume that $D_\lambda(\cdot)$ is such that $G(N,\lambda)$ is finite for all $\lambda/\mu < N$.
We are interested in determining the optimum staffing level

$$N_\lambda^* := \arg\min_{N>\lambda/\mu} C(N,\lambda) \tag{8}$$

(the minimization being over integer values). To see that N_λ^* is well-defined, notice that $\lim_{N\to\infty} F(N) = \infty$, and thus $\lim_{N\to\infty} C(N,\lambda) = \infty$. Hence, $C(N,\lambda)$ indeed achieves a minimum value.

4 Framework for asymptotic optimality

In principle, the optimum staffing level N_λ^* in equation (8) may be obtained through brute-force enumeration. Rather than determining the optimum staffing level numerically, however, we are primarily interested in gaining insight into how N_λ^* grows with the arrival intensity λ, and how it depends on the staffing and waiting cost functions $F(N)$ and $G(N,\lambda)$. In order to do so, we develop an approximate analytical approach for determining the optimum staffing level. As a first step, we translate the discrete optimization problem (8) into a continuous one. The next step is to approximate the latter problem by a related continuous version which is easier to solve. To validate the approach, we then prove that the optimal solution to the approximating continuous problem provides an asymptotically optimal solution to the original discrete problem.

We first transform the discrete optimization problem into a continuous one. Let

$$N_\lambda(x) = \lambda/\mu + x\sqrt{\lambda/\mu},$$

so that the variable $x = (N - \lambda/\mu)/\sqrt{\lambda/\mu}$ is the (normalized) number of servers in excess of the minimum number λ/μ required for stability. In terms of x, we define

$$F_\lambda(x) := F(N_\lambda(x)) - F(\lambda/\mu);$$

$$G_\lambda(x) := \lambda G(N_\lambda(x), \lambda);$$

$$C_\lambda(x) := C(N_\lambda(x), \lambda) - F(\lambda/\mu);$$

$$\pi_\lambda(x) := H(N_\lambda(x), \lambda/\mu),$$

with

$$H(M,\alpha) = \left\{ \alpha \int_0^\infty \mathrm{e}^{-\alpha t} t(1+t)^{M-1} dt \right\}^{-1}.$$

It can be verified ([10], [11]) that $\pi_\lambda(x) = \pi(N_\lambda(x), \lambda/\mu)$ for integer values of $N_\lambda(x)$. The total cost per unit of time (up to the additive constant factor $F(\lambda/\mu)$) can thus be rewritten

$$C_\lambda(x) = F_\lambda(x) + \pi_\lambda(x) G_\lambda(x).$$

Denote

$$x_\lambda^* := \arg\min_{x>0} C_\lambda(x). \tag{9}$$

To see that x_λ^* is well-defined, first notice that the function $C_\lambda(\cdot)$ is strictly convex. This follows from the assumption that $F(\cdot)$ is convex and the fact that $\pi_\lambda(\cdot)$ is convex ([10], [11]) and $G_\lambda(\cdot)$ is strictly convex (Appendix C). In addition, $\lim_{x\downarrow 0} C_\lambda(x) = \infty$, since $\lim_{N\downarrow\lambda/\mu} G(N,\lambda) = \infty$. Also, $\lim_{x\to\infty} C_\lambda(x) = \infty$, because $\lim_{N\to\infty} F(N) = \infty$. Hence, $C_\lambda(\cdot)$ is unimodal, implying that it indeed achieves a unique minimum value at $x_\lambda^* \in (0,\infty)$. Further notice that either

$N^*_\lambda = \lfloor N_\lambda(x^*_\lambda) \rfloor$ or $N^*_\lambda = \lceil N_\lambda(x^*_\lambda) \rceil$, which establishes the link between the discrete problem and the corresponding continuous problem. (Here $\lfloor u \rfloor$ and $\lceil u \rceil$ denote the largest integer smaller than or equal to u, and the smallest integer larger than or equal to u, respectively.) Next, we approximate x^*_λ in (9) by

$$z^*_\lambda := \arg\min_{z>0} C[z; \hat{F}_\lambda, \hat{\pi}_\lambda, \hat{G}_\lambda], \tag{10}$$

where

$$C[z; \hat{F}_\lambda, \hat{\pi}_\lambda, \hat{G}_\lambda] := \hat{F}_\lambda(z) + \hat{\pi}_\lambda(z)\hat{G}_\lambda(z),$$

with the functions $\hat{F}_\lambda(\cdot)$, $\hat{\pi}_\lambda(\cdot)$, $\hat{G}_\lambda(\cdot)$ 'approximating' $F_\lambda(\cdot)$, $\pi_\lambda(\cdot)$, $G_\lambda(\cdot)$, respectively. (Note that with this notation, $x^*_\lambda = \arg\min_{x>0} C[x; F_\lambda, \pi_\lambda, G_\lambda]$.) The approximating functions $\hat{F}_\lambda(\cdot)$, $\hat{G}_\lambda(\cdot)$, and $\hat{\pi}_\lambda(\cdot)$ that we consider will always be such that z^*_λ exists and is unique. If $\hat{F}_\lambda(\cdot)$, $\hat{G}_\lambda(\cdot)$, $\hat{\pi}_\lambda(\cdot)$ have a simple form, then solving for z^*_λ will be easier than determining x^*_λ. At the same time, if $\hat{F}_\lambda(\cdot)$, $\hat{G}_\lambda(\cdot)$, and $\hat{\pi}_\lambda(\cdot)$ approximate $F_\lambda(\cdot)$, $G_\lambda(\cdot)$, and $\pi_\lambda(\cdot)$ well, then it is reasonable to expect that z^*_λ provides a good approximation to x^*_λ and, moreover, $N_\lambda(z^*_\lambda)$ yields a good approximation to N^*_λ.

Before formalizing the above approximation principle, we first introduce the following notational conventions: for any pair of functions a_λ and b_λ (implicitly assuming existence of the limits), denote

$$a_\lambda \overset{\infty}{\approx} b_\lambda : \lim_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = 1; \qquad a_\lambda \overset{\infty}{\approxeq} b_\lambda : \lim_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = \gamma, \qquad 0 < \gamma < \infty;$$

$$a_\lambda \overset{\infty}{\lll} b_\lambda : \lim_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = 0; \qquad a_\lambda \overset{\infty}{\ggg} b_\lambda : \lim_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = \infty;$$

$$a_\lambda \overset{\sup}{\leq} b_\lambda : \limsup_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} \leq 1; \qquad a_\lambda \overset{\inf}{\geq} b_\lambda : \liminf_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} \geq 1;$$

$$a_\lambda \overset{\inf}{<} b_\lambda : \liminf_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} < 1; \qquad a_\lambda \overset{\sup}{>} b_\lambda : \limsup_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} > 1;$$

$$a_\lambda \overset{\inf}{\lll} b_\lambda : \liminf_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = 0; \qquad a_\lambda \overset{\sup}{\ggg} b_\lambda : \limsup_{\lambda\to\infty} \frac{a_\lambda}{b_\lambda} = \infty.$$

Lemma 4.1

Denote $\hat{C}_\lambda(z) = C[z; \hat{F}_\lambda, \hat{\pi}_\lambda, \hat{G}_\lambda]$.
Then $C_\lambda(z^*_\lambda) \overset{\infty}{\approx} C_\lambda(x^*_\lambda)$ if both $C_\lambda(x^*_\lambda) \overset{\infty}{\approx} \hat{C}_\lambda(x^*_\lambda)$ and $C_\lambda(z^*_\lambda) \overset{\infty}{\approx} \hat{C}_\lambda(z^*_\lambda)$.

Proof

By definition of x^*_λ, $C_\lambda(z^*_\lambda) \geq C_\lambda(x^*_\lambda)$, so it suffices to show that $C_\lambda(z^*_\lambda) \overset{\sup}{\leq} C_\lambda(x^*_\lambda)$, which follows directly from

$$C_\lambda(z^*_\lambda) \overset{\infty}{\approx} \hat{C}_\lambda(z^*_\lambda) \leq \hat{C}_\lambda(x^*_\lambda) \overset{\infty}{\approx} C_\lambda(x^*_\lambda).$$

$\square$

Define

$$S_\lambda(x) := \min\{C(\lfloor N_\lambda(x) \rfloor, \lambda), C(\lceil N_\lambda(x) \rceil, \lambda)\}. \tag{11}$$

Lemma 4.2
If $C_\lambda(z^*_\lambda) \overset{\infty}{\approx} C_\lambda(x^*_\lambda)$, *then* $S_\lambda(z^*_\lambda) - F(\lambda/\mu) \overset{\infty}{\approx} C(N^*_\lambda, \lambda) - F(\lambda/\mu)$.

Proof
By definition, $S_\lambda(z^*_\lambda) \geq C(N^*_\lambda, \lambda)$, so it suffices to show that $S_\lambda(z^*_\lambda) - F(\lambda/\mu) \overset{\sup}{\leq} C(N^*_\lambda, \lambda) - F(\lambda/\mu)$.
For fixed λ, we distinguish between four cases.
i. $N^*_\lambda - 1 < N_\lambda(z^*_\lambda) \leq N^*_\lambda$. Then $\lceil N_\lambda(z^*_\lambda) \rceil = N^*_\lambda$, and $S_\lambda(z^*_\lambda) = C(N^*_\lambda, \lambda)$.
ii. $N^*_\lambda \leq N_\lambda(z^*_\lambda) < N^*_\lambda + 1$. Then $\lfloor N_\lambda(z^*_\lambda) \rfloor = N^*_\lambda$, and $S_\lambda(z^*_\lambda) = C(N^*_\lambda, \lambda)$.
iii. $N_\lambda(z^*_\lambda) \leq N^*_\lambda - 1$. Then

$$z^*_\lambda \leq \frac{\lceil N_\lambda(z^*_\lambda) \rceil - \lambda/\mu}{\sqrt{\lambda/\mu}} \leq \frac{N^*_\lambda - 1 - \lambda/\mu}{\sqrt{\lambda/\mu}} \leq \frac{\lfloor N_\lambda(x^*_\lambda) \rfloor - \lambda/\mu}{\sqrt{\lambda/\mu}} \leq x^*_\lambda,$$

so that

$$S_\lambda(z^*_\lambda) - F(\lambda/\mu) \leq C(\lceil N_\lambda(z^*_\lambda) \rceil, \lambda) - F(\lambda/\mu) = C_\lambda(\frac{\lceil N_\lambda(z^*_\lambda) \rceil - \lambda/\mu}{\sqrt{\lambda/\mu}}) \leq C_\lambda(z^*_\lambda)$$

because of the unimodality of $C_\lambda(\cdot)$.
iv. $N_\lambda(z^*_\lambda) \geq N^*_\lambda + 1$. Then

$$z^*_\lambda \geq \frac{\lfloor N_\lambda(z^*_\lambda) \rfloor - \lambda/\mu}{\sqrt{\lambda/\mu}} \geq \frac{N^*_\lambda + 1 - \lambda/\mu}{\sqrt{\lambda/\mu}} \leq \frac{\lceil N_\lambda(x^*_\lambda) \rceil - \lambda/\mu}{\sqrt{\lambda/\mu}} \leq x^*_\lambda,$$

so that

$$S_\lambda(z^*_\lambda) - F(\lambda/\mu) \leq C(\lfloor N_\lambda(z^*_\lambda) \rfloor, \lambda) - F(\lambda/\mu) = C_\lambda(\frac{\lfloor N_\lambda(z^*_\lambda) \rfloor - \lambda/\mu}{\sqrt{\lambda/\mu}}) \leq C_\lambda(z^*_\lambda).$$

Thus, for all λ,

$$\begin{aligned} S_\lambda(z^*_\lambda) - F(\lambda/\mu) &\leq \max\{C(N^*_\lambda, \lambda) - F(\lambda/\mu), C_\lambda(z^*_\lambda)\} \\ &\overset{\infty}{\approx} \max\{C(N^*_\lambda, \lambda) - F(\lambda/\mu), C_\lambda(x^*_\lambda)\} \\ &= C(N^*_\lambda, \lambda) - F(\lambda/\mu). \end{aligned}$$

□

Combining Lemmas 4.1 and 4.2, we obtain the fundamental approximation principle underlying our approach:

Corollary 4.3 *(Asymptotic Optimality)*
Denote $\hat{C}_\lambda(z) = C[z; \hat{F}_\lambda, \hat{\pi}_\lambda, \hat{G}_\lambda]$. *Let* x^*_λ *and* z^*_λ *be as in (9) and (10) respectively. If* $C_\lambda(x^*_\lambda) \overset{\infty}{\approx} \hat{C}_\lambda(x^*_\lambda)$ *and* $C_\lambda(z^*_\lambda) \overset{\infty}{\approx} \hat{C}_\lambda(z^*_\lambda)$, *then the staffing function* z^*_λ *is asymptotically optimal in the sense that, as* $\lambda \to \infty$,

$$S_\lambda(z^*_\lambda) - F(\lambda/\mu) \overset{\infty}{\approx} C(N^*_\lambda, \lambda) - F(\lambda/\mu),$$

with $S_\lambda(x)$ *given in (11).*

Note that the quantities $S_\lambda(z_\lambda^*) - F(\lambda/\mu)$ and $C(N_\lambda^*, \lambda) - F(\lambda/\mu)$ may be interpreted as the total cost in excess of the minimum required staffing cost $F(\lambda/\mu)$ for the approximately optimal staffing level $N_\lambda(z_\lambda^*)$ and for the truly optimal level N_λ^*, respectively. The above corollary identifies conditions under which these two quantities are asymptotically equal, implying that the approximate solution is asymptotically optimal in a certain sense.
In the next sections, we identify 'simple' functions $\hat{F}_\lambda(\cdot)$, $\hat{G}_\lambda(\cdot)$, and $\hat{\pi}_\lambda(\cdot)$, such that $F_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(x_\lambda^*)$ and $F_\lambda(z_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(z_\lambda^*)$, $G_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(x_\lambda^*)$ and $G_\lambda(z_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(z_\lambda^*)$, $\pi_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(x_\lambda^*)$ and $\pi_\lambda(z_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(z_\lambda^*)$. This implies $C_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{C}_\lambda(x_\lambda^*)$ and $C_\lambda(z_\lambda^*) \overset{\infty}{\approx} \hat{C}_\lambda(z_\lambda^*)$ as required in the above corollary, which then enables us to gain insight into the behavior of N_λ^*, as a function of λ.

5 Some special functions

In this section we introduce some functions that will play a central role in our analysis.
For any $x > 0$, define

$$P(x) := \frac{1}{1 + \frac{x}{h(-x)}}, \tag{12}$$

where $h(\cdot)$ is the 'hazard rate' function of the standard normal distribution, namely

$$h(x) := \frac{\phi(x)}{1 - \Phi(x)},$$

with

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}; \qquad \Phi(x) = \int_{-\infty}^{x} \phi(y)dy.$$

In Lemma B.1 we prove that $P(\cdot)$ is strictly convex decreasing.
Also define

$$Q_\lambda(x) := \frac{\exp\{N_\lambda(x)[1 - r_\lambda(x) + \log r_\lambda(x)]\}}{\sqrt{2\pi N_\lambda(x)}(1 - r_\lambda(x))},$$

with

$$r_\lambda(x) := \frac{\lambda/\mu}{N_\lambda(x)},$$

and let

$$Q(x) := \frac{\phi(x)}{x} = \frac{\mathrm{e}^{-x^2/2}}{x\sqrt{2\pi}}. \tag{13}$$

The following two lemmas characterize the asymptotic behavior of $\pi_\lambda(\cdot)$, as $\lambda \to \infty$.

Lemma 5.1 *(Halfin & Whitt [9])*
For any function x_λ with $\limsup_{\lambda\to\infty} x_\lambda < \infty$,

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} P(x_\lambda).$$

If, moreover, $\lim_{\lambda\to\infty} x_\lambda = x$, then $\pi_\lambda(x_\lambda) \overset{\infty}{\approx} P(x)$, $x \geq 0$.
In particular, if $\lim_{\lambda\to\infty} x_\lambda = 0$, then $\pi_\lambda(x_\lambda) \overset{\infty}{\approx} 1$.

Proof
Suppose to the contrary. Then there must be a subsequence $\{\lambda_n\}$ with $\lim_{n\to\infty} \lambda_n = \infty$ such that $\lim_{n\to\infty} x_{\lambda_n} = \beta$ and $\lim_{n\to\infty} \pi_{\lambda_n}(x_{\lambda_n}) = \alpha$, where $0 < \beta < \infty$ and $\alpha \neq P(\beta)$. This is in contradiction with Proposition 1 of Halfin & Whitt [9], which asserts that $\alpha = P(\beta)$ must prevail for such a sequence $\{\lambda_n\}$.

□

Lemma 5.2 *(Appendix A)*
For any function x_λ with $\lim_{\lambda\to\infty} x_\lambda = \infty$,

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} Q_\lambda(x_\lambda).$$

If also $x_\lambda \overset{\sup}{\leq} \lambda^{1/6}$, then

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} Q(x_\lambda).$$

If specifically $x_\lambda = \kappa\sqrt{\lambda/\mu}$ for some constant $\kappa > 0$, then

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} \frac{1}{\kappa\sqrt{2\pi\lambda/\mu(1+\kappa)}} \left(\frac{\mathrm{e}^\kappa}{(1+\kappa)^{1+\kappa}}\right)^{\lambda/\mu}.$$

□

We conclude the section with some observations on the behavior of the functions $F_\lambda(\cdot)$, $G_\lambda(\cdot)$, and $\pi_\lambda(\cdot)$, as defined at the outset of Section 4.
Recall that the staffing cost function $F(\cdot)$ is convex increasing, which implies that the function $F_\lambda(\cdot)$ is convex increasing as well. In addition, $F_\lambda(0) = 0$. Hence, $F_\lambda(x)/F_\lambda(y) \leq x/y$ for any pair of numbers $x \leq y$. Thus,

$$a_\lambda \overset{\sup}{>} b_\lambda \Longrightarrow F_\lambda(a_\lambda) \overset{\sup}{>} F_\lambda(b_\lambda), \tag{14}$$

and

$$a_\lambda \overset{\sup}{\gg} b_\lambda \Longrightarrow F_\lambda(a_\lambda) \overset{\sup}{\gg} F_\lambda(b_\lambda). \tag{15}$$

Also, from Lemmas 5.1 and B.1, for fixed $b \geq 0$,

$$a_\lambda \overset{\inf}{<} b \Longrightarrow P(a_\lambda) \overset{\sup}{>} P(b), \pi_\lambda(a_\lambda) \overset{\sup}{>} \pi_\lambda(b), \tag{16}$$

$$a_\lambda \overset{\sup}{>} b \Longrightarrow P(a_\lambda) \overset{\inf}{<} P(b), \pi_\lambda(a_\lambda) \overset{\inf}{<} \pi_\lambda(b), \tag{17}$$

and noting that $\lim_{x\to\infty} \pi_\lambda(x) = 0$,

$$a_\lambda \overset{\sup}{\gg} b \Longrightarrow P(a_\lambda) \overset{\inf}{\ll} P(b), \pi_\lambda(a_\lambda) \overset{\inf}{\ll} \pi_\lambda(b). \tag{18}$$

6 Case I: Rationalized regime

In this section we consider what we call a *rationalized* scenario, by which we mean that, for some $\kappa > 0$,

$$F_\lambda(\kappa) \overset{\infty}{\approx} G_\lambda(\kappa), \tag{19}$$

or in words, the staffing cost $F_\lambda(\cdot)$ is comparable to the waiting cost $G_\lambda(\cdot)$, as $\lambda \to \infty$.

For any $\lambda > 0$, define

$$y^*_\lambda := \arg\min_{y>0} C[y; F_\lambda, P, G_\lambda], \tag{20}$$

with $P(\cdot)$ as in (12).

Theorem 6.1
The staffing function y^*_λ is asymptotically optimal in the sense of Corollary 4.3.

Proof
The idea of the proof may be described as follows. In order for Corollary 4.3 to apply, we need to show that the function $P(\cdot)$ is an asymptotically exact approximation for the function $\pi_\lambda(\cdot)$ around the points x^*_λ and y^*_λ. In view of Lemma 5.1, it suffices to show that $\limsup_{\lambda\to\infty} x^*_\lambda < \infty$ and $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$. We prove this by contradiction, arguing that if this were not the case, then just the staffing cost alone would be overwhelmingly larger than the total cost associated with the fixed staffing function κ, contradicting the supposed optimality of x^*_λ or y^*_λ.
We start with showing that $\limsup_{\lambda\to\infty} x^*_\lambda < \infty$. Suppose to the contrary. Then $x^*_\lambda \overset{\sup}{\gg} \kappa$, so that

$$F_\lambda(x^*_\lambda) \overset{\sup}{\gg} F_\lambda(\kappa)$$

from (15). Using (19),

$$F_\lambda(\kappa) \overset{\infty}{\approx} F_\lambda(\kappa) + G_\lambda(\kappa) \geq F_\lambda(\kappa) + \pi_\lambda(\kappa)G_\lambda(\kappa) = C_\lambda(\kappa).$$

By definition,

$$C_\lambda(x^*_\lambda) \geq F_\lambda(x^*_\lambda).$$

Combining the above relations, we deduce

$$C_\lambda(x^*_\lambda) \overset{\sup}{\gg} C_\lambda(\kappa),$$

contradicting the optimality of x^*_λ. Thus, $\limsup_{\lambda\to\infty} x^*_\lambda < \infty$. By similar arguments, $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$.
Hence, according to Lemma 5.1, $\pi_\lambda(x^*_\lambda) \overset{\infty}{\sim} \hat{\pi}_\lambda(x^*_\lambda) = P(x^*_\lambda)$ and $\pi_\lambda(y^*_\lambda) \overset{\infty}{\sim} \hat{\pi}_\lambda(y^*_\lambda) = P(y^*_\lambda)$. Applying Corollary 4.3 then completes the proof.

□

Proposition 6.2
Assume that there exist functions $f(\cdot)$, $g(\cdot)$, and H_λ such that, for any function $k_\lambda > 0$,

$$F_\lambda(k_\lambda) \overset{\infty}{\sim} f(k_\lambda)H_\lambda, \tag{21}$$

and

$$G_\lambda(k_\lambda) \overset{\infty}{\sim} g(k_\lambda)H_\lambda, \tag{22}$$

so certainly $F_\lambda(\kappa) \overset{\infty}{\approx} G_\lambda(\kappa)$ for all $\kappa > 0$.
Define

$$y^* = \arg\min_{y>0} C[y; f, P, g].$$

The staffing function y^* is then asymptotically optimal in the sense of Corollary 4.3.

Proof
Similar to that of Theorem 6.1. Note that $F_\lambda(x^*_\lambda) \overset{\infty}{\sim} \hat{F}_\lambda(x^*_\lambda) = f(x^*_\lambda)H_\lambda$, $F_\lambda(y^*) \overset{\infty}{\sim} \hat{F}_\lambda(y^*) = f(y^*)H_\lambda$, $G_\lambda(x^*_\lambda) \overset{\infty}{\sim} \hat{G}_\lambda(x^*_\lambda) = g(x^*_\lambda)H_\lambda$, and $G_\lambda(y^*) \overset{\infty}{\sim} \hat{G}_\lambda(y^*) = g(y^*)H_\lambda$.

□

Example 6.3
Assume there is a staffing cost c_λ per server per time unit, and a waiting cost a_λ per customer per time unit, as well as a fixed penalty cost b_λ when the waiting time exceeds d_λ time units, i.e., $F(N) := Nc_\lambda$ and $D_\lambda(t) = a_\lambda t + b_\lambda \mathrm{I}_{\{t>d_\lambda\}}$, so that $G(N,\lambda) = \dfrac{a_\lambda}{N\mu - \lambda} + b_\lambda\, \mathrm{e}^{-(N\mu-\lambda)d_\lambda}$.
Thus $F_\lambda(\kappa) = c_\lambda \kappa\sqrt{\lambda/\mu}$ and $G_\lambda(\kappa) = \dfrac{a_\lambda\sqrt{\lambda/\mu}}{\kappa} + b_\lambda\lambda\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}}$.
(i) First suppose that $a_\lambda \overset{\infty}{\sim} a$, $b_\lambda\sqrt{\lambda}\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \overset{\infty}{\ll} 1$, and $c_\lambda \overset{\infty}{\sim} c$. Then (21)-(22) are satisfied for $f(\kappa) = c\kappa$, $g(\kappa) = \dfrac{a}{\kappa}$, and $H_\lambda = \sqrt{\lambda/\mu}$. Proposition 6.2 says that the staffing function

$$y^* = \arg\min_{y>0}\left\{cy + \frac{aP(y)}{y}\right\} \tag{23}$$

is asymptotically optimal in the sense of Corollary 4.3. This is exactly the $y^*(\frac{a}{c})$ that appears in (1). The numerical search for y^* is straightforward, since the function that it

minimizes is unimodal. It is important enough for our purposes, as demonstrated in Section 2, that we plotted it in Figures 1 and 2.
(ii) Now suppose that $a_\lambda \overset{\infty}{\ll} 1$, $b_\lambda \overset{\infty}{\approx} b/\sqrt{\lambda}$, $c_\lambda \overset{\infty}{\approx} c$, and $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$ with $bd\mu > c$. Then (21)-(22) are satisfied for $f(\kappa) = c\kappa$, $g(\kappa) = b\sqrt{\mu}\,\mathrm{e}^{-d\kappa\sqrt{\mu}}$, and $H_\lambda = \sqrt{\lambda/\mu}$. The asymptotically optimal staffing function is

$$y^* = \arg\min_{y>0} \left\{ cy + bP(y)\sqrt{\mu}\,\mathrm{e}^{-dy\sqrt{\mu}} \right\}.$$

(iii) Finally, consider the 'combined' case where all costs show up in the limit, i.e., suppose that $a_\lambda \overset{\infty}{\approx} a$, $b_\lambda \overset{\infty}{\approx} b/\sqrt{\lambda}$, $c_\lambda \overset{\infty}{\approx} c$, and $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$. Then (21)-(22) are satisfied for $f(\kappa) = c\kappa$, $g(\kappa) = \dfrac{a}{\kappa} + b\sqrt{\mu}\,\mathrm{e}^{-d\kappa\sqrt{\mu}}$, and $H_\lambda = \sqrt{\lambda/\mu}$. The asymptotically optimal staffing function is

$$y^* = \arg\min_{y>0} \left\{ cy + P(y)\left[\frac{a}{y} + b\sqrt{\mu}\,\mathrm{e}^{-dy\sqrt{\mu}}\right] \right\}.$$

□

Remark 6.4 *(Asymptotic expansions for $y^*(\cdot)$)*
Two asymptotic expansions for $y^*(r)$ were quoted in Section 2, one for small r and the other for large. To derive the former, one simply replaces $P(y)$ in (23) by $P(0)+yP'(0)+\frac{1}{2}y^2P''(0)$, then uses the values for the derivatives from Appendix B, and finally minimizes the resulting simple function. As for large values, one uses the well-known approximation $1-\Phi(y) \approx \phi(y)/y$ to get that also $P(y) \approx \phi(y)/y$. Substituting this approximation for $P(y)$ into (23) identifies y^* as the solution y of

$$y\mathrm{e}^{y^2/2} = \frac{a}{\sqrt{2\pi}}.$$

Changing variables to $x = y^2$, then squaring, gives $x(t)\mathrm{e}^{x(t)} = t$, where $t = \frac{a^2}{2\pi}$. A complete asymptotic expansion of $x(t)$, for large t, is calculated in [3], pages 25–28. Only its first two terms were used by the approximation in the Introduction, namely

$$x(t) = \log t - \log\log t + O(\frac{\log\log t}{\log t}).$$

Our qualitative assessments in the Introduction, about the above two approximations, are based on plotting them against the true y^*.

□

7 Case II: Efficiency-driven regime

In this section we consider an *efficiency-driven* scenario, meaning that, for all $\kappa > 0$,

$$F_\lambda(\kappa) \overset{\infty}{\gg} G_\lambda(\kappa), \tag{24}$$

($\limsup_{\lambda\to\infty} \frac{G_\lambda(\kappa)}{F_\lambda(\kappa)} < \infty$ is actually sufficient) or in words, the staffing cost dominates the waiting cost, as $\lambda \to \infty$.

For any $\lambda > 0$, define

$$y_\lambda^* := \arg\min_{y>0} C[y; F_\lambda, 1, G_\lambda]. \tag{25}$$

Theorem 7.1
The staffing level y_λ^* is asymptotically optimal in the sense of Corollary 4.3.

Proof
The idea of the proof is largely similar to that of Theorem 6.1.
We start with showing that $\lim_{\lambda\to\infty} x_\lambda^* = 0$. Suppose to the contrary. Then $x_\lambda^* \overset{\sup}{>} u$ for some $u > 0$, so that

$$F_\lambda(x_\lambda^*) \overset{\sup}{>} F_\lambda(u)$$

from (14). Using (24),

$$F_\lambda(u) \overset{\infty}{\approx} F_\lambda(u) + G_\lambda(u) \geq F_\lambda(u) + \pi_\lambda(u) G_\lambda(u) = C_\lambda(u).$$

Combining the above relations, we obtain

$$C_\lambda(x_\lambda^*) \geq F_\lambda(x_\lambda^*) \overset{\sup}{>} C_\lambda(u),$$

contradicting the optimality of x_λ^*. Thus, $\lim_{\lambda\to\infty} x_\lambda^* = 0$. By similar arguments, $\lim_{\lambda\to\infty} y_\lambda^* = 0$.
Hence, according to Lemma 5.1, $\pi_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(x_\lambda^*) = P(x_\lambda^*) \overset{\infty}{\approx} 1$, and $\pi_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(y_\lambda^*) = P(y_\lambda^*) \overset{\infty}{\approx} 1$.
Applying Corollary 4.3 then completes the proof.

□

Proposition 7.2
Assume that there exist functions $f(\cdot)$, $g(\cdot)$, H_λ, and J_λ such that, for any function $k_\lambda > 0$,

$$F_\lambda(k_\lambda) \overset{\infty}{\approx} f(k_\lambda) H_\lambda, \tag{26}$$

and

$$G_\lambda(k_\lambda) \overset{\infty}{\approx} g(k_\lambda) H_\lambda J_\lambda, \tag{27}$$

with $J_\lambda \overset{\infty}{\ll} 1$, so certainly $F_\lambda(\kappa) \overset{\infty}{\gg} G_\lambda(\kappa)$ for all $\kappa > 0$.
Define

$$y_\lambda^* = \arg\min_{y>0} C[y; f, 1, g J_\lambda].$$

Assume that $f(\cdot)$ is convex increasing and that $g(\cdot)$ is strictly convex decreasing, with $\lim_{x\downarrow 0} g(x) = \infty$ so that y_λ^* exists and is unique.
The staffing function y_λ^* is then asymptotically optimal in the sense of Corollary 4.3.

Proof
Similar to that of Theorem 7.1. Note that $F_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(x_\lambda^*) = f(x_\lambda^*)H_\lambda$, $F_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(y_\lambda^*) = f(y_\lambda^*)H_\lambda$, $G_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(x_\lambda^*) = g(x_\lambda^*)H_\lambda J_\lambda$, and $G_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(y_\lambda^*) = g(y_\lambda^*)H_\lambda J_\lambda$.

□

Remark 7.3 The rationalized staffing function (20) is in fact also asymptotically optimal in the efficiency-driven regime as defined by (24). However, the staffing function (25), where $P(\cdot)$ is replaced by 1, is asymptotically optimal as well, while considerably simpler.

□

Example 7.4
Consider the same cost structure as in Example 6.3. Now however suppose that $a_\lambda \overset{\infty}{\approx} aJ_\lambda$, $b_\lambda \overset{\infty}{\approx} bJ_\lambda/\sqrt{\lambda}$, $c_\lambda \overset{\infty}{\approx} c$, $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$, and $J_\lambda \overset{\infty}{\ll} 1$. Then (26)-(27) are satisfied for $f(\kappa) = c\kappa$, $g(\kappa) = \dfrac{a}{\kappa} + b\sqrt{\mu}\,\mathrm{e}^{-d\kappa\sqrt{\mu}}$, and $H_\lambda = \sqrt{\lambda/\mu}$. The asymptotically optimal staffing function is

$$y_\lambda^* = \arg\min_{y>0} \left\{ cy + \left[\frac{a}{y} + b\sqrt{\mu}\,\mathrm{e}^{-dy\sqrt{\mu}}\right] J_\lambda \right\}.$$

In the special case where the non-linear penalty cost are asymptotically negligible, i.e., '$b = 0$' (or rather $b_\lambda\sqrt{\lambda}\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \overset{\infty}{\ll} J_\lambda$), the optimal staffing function reduces to

$$y_\lambda^* = \arg\min_{y>0} \left\{ cy + \frac{a}{y} J_\lambda \right\}.$$

Note that the extreme case where the linear waiting cost are asymptotically vanishing, i.e., '$a = 0$' (formally $a_\lambda \overset{\infty}{\ll} J_\lambda$), does not make sense, since $\lim_{\kappa\downarrow 0} g(\kappa) = b\sqrt{\mu} < \infty$. Thus, the optimization problem will not have a strictly positive solution for large λ.

□

8 Case III: Quality-driven regime

In this section we consider a *quality-driven* regime, meaning that, for all $\kappa > 0$,

$$F_\lambda(\kappa) \overset{\infty}{\ll} G_\lambda(\kappa), \tag{28}$$

or in words, the staffing cost is negligible compared to the waiting cost, as $\lambda \to \infty$.

For any $\lambda > 0$, define

$$y_\lambda^* := \arg\min_{y>0} C[y; F_\lambda, Q_\lambda, G_\lambda].$$

Theorem 8.1
The staffing function y_λ^* is asymptotically optimal in the sense of Corollary 4.3.

Proof
The idea of the proof is largely similar to that of Theorem 6.1. However, now we rely on Lemma 5.2, which indicates that the function $Q_\lambda(\cdot)$ is an asymptotically exact approximation for the function $\pi_\lambda(\cdot)$ around the points x_λ^* and y_λ^*, provided $\lim_{\lambda\to\infty} x_\lambda^* = \infty$ and $\lim_{\lambda\to\infty} y_\lambda^* = \infty$. This may be shown again by contradiction, verifying that if this were not the case then just the waiting cost alone would now be larger than the total cost associated with some fixed staffing function u.

We start with showing that $\lim_{\lambda\to\infty} x_\lambda^* = \infty$. Suppose to the contrary. Then $x_\lambda^* \overset{\inf}{<} u$ for some $u > 0$, so that

$$\pi_\lambda(x_\lambda^*)G_\lambda(x_\lambda^*) \overset{\sup}{>} \pi_\lambda(u)G_\lambda(u)$$

from (16) and the fact that $G_\lambda(\cdot)$ is decreasing. Using Lemma 5.2 and (28),

$$\pi_\lambda(u)G_\lambda(u) \overset{\infty}{\approx} P(u)G_\lambda(u) \overset{\infty}{\approx} F_\lambda(u) + P(u)G_\lambda(u) \overset{\infty}{\approx} F_\lambda(u) + \pi_\lambda(u)G_\lambda(u) = C_\lambda(u).$$

Combining the above relations, we deduce

$$C_\lambda(x_\lambda^*) \geq \pi_\lambda G_\lambda(x_\lambda^*) \overset{\sup}{>} C_\lambda(u),$$

contradicting the optimality of x_λ^*. Thus, $\lim_{\lambda\to\infty} x_\lambda^* = \infty$. By similar arguments, $\lim_{\lambda\to\infty} y_\lambda^* = \infty$. Hence, according to Lemma 5.2, $\pi_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(y_\lambda^*) = Q_\lambda(x_\lambda^*)$, and $\pi_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(y_\lambda^*) = Q_\lambda(y_\lambda^*)$. Applying Corollary 4.3 then completes the proof.

□

Proposition 8.2
Assume that there exist functions $f(\cdot)$, $g(\cdot)$, H_λ, and J_λ such that, for any function $k_\lambda > 0$,

$$F_\lambda(k_\lambda) \overset{\infty}{\approx} f(k_\lambda)H_\lambda \tag{29}$$

and

$$G_\lambda(k_\lambda) \overset{\infty}{\approx} g(k_\lambda)H_\lambda J_\lambda, \tag{30}$$

with $J_\lambda \overset{\infty}{\gg} 1$, and $g(\lambda^\alpha)J_\lambda Q(\lambda^\alpha) \overset{\infty}{\ll} f(\lambda^\alpha)$ for some $\alpha < 1/6$, so that certainly $F_\lambda(\kappa) \overset{\infty}{\ll} G_\lambda(\kappa)$ for all $\kappa > 0$.
Define

$$y_\lambda^* = \arg\min_{y>0} C[y; f, Q, gJ_\lambda]. \tag{31}$$

The staffing function y_λ^* is then asymptotically optimal in the sense of Corollary 4.3.

Proof
In a similar fashion as in the proof of Theorem 8.1 it may be shown that $\lim_{\lambda\to\infty} x_\lambda^* = \infty$ and $\lim_{\lambda\to\infty} y_\lambda^* = \infty$. However, to show that the function $Q(\cdot)$ is an asymptotically exact

approximation for the function $\pi_\lambda(\cdot)$(using Lemma 5.2), we additionally need to prove that x_λ^* and y_λ^* do not grow faster than λ^α.
We start with showing that $x_\lambda^* \overset{\sup}{\leq} \lambda^\alpha$. Suppose to the contrary. Then $x_\lambda^* \overset{\sup}{>} \lambda^\alpha$, so that

$$F_\lambda(x_\lambda^*) \overset{\sup}{>} F_\lambda(\lambda^\alpha)$$

from (14). Using Lemma 5.2 and (29), (30),

$$F_\lambda(\lambda^\alpha) \overset{\infty}{\approx} F_\lambda(\lambda^\alpha) + Q(\lambda^\alpha)G_\lambda(\lambda^\alpha) \overset{\infty}{\approx} F_\lambda(\lambda^\alpha) + \pi_\lambda(\lambda^\alpha)G_\lambda(\lambda^\alpha) = C_\lambda(\lambda^\alpha).$$

Combining the above relations, we obtain

$$C_\lambda(x_\lambda^*) \geq F_\lambda(x_\lambda^*) \overset{\sup}{>} C_\lambda(\lambda^\alpha),$$

contradicting the optimality of x_λ^*. Thus, $x_\lambda^* \overset{\sup}{\leq} \lambda^\alpha$. By similar arguments, $y_\lambda^* \overset{\sup}{\leq} \lambda^\alpha$. Hence, according to Lemma 5.2, $\pi_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(x_\lambda^*) = Q(x_\lambda^*)$, and $\pi_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(y_\lambda^*) = Q(y_\lambda^*)$. Applying Corollary 4.3 then completes the proof. Note that $F_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(x_\lambda^*) = f(x_\lambda^*)H_\lambda$, $F_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{F}_\lambda(y_\lambda^*) = f(y_\lambda^*)H_\lambda$, $G_\lambda(x_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(x_\lambda^*) = g(x_\lambda^*)H_\lambda J_\lambda$, and $G_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{G}_\lambda(y_\lambda^*) = g(y_\lambda^*)H_\lambda J_\lambda$.

□

Remark 8.3 The rationalized staffing function (20) is in fact also asymptotically optimal in the quality-driven regime as defined by (28) when $x_\lambda^* \overset{\infty}{\ll} \lambda^{1/6}$, since the proof of Theorem 8.2 then shows that $\pi_\lambda(x_\lambda^*) \overset{\infty}{\approx} Q(x_\lambda^*)$, while $Q(x_\lambda^*) \overset{\infty}{\approx} P(x_\lambda^*)$. However, the staffing function (31) is asymptotically optimal as well, while simpler.

□

Example 8.4
Consider the same cost structure as in Example 6.3. Now however suppose that $a_\lambda \overset{\infty}{\approx} aJ_\lambda$, $b_\lambda \overset{\infty}{\approx} bJ_\lambda/\sqrt{\lambda}$, $c_\lambda \overset{\infty}{\approx} c$, and $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$, with $J_\lambda \overset{\infty}{\gg} 1$. Then (29)-(30) are satisfied for $f(\kappa) = c\kappa$, $g(\kappa) = \dfrac{a}{\kappa} + b\sqrt{\mu}\,\mathrm{e}^{-d\kappa\sqrt{\mu}}$, and $H_\lambda = \sqrt{\lambda/\mu}$. The asymptotically optimal staffing function is

$$y_\lambda^* = \arg\min_{y>0} \left\{ cy + Q(y) \left[\frac{a}{y} + b\sqrt{\mu}\,\mathrm{e}^{-dy\sqrt{\mu}} \right] J_\lambda \right\}.$$

In the special case where the non-linear penalty cost are asymptotically negligible, i.e., '$b = 0$' (or rather $b_\lambda\sqrt{\lambda}\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \overset{\infty}{\ll} J_\lambda$), the optimal staffing functions reduces to

$$y_\lambda^* = \arg\min_{y>0} \left\{ cy + \frac{aQ(y)}{y} J_\lambda \right\}.$$

In the extreme case where the linear waiting cost are asymptotically insignificant, i.e., '$a = 0$' (formally, $a_\lambda \overset{\infty}{\ll} J_\lambda$), the optimal staffing functions takes the form

$$y_\lambda^* = \arg\min_{y>0} \left\{ cy + bQ(y)\sqrt{\mu}\,\mathrm{e}^{-dy\sqrt{\mu}} J_\lambda \right\}.$$

□

9 Constraint satisfaction

In the previous sections, we have considered the problem of determining the staffing level so as to minimize the total staffing and waiting cost. A closely related problem, which is in fact motivated by actual practice, is to minimize the staffing level subject to a constraint $M_\lambda > 0$ on the waiting cost.
So we are now interested in determining

$$N_\lambda^* := \min_{N > \lambda/\mu} \{N : K(N, \lambda) \leq M_\lambda\}, \tag{32}$$

with $K(N, \lambda) := \lambda \pi(N, \lambda/\mu) G(N, \lambda)$ denoting the waiting cost. Notice that $\lim_{N \to \infty} K(N, \lambda) = 0$, so N_λ^* is well-defined.
As for the cost minimization problem, our approach is to first translate the discrete problem (32) into a continuous one, and then approximate the latter problem by a related continuous problem which is easier to solve. Denote

$$x_\lambda^* := \min_{x>0} \{x : K_\lambda(x) \leq M_\lambda\},$$

with $K_\lambda(x) := \pi_\lambda(x) G_\lambda(x)$. Since $\pi_\lambda(\cdot)$ and $G_\lambda(\cdot)$ are both continuous and strictly decreasing, x_λ^* is the unique solution to the equation $K_\lambda(x) = M_\lambda$. Further notice that $N_\lambda^* = \lceil N_\lambda(x_\lambda^*) \rceil$, which establishes the link between the discrete problem and the corresponding continuous problem.
To approximate x_λ^*, define z_λ^* as the solution to the equation $\hat{\pi}_\lambda(z) \hat{G}_\lambda(z) = M_\lambda$. The functions $\hat{\pi}_\lambda(\cdot)$ and $\hat{G}_\lambda(\cdot)$ that we consider will always be such that z_λ^* exists and is unique.
We now formulate the approximation principle underlying our approach, in parallel to that for the cost minimization problem.
Define

$$T_\lambda(x) := \min\{\mid K(\lfloor N_\lambda(x) \rfloor, \lambda) - M_\lambda \mid, \mid K(\lceil N_\lambda(x) \rceil, \lambda) - M_\lambda \mid, \mid K(\lceil N_\lambda(x) \rceil, \lambda) - K(N_\lambda^*, \lambda) \mid\}. \tag{33}$$

Lemma 9.1 *(Asymptotic Optimality)*
Denote $\hat{K}_\lambda(y) = \hat{\pi}_\lambda(y) \hat{G}_\lambda(y)$. Let z_λ^* be as defined above. If $K_\lambda(z_\lambda^*) \overset{\infty}{\approx} \hat{K}_\lambda(z_\lambda^*)$, then the staffing function z_λ^* is asymptotically optimal in the sense that, as $\lambda \to \infty$, $T_\lambda(z_\lambda^*) \overset{\infty}{\ll} M_\lambda$, with $T_\lambda(\cdot)$ given in (33).

Proof
For fixed λ, we distinguish between three cases.
i. $N_\lambda^* - 1 < N_\lambda(z_\lambda^*) \leq N_\lambda^*$. Then $\lceil N_\lambda(z_\lambda^*) \rceil = N_\lambda^*$, so that $T_\lambda(z_\lambda^*) = 0$.
ii. $N_\lambda(z_\lambda^*) \leq N_\lambda^* - 1$. Then

$$M_\lambda = K_\lambda(x_\lambda^*) \leq K(N_\lambda^* - 1, \lambda) \leq K(\lceil N_\lambda(z_\lambda^*) \rceil, \lambda) \leq K_\lambda(z_\lambda^*),$$

so that

$$T_\lambda(z_\lambda^*) \leq \mid K(\lceil N_\lambda(z_\lambda^*) \rceil, \lambda) - M_\lambda \mid \leq K_\lambda(z_\lambda^*) - M_\lambda.$$

iii. $N_\lambda(z^*_\lambda) > N^*_\lambda$. Then

$$K_\lambda(z^*_\lambda) \leq K(\lfloor N_\lambda(z^*_\lambda) \rfloor, \lambda) \leq K(N^*_\lambda, \lambda) \leq K_\lambda(x^*_\lambda) = M_\lambda,$$

so that

$$T_\lambda(z^*_\lambda) \leq | \; K(\lfloor N_\lambda(z^*_\lambda) \rfloor, \lambda) - M_\lambda \; | \leq M_\lambda - K_\lambda(z^*_\lambda).$$

Thus, for all λ,

$$T_\lambda(z^*_\lambda) \leq | \; K_\lambda(z^*_\lambda) - M_\lambda \; | \stackrel{\infty}{\ll} M_\lambda,$$

as $\hat{K}_\lambda(z^*_\lambda) = M_\lambda$ by definition.

$\square$

In full generality, it seems difficult to establish a stronger optimality property than indicated in the above lemma. Under additional conditions, however, it is possible to make sharper statements. For example, a more desirable criterion for asymptotic optimality would be

$$| \; K(\lceil N_\lambda(z^*_\lambda) \rceil, \lambda) - K(N^*_\lambda, \lambda) \; | \stackrel{\infty}{\ll} M_\lambda.$$

And indeed, following the same reasoning of the proof of Lemma 9.1, it can be guaranteed to hold but under additional constraints on the oscillation of our costs.

9.1 Rationalized regime

We first consider a rationalized scenario, by which we mean that, for some $\kappa > 0$,

$$G_\lambda(\kappa) \stackrel{\infty}{\approx} M_\lambda, \tag{34}$$

($\limsup\limits_{\lambda\to\infty} \frac{G_\lambda(\kappa)}{M_\lambda} < \infty$ is actually sufficient) or in words, the waiting cost $G_\lambda(\cdot)$ is comparable to the constraint M_λ, as $\lambda \to \infty$.
For any $\lambda > 0$, define y^*_λ as the solution to the equation $P(y)G_\lambda(y) = M_\lambda$.
Note that $P(\cdot)$ and $G_\lambda(\cdot)$ are both continuous and strictly decreasing with $\lim\limits_{x\downarrow 0} P(x)G_\lambda(x) = \infty$ and $\lim\limits_{x\to\infty} P(x)G_\lambda(x) = 0$, so that y^*_λ exists and is unique.

Theorem 9.2
The staffing function y^*_λ is asymptotically optimal in the sense of Lemma 9.1.

Proof
The idea of the proof may be described as follows. In order for Lemma 9.1 to apply, we need to show that the function $P(\cdot)$ is asymptotically close to the function $\pi_\lambda(\cdot)$ around y^*_λ. In view of Lemma 5.1, it suffices to show that $\limsup\limits_{\lambda\to\infty} y^*_\lambda < \infty$. We prove this by contradiction, arguing that if this were not the case, then the incurred waiting cost would be strictly smaller than the constraint value M_λ.

We start with showing that $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$. Suppose to the contrary. Then $y^*_\lambda \stackrel{\sup}{\gg} \kappa$, so that

$$P(y^*_\lambda)G_\lambda(y^*_\lambda) \stackrel{\inf}{\ll} P(\kappa)G_\lambda(\kappa)$$

from (18) and the fact that $G_\lambda(\cdot)$ is decreasing. Using (34),

$$P(\kappa)G_\lambda(\kappa) \leq G_\lambda(\kappa) \stackrel{\infty}{\lessapprox} M_\lambda.$$

Combining the above relations, we deduce

$$P(y^*_\lambda)G_\lambda(y^*_\lambda) \stackrel{\inf}{\ll} M_\lambda,$$

contradicting the definition of y^*_λ. Thus, $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$.
Hence, according to Lemma 5.1, $\pi_\lambda(y^*_\lambda) \stackrel{\infty}{\sim} \hat{\pi}_\lambda(y^*_\lambda) = P(y^*_\lambda)$.
Applying Lemma 9.1 then completes the proof.

□

Proposition 9.3
Assume that there exists a function $g(\cdot)$, such that, for any function $k_\lambda > 0$,

$$G_\lambda(\kappa) \stackrel{\infty}{\sim} g(\kappa)M_\lambda, \tag{35}$$

so certainly $G_\lambda(\kappa) \stackrel{\infty}{\lessapprox} M_\lambda$ for all $\kappa > 0$.
Define y^* as the solution to the equation $P(y)g(y) = 1$.
Assume that $g(\cdot)$ is continuous and decreasing, with $\lim_{x\downarrow 0} g(x) > 1$ so that y^* exists and is unique.
The staffing function y^* is then asymptotically optimal in the sense of Lemma 9.1.

Proof
Similar to that of Theorem 9.2. Note that $G_\lambda(y^*) \stackrel{\infty}{\sim} \hat{G}_\lambda(y^*) = g(y^*)M_\lambda$.

□

Example 9.4
Assume there is a waiting cost a_λ per customer per time unit, as well as a fixed penalty cost b_λ when the waiting time exceeds d_λ time units, i.e., $D_\lambda(t) = a_\lambda t + b_\lambda \mathrm{I}_{\{t>d_\lambda\}}$, so that $G(N,\lambda) = \dfrac{a_\lambda}{N\mu - \lambda} + b_\lambda\, \mathrm{e}^{-(N\mu-\lambda)d_\lambda}$. Thus $G_\lambda(\kappa) = \dfrac{a_\lambda\sqrt{\lambda/\mu}}{\kappa} + b_\lambda\lambda\, \mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}}$.
(i) First suppose that $a_\lambda \stackrel{\infty}{\sim} a\sqrt{\lambda/\mu}$, $b_\lambda\, \mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \stackrel{\infty}{\ll} 1$, and $M_\lambda = M\lambda$. Then (35) is satisfied for $g(\kappa) = \dfrac{a}{\kappa\mu M}$. Proposition 9.3 says that the staffing function y^* determined as the unique solution to the equation

$$\frac{aP(y)}{\mu y} = M$$

is asymptotically optimal in the sense of Lemma 9.1.
(ii) Now suppose that $a_\lambda/\sqrt{\lambda} \overset{\infty}{\ll} 1$, $b_\lambda \overset{\infty}{\approx} b$, $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$, and $M_\lambda = M\lambda$ with $b > M$. Then (35) is satisfied for $g(\kappa) = b\,\mathrm{e}^{-d\kappa\sqrt{\mu}}/M$. The asymptotically optimal staffing function is the unique solution to the equation

$$bP(y)\,\mathrm{e}^{-dy\sqrt{\mu}} = M.$$

(iii) Finally, consider the 'combined' case where all costs show up in the limit, i.e., suppose that $a_\lambda \overset{\infty}{\approx} a\sqrt{\lambda/\mu}$, $b_\lambda \overset{\infty}{\approx} b$, $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$, and $M_\lambda = M\lambda$. Then (35) is satisfied for $g(\kappa) = \dfrac{a}{\kappa\mu M} + b\,\mathrm{e}^{-d\kappa\sqrt{\mu}}/M$. The asymptotically optimal staffing function is the unique solution to the equation

$$P(y)[\frac{a}{\mu y} + b\,\mathrm{e}^{-dy\sqrt{\mu}}] = M.$$

□

Example 9.5
An important special case is $a_\lambda = 0$, $b_\lambda = 1$, $d_\lambda = 0$, $M_\lambda = \epsilon\lambda$, which corresponds to a target waiting probability ϵ. Case ii) of the above example then shows that the staffing function $y^* = P^{-1}(\epsilon)$ is asymptotically optimal. We described this example in Section 2– see (5). It is to be compared with (6), used in [18] and [14]. On the differences and similarities between the two approximations, see Section 10.

□

9.2 Efficiency-driven regime

We now consider an efficiency-driven scenario, meaning that, for all $\kappa > 0$,

$$G_\lambda(\kappa) \overset{\infty}{\ll} M_\lambda, \tag{36}$$

(in fact $G_\lambda(\kappa) \overset{\sup}{\leq} M_\lambda$ would be sufficient for the results below to hold) or in words, the waiting cost is dominated by the target upper bound, as $\lambda \to \infty$.
For any $\lambda > 0$, define y_λ^* as the solution to the equation $G_\lambda(y) = M_\lambda$.
Note that $G_\lambda(\cdot)$ is continuous and strictly decreasing with $\lim_{x\downarrow 0} G_\lambda(x) = \infty$ and $\lim_{x\to\infty} G_\lambda(x) = 0$, so that y_λ^* exists and is unique.

Theorem 9.6
The staffing function y_λ^* is asymptotically optimal in the sense of Lemma 9.1.

Proof
The idea of the proof is largely similar to that of Theorem 9.2.
We start with showing that $\lim_{\lambda\to\infty} y_\lambda^* = 0$. Suppose to the contrary. Then $y_\lambda^* \overset{\sup}{>} u$ for some $u > 0$, so that

$$P(y_\lambda^*)G_\lambda(y_\lambda^*) \overset{\inf}{<} P(u)G_\lambda(u),$$

from (17) and the fact that $G_\lambda(\cdot)$ is decreasing. Using (36),

$$P(u)G_\lambda(u) \leq G_\lambda(u) \overset{\sup}{\leq} M_\lambda,$$

Combining the above relations, we deduce

$$P(y^*_\lambda)G_\lambda(y^*_\lambda) \overset{\inf}{<} M_\lambda,$$

contradicting the definition of y^*_λ. Thus, $\lim_{\lambda\to\infty} y^*_\lambda = 0$.
Hence, according to Lemma 5.1, $\pi_\lambda(y^*_\lambda) \overset{\infty}{\approx} \hat{\pi}_\lambda(y^*_\lambda) = P(y^*_\lambda) \overset{\infty}{\approx} 1$.
Applying Lemma 9.1 then completes the proof.

□

Proposition 9.7
Assume that there exist functions $g(\cdot)$ and J_λ such that, for any function $k_\lambda > 0$,

$$G_\lambda(k_\lambda) \overset{\infty}{\approx} g(k_\lambda)J_\lambda M_\lambda, \tag{37}$$

with $J_\lambda \overset{\infty}{\ll} 1$, so certainly $G_\lambda(\kappa) \overset{\infty}{\ll} M_\lambda$ for all $\kappa > 0$.
Define y^*_λ as the unique solution to the equation $g(y)J_\lambda = 1$.
Assume that $g(\cdot)$ is continuous and strictly decreasing, with $\lim_{x\downarrow 0} g(x) = \infty$ so that y^*_λ exists and is unique.
The staffing function y^*_λ is then asymptotically optimal in the sense of Lemma 9.1.

Proof
Similar to that of Theorem 9.6. Note that $G_\lambda(y^*_\lambda) \overset{\infty}{\approx} g(y^*_\lambda)J_\lambda M_\lambda$.

□

Example 9.8
Consider the same cost structure as in Example 9.4. Now however suppose that $a_\lambda \overset{\infty}{\approx} a\sqrt{\lambda/\mu}$, $b_\lambda \overset{\infty}{\approx} b$, $d_\lambda \overset{\infty}{\approx} d/\sqrt{\lambda}$, and $M_\lambda = M\lambda/J_\lambda$, with $J_\lambda \overset{\infty}{\ll} 1$. Then (37) is satisfied for $g(\kappa) = \dfrac{a}{\kappa\mu M} + b\,\mathrm{e}^{-d\kappa\sqrt{\mu}}/M$. The asymptotically optimal staffing function is the unique solution to the equation

$$\left[\frac{a}{\mu y} + b\,\mathrm{e}^{-dy\sqrt{\mu}}\right] J_\lambda = M.$$

In the special case where the non-linear penalty cost are asymptotically negligible, i.e., '$b = 0$' (or rather $b_\lambda\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \overset{\infty}{\ll} 1$), the equation for the optimal staffing function reduces to

$$y^*_\lambda = \frac{a}{\mu M} J_\lambda.$$

As noted earlier for the optimization problem, the extreme case where the linear waiting cost are asymptotically vanishing, i.e., '$a = 0$' (formally, $a_\lambda\sqrt{\lambda} \overset{\infty}{\ll} 1$) does not make sense, since $\lim_{\kappa\downarrow 0} g(\kappa) = b/M < \infty$. Thus, the equation for the optimal staffing function will not have a positive solution for large λ.

□

9.3 Quality-driven regime

We finally consider a quality-driven scenario, meaning that, for all $\kappa > 0$,

$$G_\lambda(\kappa) \overset{\infty}{\gg} M_\lambda, \tag{38}$$

or in words, the waiting cost dominates the target upper bound, as $\lambda \to \infty$.
For any $\lambda > 0$, define y_λ^* as the solution to the equation $G_\lambda(y)Q_\lambda(y) = M_\lambda$.
Note that $G_\lambda(\cdot)$ and $Q_\lambda(\cdot)$ are continuous and strictly decreasing with $\lim\limits_{x\downarrow 0} G_\lambda(x)Q_\lambda(x) = \infty$ and $\lim\limits_{x\to\infty} G_\lambda(x)Q_\lambda(x) = 0$, so that y_λ^* exists and is unique.

Theorem 9.9
The staffing function y_λ^* is asymptotically optimal in the sense of Lemma 9.1.

Proof
The proof is largely similar to that of Theorem 9.2. However, now we rely on Lemma 5.2, which indicates that the function $Q_\lambda(\cdot)$ is asymptotically close to the function $\pi_\lambda(\cdot)$, provided $\lim\limits_{\lambda\to\infty} y_\lambda^* = \infty$. This may be shown again by contradiction, verifying that if this were not the case, then the incurred waiting cost would now strictly exceed the constraint value M_λ.
We start with showing that $\lim\limits_{\lambda\to\infty} y_\lambda^* = \infty$. Suppose to the contrary. Then $y_\lambda^* \overset{\inf}{<} u$ for some $u > 0$, so that

$$P(y_\lambda^*)G_\lambda(y_\lambda^*) \overset{\sup}{>} P(u)G_\lambda(u)$$

from (16) and the fact that $G_\lambda(\cdot)$ is decreasing. Using Lemma 5.2 and (38),

$$P(u)G_\lambda(u) \overset{\infty}{\approx} P(u)G_\lambda(u) \overset{\infty}{\gg} M_\lambda.$$

Combining the above relations, we deduce

$$P(y_\lambda^*)G_\lambda(y_\lambda^*) \overset{\sup}{>} M_\lambda,$$

contradicting the definition of y_λ^*. Thus, $\lim\limits_{\lambda\to\infty} y_\lambda^* = \infty$.
Hence, according to Lemma 5.2, $\pi_\lambda(y_\lambda^*) \overset{\infty}{\approx} \hat{\pi}_\lambda(y_\lambda^*) = Q_\lambda(y_\lambda^*)$.
Applying Lemma 9.1 then completes the proof.

□

Proposition 9.10
Assume that there exist functions $g(\cdot)$ and J_λ such that, for any function $k_\lambda > 0$,

$$G_\lambda(k_\lambda) \overset{\infty}{\approx} g(k_\lambda)J_\lambda M_\lambda, \tag{39}$$

with $J_\lambda \overset{\infty}{\gg} 1$, and $g(\lambda^\alpha)J_\lambda Q(\lambda^\alpha) \overset{\infty}{\ll} 1$ for some $\alpha < 1/6$, so that certainly $G_\lambda(\kappa) \overset{\infty}{\gg} M_\lambda$ for all $\kappa > 0$.
Define y_λ^* as the unique solution to the equation

$$Q(y)g(y)J_\lambda = 1.$$

Assume that $g(\cdot)$ is continuous and strictly decreasing, with $\lim_{x\downarrow 0} g(x) > 0$ so that y_λ^* exists and is unique.
The staffing function y_λ^* is then asymptotically optimal in the sense of Lemma 9.1.

Proof
It may easily be shown that $y_\lambda^* \overset{\sup}{\leq} \lambda^\alpha$. Hence, according to Lemma 5.2, $\pi_\lambda(y_\lambda^*) \overset{\infty}{\sim} \hat{\pi}_\lambda(y_\lambda^*) = Q(y_\lambda^*)$.
The proof is further similar to that of Theorem 9.9. Note that $G_\lambda(y_\lambda^*) \overset{\infty}{\sim} \hat{G}_\lambda(y_\lambda^*) = g(y_\lambda^*)M_\lambda J_\lambda$.

□

Example 9.11
Consider the same cost structure as in Example 9.4. Now however suppose that $a_\lambda \overset{\infty}{\sim} a\sqrt{\lambda/\mu}$, $b_\lambda \overset{\infty}{\sim} b$, $d_\lambda \overset{\infty}{\sim} d/\sqrt{\lambda}$, and $M_\lambda = M\lambda/J_\lambda$, with $J_\lambda \overset{\infty}{\gg} 1$. Then (39) is satisfied for $g(\kappa) = \dfrac{a}{\kappa\mu M} + b\,\mathrm{e}^{-d\kappa\sqrt{\mu}}/M$. The asymptotically optimal staffing function is the unique solution of the equation

$$Q(y)[\frac{a}{\mu y} + b\,\mathrm{e}^{-dy\sqrt{\mu}}]J_\lambda = M.$$

In the special case where the non-linear penalty cost are asymptotically negligible, i.e., '$b = 0$' (or rather $b_\lambda\,\mathrm{e}^{-\mu\kappa d_\lambda\sqrt{\lambda/\mu}} \overset{\infty}{\ll} 1$), the equation for the optimal staffing function reduces to

$$\frac{aQ(y)}{\mu y}J_\lambda = M.$$

In the extreme case where the linear waiting cost are asymptotically insignificant, i.e., '$a = 0$' (formally, $a_\lambda \overset{\infty}{\ll} \sqrt{\lambda}$), the equation for the optimal staffing function takes the form

$$bQ(y)\,\mathrm{e}^{-dy\sqrt{\mu}}J_\lambda = M.$$

□

Remark 9.12 Notice that the results for the constraint satisfaction problem closely mirror those for the cost minimization problem. In fact, the two problems may be formally related as follows. Consider a strictly decreasing function $M(\cdot)$ on $(0,\infty)$ with $\lim_{x\downarrow 0} M(x) = \infty$ and $\lim_{x\to\infty} M(x) = 0$. Then the (unique) solution to the equation $M(x) = 1$ is $\arg\min_{x>0}\{M(x) + 1/M(x)\}$. Thus, the solution to the constraint satisfaction problem $\pi_\lambda(x)G_\lambda(x) = M_\lambda$ may also be represented as $\arg\min_{x>0}\{\frac{M_\lambda^2}{\pi_\lambda(x)G_\lambda(x)} + \pi_\lambda(x)G_\lambda(x)\}$, which has the form of the cost minimization problem. The relation thus established is only formal. We could not utilize it to derive the results for constraint satisfaction from the optimization results.

□

10 Numerical experiments

In this section we present the results of some numerical experiments that we carried out. The main purpose of the numerical experiments was to test the accuracy of the approximations that arise from our asymptotically optimal staffing levels. The numerical results indicate that the rationalized approximation performs exceptionally well in all regimes. By Remark 7.3 we know that the rationalized approximation is in fact asymptotically optimal in all regimes. On the other hand, the accuracy displayed by the rationalized approximation is astonishingly better than our rigorous results lead us to believe.
Our first two experiments address [8] and [14], which correspond to Examples 6.3 and 9.5, respectively.
Grassmann [8], Table 3: Grassmann calculates the optimal staffing level N^* for the M/M/N queue, with offered loads $R = 1, 3, 10, 30, 100$ and costs $r = \frac{a}{c} = 10, 20, 100, 200$. While the latter are rather extreme values, our approximation (1) is nevertheless accurate: it is exact in 7 cases and off by only 1 agent in the other 13 cases.

□

Kolesar & Green [14], Table 1: Kolesar & Green use (6) in order to calculate N^* that achieves $\Pr\{\text{Wait} > 0\} = \epsilon$, for $\epsilon = 0.2, 0.1, 0.05, 0.025, 0.01, 0.001$, and offered loads $R = 2^m$, $m = 0, 1, \ldots, 10$. The approximation (1) is superior to (6). This is to be expected in view of the theory that supports the former, while the latter is heuristically based. Indeed, for $\epsilon = 0.2$, (1) is exact for 9 cases and misses by 1 agent for the other 2. In contrast, (6) misses by up to 7 agents. But more significantly, the misses in staffing levels lead to misses in the target delay probabilities, off by 25-75% in 8 out of the 11 cases.
The approximation (6) improves as ϵ decreases, until eventually it coincides with (1). This is understood as follows: small values of ϵ give rise to large y^* (quality-driven), for which

$$\bar{\Phi}(y) \approx \frac{\phi(y)}{y} \approx P(y).$$

□

We now turn to numerical experiments related to Examples 6.3, 7.4, and 8.4. In all cases we compared an approximation to the optimal staffing level obtained from the asymptotics with the exact optimal staffing level, which we obtained through a simple search procedure. (The unimodality of $C(N, \lambda)$ in N makes such a search simple.)
In all of the examples we use $c_\lambda = c = 1$ and $\mu = 1$. Once we set $c_\lambda = c$, taking $c = 1$ is without loss of generality because we can take this as the definition of the monetary unit. Similarly, taking $\mu = 1$ is also without loss of generality because we can take $1/\mu$ as the definition of the time unit. The parameter settings for the experiments are summarized in Table 2.
We first describe the numerical results related to Example 6.3. We considered three cases:

i) $a_\lambda = a, \quad b_\lambda = 0$

ii) $a_\lambda = 0, \quad b_\lambda = b/\sqrt{\lambda}, \quad d_\lambda = d/\sqrt{\lambda}$

$\mu = 1$, $c_\lambda = c = 1$						
Example 6.3					Example 7.4	Example 8.4
i.		ii.		iii.		
$\lambda = 100$	$a = 2$	$\lambda = 100$, $d = 0.1$	$b = 5$, $d = 1$	$a = 2$, $b = 2.5$, $d = 0.1$	$a = 1$, $b_\lambda = 0$	$a = 1$, $b_\lambda = 0$
a	λ	b	λ	λ	λ	λ
0.1	5	0.1	5	5	10	10
0.25	6	0.25	6	6	20	20
0.5	7	0.5	7	7	30	30
1	$\vdots$	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	$\vdots$	2	$\vdots$	$\vdots$	$\vdots$	$\vdots$
4	99	4	99	99	190	190
10	100	10	100	100	200	200

Table 2: Overview of the parameter settings for the numerical experiments

iii) $a_\lambda = a, \quad b_\lambda = b/\sqrt{\lambda}, \quad d_\lambda = d/\sqrt{\lambda}.$

These correspond to the three cases in Example 6.3. First consider case i). For $r > 0$, let

$$y^*(r) = \arg\min_{y>0} \left\{ y + \frac{rP(y)}{y} \right\}. \tag{40}$$

We plot $y^*(r)$, $0 \le r \le 10$ in Figure 1.
Let

$$n^* = \lambda/\mu + y^*(r)\sqrt{\lambda/\mu}.$$

The rationalized approximation for case i) of Example 6.3 is obtained by rounding n^* to the nearest integer. (The asymptotic analysis gives no guidance of how to go from n^* to an integer staffing level. Preliminary numerical calculations comparing rounding up, rounding down, and rounding off showed that rounding off is generally superior. So all of our numerical results involve rounding off. Of course, if rounding off n^* yields a value smaller than or equal to λ/μ, the staffing level must be increased to be strictly greater λ/μ to avoid an unstable system.)
To check this approximation we first tried $\lambda = 100$ and the seven different values of a indicated in Table 2. In all of these cases rounding off n^* gave the exact optimal staffing level. We next set $a = 2$ and tried all integer values of λ between 5 and 100. Here rounding off n^* is never off by more than 1, and is usually exact (83 out of 96 cases).
For case ii) we let

$$n^* = \lambda/\mu + y^*(b,d)\sqrt{\lambda/\mu},$$

where

$$y^*(b,d) = \arg\min_{y>0} \left\{ y + bP(y)\,\mathrm{e}^{-dy} \right\}.$$

Here we first tried $\lambda = 100$ and $d = 0.1$ with the seven different values of b indicated in Table 2. Again we found that in all of these cases rounding off n^* gave the exact optimal staffing level. We next set $b = 5$ and $d = 1$ and tried all integer values between 5 and 100 for λ. Rounding off n^* is almost always exact (84 out of 96 times), and is never off by more than 1.

For case iii) we let

$$n^* = \lambda/\mu + y^*(a, b, d)\sqrt{\lambda/\mu},$$

where

$$y^*(a, b, d) = \arg\min_{y>0} \left\{ y + P(y)[\frac{a}{y} + b\,\mathrm{e}^{-dy}] \right\}.$$

Here we set $a = 2$, $b = 2.5$, and $d = 0.1$, and tried all integer values between 5 and 100 for λ (see Table 2). Again, rounding off n^* is almost always exact (80 out of 96 cases), and is never off by more than 1.

For Example 7.4 we restricted our attention to $b_\lambda = 0$. We initially set

$$a_\lambda = a\lambda^{-1/2}. \tag{41}$$

Defining

$$y^*_\lambda(a) = \arg\min_{y>0} \left\{ y + \frac{a}{y\sqrt{\lambda}} \right\},$$

we can solve explicitly to obtain $y^*_\lambda(a) = \sqrt{a}\lambda^{-1/4}$. We thus let

$$n^* = \lambda/\mu + \sqrt{a/\mu}\lambda^{1/4}.$$

For the numerical test of this approximation we set $a = 1$ and tried all integer multiples of 10 between 10 and 200 for λ (see Table 2), using (41) to determine a_λ. Rounding off n^* to the nearest integer is almost always exact (19 out of 20 times), and is never off by more than 1.

In Example 8.4 we again restricted our attention to $b_\lambda = 0$. We took

$$a_\lambda = a\sqrt{\lambda}. \tag{42}$$

Let

$$n^* = \lambda/\mu + y^*_\lambda(a)\sqrt{\lambda/\mu},$$

where

$$y^*_\lambda(a) = \arg\min_{y>0} \left\{ y + \frac{aQ(y)}{\sqrt{\lambda}} y \right\}, \tag{43}$$

and $Q(y)$ is given by (13).

For the numerical test of this approximation we again set $a = 1$ and tried all integer multiples of 10 between 10 and 200 for λ (see Table 2), using (42) to determine a_λ. Once

$\mu = 1$, $a = 1$, $b_\lambda = 0$, $c_\lambda = c = 1$						
Scaling	a_λ	Approx. regime	$n^* - \lambda/\mu$	# Correct	# Off by 1	# Off by 2
Efficiency	$a\lambda^{-1/2}$	Rationalized	$y^*(a\lambda^{-1/2})\sqrt{\lambda/\mu}$	19	1	0
Efficiency	$a\lambda^{-1/2}$	Quality	$y^*_\lambda\left(\frac{a}{\lambda}\right)\sqrt{\lambda/\mu}$	2	14	4
Quality	$a\sqrt{\lambda}$	Rationalized	$y^*\left(a\sqrt{\lambda}\right)\sqrt{\lambda/\mu}$	14	6	0
Quality	$a\sqrt{\lambda}$	Efficiency	$\sqrt{a/\mu}\lambda^{3/4}$	Poor – overstaffing by 10-15%		
Rationalized	a	Quality	$y^*_\lambda\left(\frac{a}{\sqrt{\lambda}}\right)\sqrt{\lambda/\mu}$	8	12	0
Rationalized	a	Efficiency	$\sqrt{a/\mu}\sqrt{\lambda}$	3	8	9

Table 3: Overview of the 'wrong-regime tests'

again, rounding off n^* to the nearest integer is almost always exact (16 out of the 20 times), and is never off by more than 1.

The above tests were run under favorable conditions: The asymptotic scaling of the parameters was known, and the approximation used was that associated with the regime corresponding to the parameter scaling. On the other hand, the results of the test are astoundingly good: Rounding n^* to the nearest integer is almost always exact, and is never off by more than 1. Note that these tests include values of λ that do not appear to be very large.

It is clear that more numerical testing is in order. There are two aims to this testing: 1) Find parameter values that 'break' the approximations, and 2) Determine if any of the asymptotic approximations is robust enough to work outside of its regime and/or determine rules of thumb for when each approximation should be used. (Indeed, this last point is central for obtaining a practically useful approximation.)

The additional testing takes the form of 'wrong-regime' testing. In all of these tests we set $c_\lambda = c = 1$, $\mu = 1$, $b_\lambda = 0$, and $a = 1$. We used all integer multiples of 10 between 10 and 200 for λ. The results of these tests are summarized in Table 3. The first 'wrong-regime' test we conducted involved scaling the parameters as in the efficiency-driven regime with $a_\lambda = a\lambda^{-1/2}$, and using the approximation from the rationalized regime. Thus our approximation rounded off

$$n^* = \lambda/\mu + y^*(a\lambda^{-1/2})\sqrt{\lambda/\mu}$$

to obtain the staffing level. The approximation was exact in all but one case ($\lambda = 160$), where it was off by 1.

Using the same parameters as in the preceding example we used the approximation from the quality-driven regime (as if $a_\lambda = a\sqrt{\lambda}$). Thus our approximation rounded off

$$n^* = \lambda/\mu + y^*_\lambda\left(\frac{a}{\lambda}\right)\sqrt{\lambda/\mu},$$

where y^*_λ is given by (43), to obtain the staffing level. The approximation was good, but not as good as the rationalized regime: it was off by 1 in 14 cases and off by 2 in 4 cases.

We next scaled the parameters as in the quality-driven regime with $a_\lambda = a\sqrt{\lambda}$, and used the approximation from the rationalized regime, rounding off

$$n^* = \lambda/\mu + y^*\left(a\sqrt{\lambda}\right)\sqrt{\lambda/\mu},$$

where $y^*(\cdot)$ is given by (40), to obtain the staffing level. The approximation was off by 1 in 6 cases and exact in the others.

Using the same parameters as in the preceding example we used the approximation from the efficiency-driven regime (as if $a_\lambda = a\lambda^{-1/2}$). Thus our approximation rounded off

$$n^* = \lambda/\mu + \sqrt{a/\mu}\lambda^{3/4}$$

to obtain the staffing level. This approximation did not perform well, leading to overstaffing of 10–15%.

We next scaled the parameters as in the rationalized regime, with $a_\lambda = a$, using the approximation from the quality-driven regime (as if $a_\lambda = a\sqrt{\lambda}$). Thus our approximation rounded off

$$n^* = \lambda/\mu + y^*_\lambda\left(\frac{a}{\sqrt{\lambda}}\right)\sqrt{\lambda/\mu},$$

where $y^*_\lambda(\cdot)$ is given by (43), to obtain the staffing level. The approximation here was exact in 8 cases and off by 1 in 12 cases.

Using the same parameters as in the preceding example we used the approximation from the efficiency-driven regime (as if $a_\lambda = a\lambda^{-1/2}$). Thus our approximation rounded off

$$n^* = \lambda/\mu + \sqrt{a/\mu}\sqrt{\lambda}$$

to obtain the staffing level. The approximation is not as good as that of the rationalized or quality-driven regime: 3 cases were exact, 8 were off by 1, and 9 were off by 2.

In the tests above that involve scaling parameters for the efficiency-driven and quality-driven regimes we used $a_\lambda = a\lambda^{-1/2}$ and $a_\lambda = a\sqrt{\lambda}$, respectively. These regimes hold more generally with $a_\lambda = a\lambda^{-\alpha}$ and $a_\lambda = a\lambda^{\alpha}$ respectively, for $\alpha > 0$. In addition to $\alpha = 1/2$, we also tried values of $\alpha = 1/4$ and 1. The runs for $\alpha = 1$ involved multiples of 50 from 50 to 1000 for λ. (The approximations used the same value of α as used to scale the actual parameters.) The results can be summarized as follows. The rationalized approximation was excellent: it was mostly exact, and when it was wrong it was typically off by 1 and never off by more than 3. The quality-driven approximation was good, but not as good as the rationalized approximation (even in the quality-driven regime!). The efficiency-driven solution was the worst of the three, and substantially overstaffed in the quality-driven regime. In the efficiency-driven regime with $\alpha = 1$, the efficiency-driven solution provided the infeasible solution of λ as the staffing level. Of course this can be corrected by simply requiring that the staffing level must be strictly greater than λ.

11 Future research

There are a few directions of research that suggest themselves. First, we would like to explain theoretically the extreme accuracy of our approximations. This cannot be anticipated from the corresponding asymptotic approximations. Next, the call center environment enjoys many features that are not captured by the M/M/N (Erlang-C) model. Important examples are customer abandonment, time-varying arrival rates, non-exponential service times, and multiple skill classes. The goal is to incorporate such features into our framework, which we now elaborate on.

A justification of the square-root safety staffing principle has been pursued at three levels: heuristic, conceptual and explicit. The *heuristic* level, as in [14], is based on an infinite-server approximation. The *conceptual* level, as in [9], is founded on a formal limit theorem, the limit taken as the number of servers increases with the offered load in a precise manner. These two levels motivate staffing levels of the form $N \approx R + y^* \sqrt{R}$. The *explicit* level, as in the present work, involves calculation of the constant y^* as a function of model parameters.
(i) For models with abandonment (Erlang-A perhaps would be an appropriate term), the conceptual level was analyzed in [6]. Notably, abandonment renders the queue always stable, hence y^* can also take negative values. The explicit level for Erlang-A is now under study by the present authors.
(ii) The heuristic approach for time-varying arrival rates proved successful in [12]. The corresponding conceptual level can be pursued within the service network framework of [15].
(iii) As for general service times, the M/G/N queue is not amenable to exact analysis, and letting $N \to \infty$ does not make things easier. Indeed, the accuracy of the standard multi-server approximations turns out questionable as N becomes large, which is the relevant regime for call centers. A key challenge is the calculation of the Halfin-Whitt delay function for the M/G/N queue. To this end, one could first attempt the M/PH/N queue, following [17].
(iv) A first study on staffing algorithms for multi-skill scenarios may be found in [2].

In the present paper we assumed that the offered traffic parameters are exactly known. We then focused on determining the amount of safety staffing needed to deal with stochastic variability only. In practice, however, the offered traffic forecasts are typically not completely accurate [13], and additional staffing may be required in view of the inherent uncertainty in the parameters. While the need for accurate forecasts becomes even more pronounced with the high-level of utilization advocated by the present work, we still feel that our analysis for known parameters constitutes an essential first step.
Lastly, one could perhaps use duality theory from Mathematical Programming to relate the optimization approach to constraint satisfaction. This could perhaps add insight on the optimality criteria for constraint satisfaction, which should be sharpened.

Acknowledgment The authors gratefully acknowledge useful comments from the anonymous referee and associate editor which helped improve the presentation of the results.

References

[1] Andrews, B. and Parsons, H. (1993). Establishing telephone-agent staffing levels through economic optimizations. *Interfaces* **23:2**, 14–20.

[2] Borst, S.C., Seri, P.F. (1997). Robust algorithms for staffing agents with multiple skills. Unpublished manuscript.

[3] De Bruijn, N.G. (1981). *Asymptotic Methods in Analysis.* Dover.

[4] Cooper, R.B. (1981). *Introduction to Queueing Theory.* 2nd Edition, North Holland.

[5] Call Center Statistics (WEB site), September 2000. *http://www.callcenternews.com/resources/statistics.shtml.*

[6] Garnet, O., Mandelbaum, A., Reiman, M.I. (1999). Designing a telephone call center with impatient customers. Submitted for publication in MSOM.

[7] Grassmann, W.K. (1986). Is the fact that the emperor wears no clothes a subject worthy of publication? *Interfaces* **16:2**, 43–51.

[8] Grassmann, W.K. (1988). Finding the right number of servers in real-world queueing systems. *Interfaces* **18:2**, 94–104.

[9] Halfin, S., Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**, 567–588.

[10] Jagers, A.A., Van Doorn, E.A. (1991). Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review* **33**, 281–282.

[11] Jagers, A.A., Van Doorn, E.A. (1986). On the continued Erlang loss function. *Oper. Res. Lett.* **5**, 43–46.

[12] Jennings, O.B., Mandelbaum, A., Massey, W.A., and Whitt, W. (1996). Service staffing to meet time-varying demand. *Mgmt. Sc.* **42**, 1383–1394.

[13] Jongbloed, G., Koole, G.M. (2000). Managing uncertainty in call centers using Poisson mixtures. Technical Report 2000-3, Department of Stochastics, Vrije Universiteit Amsterdam.

[14] Kolesar, P.J., Green, L.V. (1998). Insights on service system design from a normal approximation to Erlang's delay formula. *Prod. Oper. Mgmt.* **7**, 282–293.

[15] Mandelbaum, A., Massey, W.A., Reiman, M.I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* **30**, 149–201.

[16] Mandelbaum, A., Massey, W.A., Reiman, M.I., Rider, B., Stolyar, A. (2000). Queue length and waiting times for multiserver queues with abandonment and retrials. Selected Proceedings of the Fifth INFORMS Telecommunications Conference.

[17] Puhalskii, A.A., Reiman, M.I. (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* **32**, 564–595.

[18] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Mgmt. Sc.* **38**, 708–723.

A Proof of Lemma 5.2

Lemma 5.2
For any function x_λ with $\lim_{\lambda\to\infty} x_\lambda = \infty$,

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} Q_\lambda(x_\lambda).$$

If also $x_\lambda \overset{\sup}{\leq} \lambda^{1/6}$, then

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} Q(x_\lambda).$$

If specifically $x_\lambda = \kappa\sqrt{\lambda/\mu}$ for some constant $\kappa > 0$, then

$$\pi_\lambda(x_\lambda) \overset{\infty}{\approx} \frac{1}{\kappa\sqrt{2\pi\lambda/\mu(1+\kappa)}}\left(\frac{e^\kappa}{(1+\kappa)^{1+\kappa}}\right)^{\lambda/\mu}.$$

Proof
The first statement follows after some manipulations from the proof of Proposition 1 of Halfin & Whitt [9].
If $x_\lambda = \kappa\sqrt{\lambda/\mu}$ for some constant $\kappa > 0$, then $N_\lambda(x_\lambda) = (1+\kappa)\lambda/\mu$, $r_\lambda(x_\lambda) = 1/(1+\kappa)$, and $1 - r_\lambda(x_\lambda) = \kappa/(1+\kappa)$, so that

$$Q_\lambda(x_\lambda) = \frac{1}{\kappa\sqrt{2\pi\lambda/\mu(1+\kappa)}}\left(\frac{e^\kappa}{(1+\kappa)^{1+\kappa}}\right)^{\lambda/\mu}.$$

We now prove the third statement. Using the Taylor series expansion

$$\log u = \log(1-(1-u)) = -\sum_{m=1}^{\infty}\frac{(1-u)^m}{m} = -(1-u) - \sum_{m=2}^{\infty}\frac{(1-u)^m}{m},$$

we obtain

$$\frac{\exp\{N_\lambda(x)[1-r_\lambda(x)+\log r_\lambda(x)]\}}{\sqrt{2\pi N_\lambda(x)}(1-r_\lambda(x))} = \frac{\exp\left\{-N_\lambda(x)\sum_{m=2}^{\infty}\frac{(1-r_\lambda(x))^m}{m}\right\}}{\sqrt{2\pi N_\lambda(x)}(1-r_\lambda(x))} =$$

$$\frac{\exp\left\{-N_\lambda(x)(1-r_\lambda(x))^2/2 - N_\lambda(x)\sum_{m=3}^{\infty}\frac{(1-r_\lambda(x))^m}{m}\right\}}{\sqrt{2\pi N_\lambda(x)}(1-r_\lambda(x))} =$$

$$Q(x)\frac{x}{\sqrt{N_\lambda(x)}(1-r_\lambda(x))}\exp\left\{[x^2 - N_\lambda(x)(1-r_\lambda(x))^2]/2\right\}\exp\left\{-N_\lambda(x)\sum_{m=3}^{\infty}\frac{(1-r_\lambda(x))^m}{m}\right\}.$$

Thus it remains to be shown that

$$\frac{x_\lambda}{\sqrt{N_\lambda(x_\lambda)}(1-r_\lambda(x_\lambda))}\exp\{[x_\lambda^2 - N_\lambda(x_\lambda)(1-r_\lambda(x_\lambda))^2]/2\}\exp\{-N_\lambda(x_\lambda)\sum_{m=3}^{\infty}\frac{(1-r_\lambda(x_\lambda))^m}{m}\} \overset{\infty}{\approx} 1$$

if $x_\lambda \overset{\infty}{\ll} \lambda^{1/6}$.
Note that

$$\sqrt{N_\lambda(x_\lambda)}(1-r_\lambda(x_\lambda)) = \frac{x_\lambda\sqrt{\lambda/\mu}}{\sqrt{\lambda/\mu + x_\lambda\sqrt{\lambda/\mu}}} \overset{\infty}{\approx} x_\lambda,$$

and

$$N_\lambda(x_\lambda)(1-r_\lambda(x_\lambda))^2 = \frac{x_\lambda^2\lambda/\mu}{\lambda/\mu + x_\lambda\sqrt{\lambda/\mu}},$$

so that

$$0 \leq x_\lambda^2 - N_\lambda(x_\lambda)(1-r_\lambda(x_\lambda))^2 = \frac{x_\lambda^3\sqrt{\lambda/\mu}}{\lambda/\mu + x_\lambda\sqrt{\lambda/\mu}} \leq \frac{x_\lambda^3}{\sqrt{\lambda/\mu}} \overset{\infty}{\ll} 1.$$

Finally,

$$0 \leq N_\lambda(x_\lambda)\sum_{m=3}^{\infty}\frac{(1-r_\lambda(x_\lambda))^m}{m} \leq N_\lambda(x_\lambda)\sum_{m=3}^{\infty}(1-r_\lambda(x_\lambda))^m =$$

$$\frac{N_\lambda(x_\lambda)(1-r_\lambda(x_\lambda))^3}{r_\lambda(x_\lambda)} \leq \frac{x_\lambda^3}{\sqrt{\lambda/\mu}} \overset{\infty}{\ll} 1,$$

which completes the proof.

□

B Properties of $P(\cdot)$

Lemma B.1
The function $P(\cdot)$ is strictly convex decreasing.

Proof
The function $P(\cdot)$ may be written

$$P(x) = \frac{1}{1+U(x)},$$

with

$$U(x) := x\,\mathrm{e}^{x^2/2}V(x),$$

and

$$V(x) := \int_{-\infty}^{x} \mathrm{e}^{-y^2/2} dy.$$

Differentiating,

$$P'(x) = \frac{-U'(x)}{(1+U(x))^2},$$

and

$$P''(x) = \frac{2U'(x)^2 - U''(x)(1+U(x))}{(1+U(x))^3}.$$

Observing that $V'(x) = \mathrm{e}^{-x^2/2}$ and $V''(x) = -x\,\mathrm{e}^{-x^2/2}$,

$$U'(x) = x + (x^2+1)\,\mathrm{e}^{x^2/2} V(x),$$

and

$$U''(x) = x^2 + 2 + (x^3+3x)\,\mathrm{e}^{x^2/2} V(x).$$

Thus, $P(\cdot)$ is decreasing since $U'(x) > 0$ for any $x > 0$.
Also, $P(\cdot)$ is strictly convex because for any $x > 0$,

$$2U'(x)^2 - U''(x)(1+U(x)) =$$

$$2\left[x + (x^2+1)\,\mathrm{e}^{x^2/2}V(x)\right]^2 - \left[x^2 + 2 + (x^3+3x)\,\mathrm{e}^{x^2/2}V(x)\right]\left[1 + x\,\mathrm{e}^{x^2/2}V(x)\right] =$$

$$x^2 - 2 + \left[2x^3 - x\right]\,\mathrm{e}^{x^2/2}V(x) + \left[x^4 + x^2 + 2\right]\left[\mathrm{e}^{x^2/2}V(x)\right]^2 >$$

$$x^2 - 2 + \left[x^4 + 2x^3 + x^2 - x + 2\mathrm{e}^{x^2/2}V(x)\right]\,\mathrm{e}^{x^2/2}V(x) >$$

$$x^2 - 2 + \left[x^4 + 2x^3 + x^2 - x + 2\sqrt{\pi/2}\right]\,\mathrm{e}^{x^2/2}V(x) >$$

$$x^2 + \left[x^4 + 2x^3 + x^2 - x + 2(\sqrt{\pi/2} - 1)\right]\,\mathrm{e}^{x^2/2}V(x) >$$

$$\left[x^2 - x + 2(\sqrt{\pi/2} - 1)\right]\,\mathrm{e}^{x^2/2}V(x) =$$

$$\left[(x-1/2)^2 + 2(\sqrt{\pi/2} - 9/8)\right]\,\mathrm{e}^{x^2/2}V(x) > 0.$$

□

C Properties of $G_\lambda(\cdot)$

Lemma C.1
The function $G_\lambda(\cdot)$ is strictly convex decreasing.

Proof
It suffices to show that $K_\lambda(\cdot)$ is strictly convex decreasing with

$$K_\lambda(\omega) := \omega \int_0^\infty D_\lambda(t)\,\mathrm{e}^{-\omega t} dt.$$

Differentiating,

$$K'_\lambda(\omega) = \int_0^\infty [1-\omega t]\,\mathrm{e}^{-\omega t} D_\lambda(t) dt,$$

and

$$K''_\lambda(\omega) = \int_0^\infty [\omega t - 2]\,t\,\mathrm{e}^{-\omega t} D_\lambda(t) dt,$$

for all $\omega > 0$.
Since $D_\lambda(\cdot)$ is strictly increasing, we have

$$[1-\omega t] D_\lambda(t) \leq [1-\omega t] D_\lambda(\omega)$$

for all $\omega > 0$, $t > 0$, with *strict* inequality for $t \neq 1/\omega$, so that

$$\int_0^\infty [1-\omega t]\,\mathrm{e}^{-\omega t} D_\lambda(t) dt < D_\lambda(1/\omega) \int_0^\infty [1-\omega t]\,\mathrm{e}^{-\omega t} dt = 0,$$

and

$$\int_0^\infty [\omega t - 2]\,t\,\mathrm{e}^{-\omega t} D_\lambda(t) dt > D_\lambda(2/\omega) \int_0^\infty [\omega t - 2]\,t\,\mathrm{e}^{-\omega t} dt = 0,$$

for all $\omega > 0$.

$\square$