Motivated by call center practice, we study asymptotically optimal staffing of many-server queues with abandonment. A call center is modelled as an M/M/n+G queue, which is characterized by Poisson arrivals, exponential service times, n servers and generally distributed patience times of customers. Our asymptotic analysis is performed as the arrival rate, and hence the number of servers n, increase indefinitely.

We consider a constraint satisfaction problem, where one chooses the minimal staffing level n that adheres to a given cost constraint. The cost can incorporate the fraction abandoning, average wait and tail probabilities of wait. Depending on the cost, several operational regimes arise as asymptotically optimal: Efficiency-Driven (ED), Quality and Efficiency Driven (QED) and also a new ED+QED operational regime that enables QED tuning of the ED regime. Numerical experiments demonstrate that, over a wide range of system parameters, our approximations provide useful insight as well as excellent fit to exact optimal solutions. It turns out that the QED regime is preferable either for small-to-moderate call centers or for large call centers with relatively strict performance constraints. The other two regimes are more appropriate for large call centers with loose constraints.

We consider two versions of the constraint satisfaction problem. The main one is constraint satisfaction on a single time-interval, say one hour, which is common in practice. Of special interest is a constraint on the tail probability, in which case our new ED+QED staffing turns out asymptotically optimal. We also address a global constraint problem, say over a full day. Here several time intervals, say 24 hours, are considered, with interval-dependent staffing levels allowed; one seeks to minimize staffing levels, or more generally costs, given overall performance constraint. In this case, there is the added flexibility of trading service levels among time intervals, but we demonstrate that only little gain is associated with this flexibility.

Acknowledgements. The research of both authors was supported by BSF (Binational Science Foundation) grant 2001685/2005175, ISF (Israeli Science Foundation) grants 388/99, 126/02, 1046/04 and by the Technion funds for the promotion of research and sponsored research. This paper grew out of joint research with Sam Borst and Marty Reiman – their contribution and encouragement are greatly appreciated.
### Contents

1 **Introduction**  
   1.1 Taking abandonment into account: M/M/n+G and Erlang-A  
   1.2 Performance measures and types of constraints  
   1.3 Main contributions  
   1.4 Structure of the paper

2 **Asymptotic operational regimes: introduction and examples**  
   2.1 The QED (Quality and Efficiency-Driven) operational regime  
   2.2 The ED (Efficiency-Driven) operational regime  
   2.3 The ED+QED operational regime  
   2.4 Comparison between operational regimes  
   2.5 Global constraint satisfaction

3 **Related literature**

4 **Constraint satisfaction on a single interval**  
   4.1 General formulation of the problem  
   4.2 QED  
   4.3 ED  
   4.4 ED+QED

5 **Global constraint satisfaction**  
   5.1 Global constraint on the delay probability  
   5.2 Global constraint on the probability to abandon

6 **Numerical experiments**  
   6.1 QED approximations  
   6.2 ED approximations  
   6.3 ED+QED approximation for the tail probability  
   6.4 Comparison between operational regimes  
   6.5 Experiments on global constraint

7 **Proofs**  
   7.1 Proofs of Theorems 4.1-4.4  
   7.2 Proof of Theorem 5.1

8 **Possible future research**
1 Introduction

During the last two decades, one observes an explosive growth in the number of companies that provide services via the telephone, as well as in the variety of telephone services provided. According to some estimates, call center employees constitute 0.5% of the workforce in France, 1.5% in Great Britain and 4% in the USA [1] and worldwide expenditure on call centers exceeds $300 billion [19].

A central challenge in designing and managing a service operation in general, and a call center in particular, is to achieve a desired balance between operational efficiency and service quality. Here we consider the staffing aspects of this problem, namely having the right number of agents in place.

“The right number” means, first of all, not too many, thus avoiding overstaffing. That is a crucial consideration since personnel costs (e.g. salaries of operators and spending on training) typically constitute about 70% of a call center’s expenditure.

“The right number”, however, also means not too few, thus avoiding understaffing and consequent poor service quality. Indeed, understaffing would imply excessive customers’ wait in tele-queues which is unpleasant in itself and, moreover, is likely to lead to abandonment of frustrated customers. (According to a Purdue University study [13], 63% of the customers name a negative call center experience as their main reason for stopping transactions with a company.)

One could consider the following two approaches to the quality/efficiency tradeoff. The first one is widely used in practice. A manager specifies performance constraint(s) and then assigns the least staffing level that satisfies these constraints, over a pre-determined time interval. In the second approach, one assigns revenues to service completions and costs to delay factors such as wait and abandonment, as well as to staffing. The goal is then to identify the staffing level that maximizes profit. Borst, Mandelbaum and Reiman [7] pursued both approaches in the context of classical queues without abandonment (M/M/n, or Erlang-C). In the present paper we focus on the constraint satisfaction problem, with customers’ abandonment taken into account. A subsequent paper [27] will deal with the cost or revenue optimization problem.

1.1 Taking abandonment into account: M/M/n+G and Erlang-A

The M/M/n (Erlang-C) model was introduced by Erlang [14], the founder of queueing theory. It has been prevalent in call center applications for many years, being the mathematical engine of Workforce Management (WFM). Erlang-C assumes Poisson arrivals at a constant rate $\lambda$, exponentially distributed service times with rate $\mu$, and $n$ independent statistically-identical agents. Erlang-C implicitly assumes infinite patience of customers: all of them are willing to wait indefinitely until they get service (which implies system instability for $\frac{\lambda}{n\mu} \geq 1$).

However, an increasing number of present call centers incorporate customers’ abandonment
in their staffing/scheduling software and performance goals, and rightly so. In no little measure, this is a consequence of a growing body of research on queues with impatient customers, which has recently taken place in academia and industry. (See Gans et al. [17] and Zeltyn and Mandelbaum [40] for surveys.) It is the goal of the present paper to contribute to this research, by developing theory that both supports staffing practice and enhances our understanding of it.

Why take abandonment into account? In addition to the obvious factors, such as customers’ frustration and lost business, ignoring abandonment can imply a totally distorted picture of a call center’s operation. For example, consider a call center with 2,000 inbound calls per hour and average service time of 3 minutes. Assume that one uses the prevalent form of a service-level constraint “80% of customers should be served within 30 seconds”. Under the classical Erlang-C model, the call center would need 106 agents in order to reach this goal. However, if we assume that customers are willing to wait for service only 3 minutes on average and their patience is exponentially distributed, 95 agents (approximately, 10% reduction) would be enough. Moreover, if average patience is in fact only 1 minute, 84 agents would be able to provide the desired service level. We observe that, as impatience grows, a given service goal is reached with increasingly less agents; or, which is the same, a given staffing level yields smaller waiting times as impatience grows - those who abandon enable a higher, sometimes much higher, service level for those who “survive” the wait.

In our research, M/M/n is replaced by M/M/n+G, assigning to each inbound call a generally distributed patience time τ with a common distribution G (Figure 1). An arriving customer encounters an offered waiting time V, defined as the time that this customer would have to wait given that her patience is infinite. If the offered wait exceeds the customer’s patience time, the call abandons, otherwise the customer eventually gets service. In both cases, the actual waiting time W is equal to min(V, τ). Throughout the paper, we assume that serviced is rendered according to the First-Come-First-Served (FCFS) discipline.

Figure 1: Schematic representation of the M/M/n+G queue
M/M/n+G generalizes the M/M/n+M (Erlang-A) model, first introduced in Palm [28], which has exponentially distributed patience times. Erlang-A is the most computationally tractable model with abandonment (see [16] for a free software). And indeed, based on our experience, it is increasingly becoming the model of choice in support of WFM.

1.2 Performance measures and types of constraints

In order to apply a queueing model, one must first define relevant performance measures and then be able to calculate them. Moreover, since call centers can get very large (up to many thousands of agents), the implementation of these calculations must be scalable (numerically stable).

In this research, we accommodate the following three performance measures: the fraction of abandoning customers \( P\{\text{Ab} \} \), average wait in queue \( E[W] \) and the tail probability to exceed a deadline \( P\{W > T\} \).

The choice of the probability to abandon is natural in models with customers’ impatience. In addition to the reasons explained above, abandonment statistics constitutes the only commonly available measurement that is customer-subjective: those who abandon declare that the service offered is not worth its wait.

Average wait is also useful and taken into account in practice. Finally, the probabilities to exceed deadlines provide us with the distribution of customers’ wait. This and similar performance measures are widely used. Indeed, \( P\{W < T, \text{Sr}\} \), the fraction of served customers that wait less than \( T \) is often referred to in practice as service level.

The special case \( P\{W > 0\} \), the delay probability, is an important measure but it is rarely recorded in practice. Based on previous research [23, 18, 40], we suggest it as a useful robust measure of congestion that distinguishes between different operational regimes. For example, a delay probability close to 1 indicates an Efficiency-Driven (ED) operation, with necessarily very high utilization of agents. Alternatively, intermediate values of \( P\{W > 0\} \) (for example 0.5) indicate a careful balance between Quality and Efficiency, which we abbreviate to \( QED = \text{Quality & Efficiency Driven} \) operation; see [40] or Sections 2.1 and 4.2 for details.

Once an appropriate performance measure is chosen, a constraint satisfaction problem could be defined. The main approach of this paper is one that is common in the practice of call centers: constraint satisfaction on a single interval (Section 4). Specifically, first consider a time interval during which the staffing level is to be kept constant (15, 30 or 60 minutes), then specify a service-level constraint (for example, less than 3% of customers should abandon) and finally find the minimal staffing level \( n^* \) that guarantees this desired service-level in steady state. An alternative approach, treated in Section 5, considers several time intervals, for example a whole day, with different staffing levels allowed per interval. The goal now is to minimize the staffing levels, or, more generally, staffing costs, given an overall constraint on the performance level.
(For example, a constraint on the average wait over a full day of work which is divided into half-hour intervals.) In this case, an optimal solution is expected to compromise the service level at some intervals in order to do better at others.

1.3 Main contributions

As we view it, our main contributions to queueing theory and call center applications are as follows:

- Three operational regimes are studied within the M/M/n+G framework: Quality and Efficiency Driven (QED), Efficiency-Driven (ED) and also a new ED+QED operational regime; staffing in each regime arises as asymptotically optimal for a special case of cost constraints. Asymptotic statements on these regimes are formulated and proved in Theorems 4.1, 4.2 and 4.4, respectively.

- We elaborate on the new ED+QED operational regime: it is supported by Theorem 4.3 in Section 4.4, on the asymptotic behavior of its major performance measures.

- Practical recommendations on applications of the three operational regimes and corresponding approximations are provided. Quality of the approximations in different circumstances is compared. Their accuracy is found to be excellent. These recommendations are summarized in Section 2.4 and, then, substantiated and elaborated on in Section 6.

- We study a global constraint problem, where several time intervals with interval-dependent staffing levels are considered. It turns out that in some cases the structure of the global solution is different from the single-interval case (Section 5.1), and in others it is similar (QED staffing in Section 5.2). In the latter case, numerical experiments show that staffing flexibility per interval provides only little gain with respect to the single-interval approach.

- We extend the framework developed in Borst, Mandelbaum and Reiman [7] to cover abandonment and cost constraints. Our asymptotic analysis differs from [7], and it requires an extension of the Laplace method used in [27]. In this new framework, many further interesting questions can be addressed, for example, optimal staffing with respect to multiple service constraints.

1.4 Structure of the paper

Sections 2.1-2.3 contain detailed introduction to the three operational regimes, Section 2.4 compares between the regimes and Section 2.5 is dedicated to a preliminary discussion on global constraints. Section 3 surveys some related literature. Sections 4 and 5 present our theoretical
results for constraint satisfaction on a single interval and for global constraint satisfaction, respectively. Section 6 presents numerical experiments and Section 7 provides proofs. We conclude in Section 8 with some possible directions worthy of future research.

2 Asymptotic operational regimes: introduction and examples

How to solve the constraint satisfaction problem on a single interval? A straightforward approach is to apply exact formulae for performance measures of the $M/M/n+G$ queue, developed in Baccelli and Hebuterne [2], Brandt and Brandt [9, 10] and Zeltyn and Mandelbaum [40]. However, this approach has several drawbacks. These formulae for performance measures are relatively complicated, involving double integration of the patience distribution. They provide no intuition and give rise to numerical problems for large $n$ (number of servers). In addition, the calculations require the whole patience distribution but its estimation is typically a very complicated, sometimes impossible, task (see Brown et al. [11]).

In this research, an alternative approach is pursued. Depending on the structure of the cost function, several operational (staffing) regimes arise as asymptotically optimal. Each regime corresponds to a different approximate solution of the constraint satisfaction problem. These approximations are theoretically validated for large systems but they also provide excellent fit for moderate and even small ones. The final outcome are regime-specific staffing rules that are highly useful for call centers management.

The operational regimes are described in terms of the offered load parameter $R$, which is defined as

$$R = \frac{\lambda}{\mu} = \lambda \cdot E[S];$$

here $\lambda$ is the arrival rate and $\mu$ is the service rate, or the reciprocal of the average service time $E[S]$. (As customary in industry, we shall measure $R$ in units of Erlangs.) The quantity $R$ represents the amount of work, measured in time-units of service, that arrives to the system per unit of time. It is significant to the staffing problem since $R$ and its neighborhood provide nominal staffing levels, deviations from which could result in extreme performance: staffing high above $R$ would result in a very high quality of service, and staffing far below $R$ would result in a very high utilization of servers.

An important goal of our paper is to have these last statements quantified. To this end, we now present three operational regimes that arise in our research as asymptotically optimal. For each regime we give an example that illustrates its application to constraint satisfaction problems. We continue with comments on the quality of approximations based on these regimes and provide recommendations for their use. Finally, Section 2.5 provides a brief introduction to global constraint satisfaction.
2.1 The QED (Quality and Efficiency-Driven) operational regime

The QED regime corresponds to the least staffing level that adheres to a constraint, \( P\{W > 0\} \leq \alpha \), on the delay probability, given that \( \alpha \) is neither too close to 0 nor to 1. It is characterized by the so-called Square-Root Staffing Rule:

\[
n_{QED} = R + \beta \sqrt{R} + o(\sqrt{R}), \quad -\infty < \beta < \infty,
\]

(2.1)

where \( \beta \) is a Quality-of-Service (QoS) parameter – the larger it is, the better is the operational service-level. The rule was already described by Erlang [14], as early as 1924, with its formal analysis awaiting Halfin and Whitt [23] for Erlang-C. Recently, Garnett et al. [18] (Erlang-A) and Zeltyn and Mandelbaum [40] (M/M/n+G) studied QED queues with impatient customers. The QED regime enables one to combine high levels of efficiency (agents’ utilization over 90%) and service quality (short delays, scarce abandonment), given sufficient scale.

If we fix the service rate \( \mu \) and patience distribution \( G \) and let \( \lambda \) and hence \( n \) converge to infinity according to (2.1), the delay probability \( P\{W > 0\} \) converges to a constant strictly between 0 and 1. Also, the probability to abandon and average wait vanish at rate \( \frac{1}{\sqrt{n}} \). Formulae for different performance measures for queues with abandonment can be found in [18] and [40]. These approximations turn out to be very accurate even for small to moderate-size systems (few 10’s of agents).

Example 2.1 (Constraint satisfaction in the QED regime) Consider a moderate-size call center with an average of 20 arrivals per minute over a given time interval. Let the average service time be 3 minutes (\( \mu = 1/3 \)). Therefore, the offered load is equal to \( R = \lambda/\mu = 20 \cdot 3 = 60 \) Erlangs.

Assume that the customers of this call center constitute a 50-50% mixture of impatient customers with patience that is distributed \( \text{Exp(mean}=1) \) and patient ones with \( \text{Exp(mean}=5) \) patience. Formally, the distribution of overall patience is hyperexponential.

Suppose that the call center would like to maintain a high level of service. Three possible performance constraints are considered, unilaterally:

1. Probability to abandon should be less than 2%;
2. Average wait should be less than 5 seconds;
3. At least 90% of the customers should wait less than 20 seconds.

As explained in Section 2.4, the parameters of the problem (high service level, around \( R=60 \) agents needed) are appropriate for application of the QED regime (2.1). We thus calculate...
the optimal QoS parameter $\beta^*$ via equation (4.17) from Section 4.2. Then the approximately optimal staffing level is given by:

$$n_{QED}^* = \lceil R + \beta^* \sqrt{R} \rceil.$$  
(2.2)

The optimal staffing $n_{QED}^*$ for the three constraints above turns out to be 67, 62 and 61 agents, with the optimal QoS parameters $\beta^*$ being 0.79, 0.14 and 0.04, respectively.

If we calculate the optimal staffing via exact M/M/$n$+G formulae \cite{40}, it turns out that the fit is perfect: the exact optimal staffing $n^*$ is equal to $n_{QED}^*$ in all three cases.

How does the patience distribution affect the optimal staffing level? Maintaining a mean patience of 3 minutes, assume that patience times are now uniformly distributed between 0 and 6 minutes. (This could correspond to a situation when after 6 minutes of wait customers are routed to other location.) Then both $n^*$ and $n_{QED}^*$ for the three constraints are 64, 66 and 63, respectively, again a perfect fit. We observe that if the P\{Ab\} constraint is considered, more agents (67 vs. 64) are needed in the case of hyperexponential patience. However, if the wait in queue is controlled, the staffing level should be higher (63 vs. 61) for uniform patience.

As our theory reveals \cite{40}, the patience density near the origin is a key characteristic that determines performance of queues with a high service level. Higher density near the origin implies more abandonment and smaller wait. The tail of the patience distribution and even its mean are less important. One can check that, of the two distributions mentioned above, the hyperexponential has a higher density near the origin (3/5 vs. 1/6). Therefore, the staffing recommendations above (67 vs. 64) are consistent with \cite{40}.

The QED approximation (2.2) provides us with an additional important insight. Assume $\beta = 0$, which corresponds to the staffing level $n = R$: assign a number of agents that is equal to the offered load. It turns out that this deterministic “naive” approach can in fact yield good to very-good results. For example, in our case, if patience times are exponential with 3 minutes mean, one gets $P\{Ab\} \approx 5\%$, $E[W] \approx 9$ seconds and $P\{W < 20 \text{ sec}\} \approx 81\%$. For a large call center, performance would be much better than that, due to Economies-of-Scale. For example, if $\lambda = 200$, $n = R = 600$ and the other parameters are the same, $P\{Ab\} \approx 1.6\%$ and $E[W] \approx 3$ seconds. (In the QED regime, both $P\{Ab\}$ and $E[W]$ decrease at rate $1/\sqrt{\lambda}$.)

Note that the staffing rule $n = R$ is not feasible in queues without abandonment (e.g. Erlang-C) since its implementation would lead to instability of the system.

Summarizing, in this example we showed that:

- Staffing recommendations depend on the measure of performance to be controlled, as well as on the patience distribution beyond the mean;

- A common naive “deterministic” approach to staffing can yield good to very-good results, in the presence of abandonment.
2.2 The ED (Efficiency-Driven) operational regime

The ED regime corresponds to the least staffing level that adheres to a constraint on the fraction abandoning, or on the average wait, given that the constraints are relatively loose; for example, the former is to be above 10%, and the latter in the order of average service time. The ED regime is characterized by the staffing rule

\[ n_{ED} = (1 - \gamma) \cdot R + o(R), \quad \gamma > 0, \]  

(2.3)

which implies understaffing with respect to the offered load. In this case, as shown in [40], with \( n \) and \( \lambda \) increasing indefinitely, virtually all customers wait (\( P\{W > 0\} \) converges to 1), the probability to abandon always converges to \( \gamma \) and average wait converges to a constant that depends on the patience distribution. As a rule, ED approximations need relatively large \( n \) (more than 100) in order to provide a satisfactory fit; see the numerical experiments in [40, 41].

Example 2.2 (Constraint satisfaction in the ED regime) Consider a very large call center with 400 arrivals per minute, average service time 3 minutes, hence \( R = 1200 \). Assume a hyperexponential patience distribution with the parameters of Example 2.1. Assume that management has an efficiency-driven view of the call center operations: utilization of agents should be close to 100% but at the cost of a certain compromise on service level.

Two alternative performance constraints are considered here:

1. Probability to abandon should be less than 10%;
2. Average wait should be less than 20 seconds.

For these parameters (large number of agents, “loose” service-level constraints) it is reasonable to apply ED staffing:

\[ n^*_{ED} = \lceil (1 - \gamma^*) \cdot R \rceil, \]  

(2.4)

where the values of \( \gamma^* \) are established via equation (4.21) from Theorem 4.2. We get \( n^*_{ED} = 1,080 \) agents for the first constraint and 972 agents for the second one. The exact optimal solutions \( n^* \) are 1,081 and 972 agents, respectively.

Now consider a \( U(0, 6) \) patience distribution instead. In this case, our ED approximations prescribe 1,080 and 1,132 agents, and the exact solutions are 1,081 and 1,132. We observe the phenomenon that was mentioned above: staffing with respect to the \( P\{\text{Ab}\} \) constraint in the ED regime does not depend on the patience distribution. However, if average wait is controlled, the influence of the patience distribution can be very significant: 972 vs. 1,132 agents.

Note that, in this example, we used only two types of constraints, as opposed to three in Section 2.1: the constraint on the tail probability \( P\{W > T\} \) is not treated. The reason is that the ED regime does not provide an applicable approximation for the distribution of waiting time. We now present a refinement of the ED regime that enables such approximations.
2.3 The ED+QED operational regime

The following new operational regime combines the ED and QED staffing rules described above:

\[ n_{ED+QED} = (1 - \gamma) \cdot R + \delta \sqrt{R} + o(\sqrt{R}), \quad \gamma \geq 0. \]  

(2.5)

This ED+QED regime corresponds to the least staffing level that adheres to a constraint, \( P\{W > T\} \leq \alpha \), on the tail probability of delay, given that \( T \) is in the order of an average service time and \( \alpha \) is neither too close to 0 nor to 1. In words, we apply ED staffing and, then, fine-tune according to the square-root QED rule.

Why is a new staffing regime needed? Assume that we vary the number of servers according to the ED staffing rule (2.3) fixing other system parameters. It follows from [40] that there exists an ED parameter \( \gamma^* \) such that the tail probability \( P\{W > T\} \) converges to zero for \( \gamma < \gamma^* \) and to \( 1 - G(T) \) for \( \gamma > \gamma^* \). The ED approximation is thus too “crude” for the constraint satisfaction problem \( P\{W > T\} \leq \alpha \). However, it turns out that QED fine-tuning around the ED staffing level \( (1 - \gamma^*)R \), taking into account \( \alpha \), provides one with the proper \( \delta^* \), thus determining \( n_{ED+QED}^* \). See Section 4.4 and Theorems 4.3 and 4.4 for rigorous results on the ED+QED regime and their clarification.

Example 2.3 (Constraint satisfaction in the ED+QED regime) Consider the large call center from Example 2.2 with offered load \( R = 1,200 \). Assume the commonly-used service-level constraint: “at least 80% of the customers should wait less than 20 seconds”. Consider three patience distributions with the same mean: \( \text{Exp}(\text{mean}=3) \), \( U(0, 6) \) and our previous hyperexponential mixture of \( \text{Exp}(\text{mean}=1) \) and \( \text{Exp}(\text{mean}=5) \). Applying the staffing formula

\[ n_{ED+QED}^* = \lceil (1 - \gamma^*) \cdot R + \delta^* \sqrt{R} \rceil, \]

with values of \( \gamma^* \) and \( \delta^* \) derived from Theorem 4.4, we get \( n_{ED+QED}^* \) equal to 1,099, 1,153 and 1,020, respectively. (The exact optimal values are 1,100, 1,153 and 1,021.) Theorem 4.4 implies that the \( \gamma^* \)'s are different for the three distributions, hence the large variations in staffing levels. One concludes that the use of the exponential assumption on patience (the Erlang-A model), which is slowly becoming standard in call centers, can imply significant deviations from the optimum, under some circumstances.

Another important insight from Examples 2.2 and 2.3 is that a reasonable service level and beyond can be reached even if significant understaffing with respect to \( R \) takes place – given sufficient scale.

Remark 2.1 (ED+QED regime in Erlang-A) Assume that the patience times are exponential with rate \( \theta \), namely mean \( 1/\theta \). (According to our experience, typical mean patience is...
around 5-10 minutes.) Fix the deadline to be $c\%$ of the patience mean: $T = c/\theta$, for some $0 < c < 1$. Then Theorem 4.4 implies that the ED parameter is equal to

$$\gamma = G(T) = 1 - e^{-\theta T} = 1 - e^{-c} \approx c,$$

where the last approximation is reasonable for $c$ around 0.1. In this case, according to the ED approximation

$$P\{\text{Ab}\} \approx \gamma \approx c.$$

This argument provides us with a useful rule-of-thumb for Erlang-A: a deadline $T$ of around 10\% of the average patience (namely $T$ between 30 seconds and 1 minute) corresponds to approximately 10\% abandonment.

**Remark 2.2 (Quality-Driven regime)** Zeltyn and Mandelbaum [40] also analyzed the QD (Quality-Driven) operational regime, with staffing levels

$$n_{QD} = \lceil (1 + \gamma) \cdot R + o(\sqrt{R}) \rceil, \quad \gamma > 0.$$

In this regime, the main performance characteristics, namely $P\{\text{Ab}\}$, $E[W]$ and $P\{W > 0\}$, converge to zero exponentially fast, as $\lambda, n \to \infty$. A similar regime was discussed also in Borst et al. [7] for Erlang-C. The QD regime can become relevant for extreme constraints on service level, say $P\{W > 0\} \leq 2\%$, which are appropriate, for example, in amply-staffed emergency operations. In this paper, we are interested in queues with non-negligible wait and abandonment, hence the QD regime will not be considered.

### 2.4 Comparison between operational regimes

In Section 2 we introduced three operational regimes and presented examples where these regimes can be successfully implemented. Note that for a given constraint satisfaction problem one can fit to it several different approximations, based on the operational regimes from Sections 2.1-2.3.

A natural question arises as to the existence of a single operational regime that is preferable over the two others, at least for all constraint satisfaction problems of practical interest. For Erlang-C, Borst et al. [7] discovered that the QED regime qualifies: it is extremely robust and can be applied for a very wide range of system parameters so as to render the ED and QD regimes almost practically useless. In Section 6 we analyze this question for M/M/n+G. (For example, Section 6.4 compares the QED, ED and ED+QED regimes using the setup of Examples 2.1-2.3 above.) It turns out that, in contrast to Erlang-C, there is no best single operational regime.

Specifically, the QED approximations work very well either for small-to-moderate call centers or for large call centers with strict performance constraints. Note that in both cases the optimal staffing level is not too far from the offered load.
However, in the case of significant understaffing, which arises with loose constraints (for example, average wait of more than 10% average service time) and large offered load (several hundreds of Erlangs and more), the situation is more complicated. If a constraint on the probability to abandon is considered, the QED approximations are excellent. If one controls average wait, QED staffing is very precise for exponential patience but could be far from optimal for other distributions, for example uniform and hyperexponential. Therefore, ED staffing should be applied for non-exponential patience. Finally, if a constraint on \( P\{W > T\} \) should be satisfied, the ED+QED approximations turn out preferable.

2.5 Global constraint satisfaction

In Sections 2.1-2.4 we described our approach for constraint satisfaction on a single time-interval. In Section 5 we shall study approximate solutions of some global constraint problems where one seeks to minimize staffing costs over several time intervals, with differing staffing levels allowed per interval. If the staffing costs are equal at all intervals, the problem reduces to minimizing the total staffing level. We then expect a smaller overall number of servers relative to interval-by-interval optimization, due to the added flexibility of trading service levels among intervals.

Figure 2: **Comparison between local and global constraints.**

![Figure 2: Comparison between local and global constraints.](image)

Example 2.4 (Comparison between global and local constraint satisfaction) Consider a weekday arrival pattern to the call center of some Israeli cellular company (left plot in Figure 2). This pattern is rather typical in call centers; see for example Brown et al. [11] and Green et al. [22]. Assume a constant arrival rate during one-hour intervals. The average service time is
equal to 218 seconds. Mean patience is taken to be 6 minutes (a reasonable number, according to our experience in data analysis; see, for example [11]). Assume that both service and patience times are exponential.

Consider the constraint $P\{\text{Ab}\} \leq 1\%$. We compare QED staffing (2.2) that seeks to sustain this service level over each one-hour interval, against QED staffing that adheres to a global daily (24 hours) service level. The latter staffing levels are derived via Theorem 5.2.

Overall, the two staffing levels are rather close. (Their pattern is also indistinguishable from the pattern of the arrival rate in Figure 2.) The right plot of Figure 2 presents their difference: the global staffing assigns more servers to heavily loaded intervals and less to lightly loaded. Overall, using global constraint saves only 9 work-hours (out of more than 3,000 hours over the day).

Figure 3 demonstrates the dynamics of $P\{\text{Ab}\}$ over the day for the two staffing strategies. ($P\{\text{Ab}\}$ is calculated via exact M/M/$n$+G formulae [40].) As expected, we observe a stable pattern in the first plot. In the second plot, however, a deterioration of the service level at night takes place (which, indeed, conforms to our personal experience with the way that many 24-hour call centers are run). Since minor savings in the number-of-agents probably do not justify instability of the service level, the staffing level that arises from the hourly constraint seems preferable.

![Figure 3: Dynamics of the probability to abandon.](image-url)
3 Related literature

For a comprehensive summary on queues with impatient customers, operational regimes and dimensioning, readers are referred to Sections 4.1 and 4.2 in Gans et al. [17]. Here we summarize research that is most relevant to ours.

**Queues with impatient customers.** The seminal work on queueing systems with impatient customers is Palm [28], where he introduced the basic Erlang-A (M/M/n+M) queue with exponential patience times. See Mandelbaum and Zeltyn [26] for a recent summary of this model. Erlang-A was generalized to M/M/n+G (General patience) by Baccelli and Hebuterne [2], Brandt and Brandt [9, 10] and Zeltyn and Mandelbaum [40]. In the present research we adopt the theoretical approach of [2] and [40] to the M/M/n+G model.

If the service distribution is not exponential, as is often the case in practice (see Brown et al. [11]), exact theoretical solutions are not available and one has to resort to approximations and simulation. In addition to ED approximations, as discussed below, one should mention the papers of Whitt [37, 38] that develop and validate an approximation for the M/G/n+G model with generally distributed iid service times. Finally, Boxma and de Waal [8] addressed the problem of cost optimization in the M/G/n+G queue via simulation and interpolation between M/M/n+D and Erlang-A.

**Remark 3.1 (Types of approximations)** Below we survey several types of approximations to the queueing systems that arise in our research. One distinguishes between two main types of approximations: *steady-state* (asymptotic expressions for steady-state performance measures like $P\{Ab\}$ or $E[W]$) and *process-limit* (asymptotics for stochastic processes such as the queue-length process). In our paper, we are mainly interested in steady-state approximations. However, many papers referred to below present also process-limit approximations.

**QED operational regime.** As mentioned above, the square-root staffing rule (2.1) was first introduced by Erlang [14]. He reports that it had in fact been in use at the Copenhagen Telephone Company since 1913. A formal analysis for the Erlang-C queue appeared only in 1981, in the seminal paper of Halfin and Whitt [23]. In that paper, the authors establish an important relation: as $\lambda$ increase indefinitely, sustaining the QED operational regime (2.1) with fixed $\beta > 0$ is equivalent to the delay probability converging to a fixed level $\alpha$, $0 < \alpha < 1$. Whitt [33] surveys QED approximations for various classical queues without abandonment.

Garnett, Mandelbaum and Reiman [18] studied the QED regime for Erlang-A with exponential abandonment, establishing results that are analogous to [23] and complemented also by the ED and QD regimes. Zeltyn and Mandelbaum [40] presented a comprehensive treatment of the QED, ED and QD regimes in steady-state for the M/M/n+G queue.
ED operational regime. ED approximations are cruder than the QED approximations, hence they enable the analysis of very general models. For example, Whitt [34] presents a general fluid model (the ED approximation) for the G/G/n+G queue with general distributions of arrivals, services and patience times and Whitt [39] presents a multi-class fluid model that takes skills-based routing into account.

Another important family of models that can be treated in the ED regime are those with uncertainty about the arrival-rate. (Whitt [35] showed that Erlang-A and other queues with abandonment are sensitive to changes in the arrival rate.) Recent papers of Whitt [36] and Harrison and Zeevi [24] study ED approximations for such models and develop asymptotic rules for optimal staffing. In addition, Bassamboo, Harrison and Zeevi [4, 5] provide asymptotic methods of routing and admission control.

ED+QED operational regime. The only reference to this regime that is known to us is Baron and Milner [3]. That work is motivated by Service Level Agreements that arise in outsourcing contracts. It includes an approximation for the tail probability of wait in Erlang-A. This approximation is a special case of our approximation for M/M/n+G, which is covered in Theorem 4.4.

It is worth mentioning that the QED staffing rule that arose in models of membership (subscriber) services is (only formally) similar to our ED+QED staffing rule (2.5); see Randhawa and Kumar [29, 30] and de Véricourt and Jennings [31].

Dimensioning Erlang-C: cost optimization and constraint satisfaction. Borst, Mandelbaum and Reiman [7] developed a mathematical framework for the problem of optimal staffing in the Erlang-C queue. The main focus of the paper is on cost optimization with convex staffing costs and general increasing waiting costs. Depending on the relative importance of these costs, [7] identifies the QED, ED and QD regimes as asymptotically optimal. It is shown that the QED regime balances and, in fact, unifies the other two regimes. In the case of linear costs, a relation between the waiting/staffing costs ratio and the QoS parameter $\beta$ from (2.1) is established. In addition, the constraint satisfaction problem is also analyzed, with the QED, ED and QD regimes arising as well.

Global constraint satisfaction. In [7], the problem of optimal staffing was studied on a single interval in steady-state, as conventionally assumed in the literature and practice. Koole and van der Sluis [25] analyze a shift scheduling problem with overall, say daily, service-level objective. They prove a useful property, called multimodularity, which, if prevailing, significantly facilitates the search for the exact optimal staffing levels.
SIPP staffing. Both the single- and multiple-interval approaches can be viewed as the SIPP (Stationary Independent Period by Period) method for staffing; see Green, Kolesar and Soares [20]. Advantages, drawbacks and possible modifications of SIPP were studied in Green, Kolesar and Soares [20, 21] and Green, Kolesar and Whitt [22]. Overall, SIPP works well if the arrival rate is slow-varying with respect to the durations of services. Otherwise, time-dependent models should be used; see [20, 21, 22] and Feldman et al. [15].

4 Constraint satisfaction on a single interval

4.1 General formulation of the problem

We analyze the M/M/n+G queue with a fixed service rate \( \mu \) and patience distribution \( G \). Let the arrival rate \( \lambda \to \infty \). We determine staffing levels (number of agents) \( n_{\lambda}, \lambda \geq 0 \), as a function of \( \lambda \).

Consider a customer with patience time \( \tau \), offered wait \( V \) and hence actual wait \( W = \min(\tau, V) \). The disutility function \( D_{\lambda}(\tau, V) \) of this customer is taken to be

\[
D_{\lambda}(\tau, V) = C_{ab}(\lambda) \cdot I_{\{\tau < V\}} + C_w(\lambda) \cdot W + C_b(\lambda) \cdot I_{\{W > d_{\lambda}\}}.
\]

The coefficients \( C_{ab}(\lambda) \), \( C_w(\lambda) \) and \( C_b(\lambda) \) in (4.1) are abandonment cost, waiting cost and deadline cost, respectively. The deadline \( d_{\lambda} \) can depend on the arrival rate. Then, the expected disutility per customer is given by

\[
U(n, \lambda) \triangleq E_{n,\lambda}[D(\tau, V)] = C_{ab}(\lambda) \cdot P_{n,\lambda}\{\text{Ab}\} + C_w(\lambda) \cdot E_{n,\lambda}[W] + C_b(\lambda) \cdot P_{n,\lambda}\{W > d_{\lambda}\}.
\]

Define the optimal staffing level by

\[
n_{\lambda}^* = \arg \min_n \{\lambda \cdot U(n, \lambda) \leq M_{\lambda}\},
\]

where \( M_{\lambda} \) is a cost constraint. In this research, we make a natural assumption that the cost constraint is proportional to the number of customers, i.e. linear in \( \lambda \): \( M_{\lambda} = M \cdot \lambda, \ M > 0 \), which allows one to replace (4.3) by

\[
n_{\lambda}^* = \arg \min_n \{U(n, \lambda) \leq M\}.
\]

Since we study different types of asymptotic solutions to (4.4), the following two additional definitions turn out natural and useful. The staffing level \( n_{\lambda} \) is called asymptotically feasible if

\[
\limsup_{\lambda} U(n_{\lambda}, \lambda) \leq M.
\]
In addition, \( n_\lambda \) is asymptotically optimal if

\[
|n_\lambda^* - n_\lambda| = o(f(\lambda)) = o(R), \quad (4.6)
\]

where a specific function for \( f(\cdot) \) will be defined separately and naturally for every special case. (For example, it can be equal to \( \lambda, \sqrt{\lambda} \) etc.)

For general \( C_{ab}(\cdot), C_w(\cdot) \) or \( C_b(\cdot) \), the problem (4.4) is rather complicated and uninsightful. Instead, we explore three basic important special cases that give rise to the three operational regimes introduced in Section 2.

**Remark 4.1** In this research, the staffing level \( n \) depends on the arrival rate \( \lambda \) and, consequently, performance measures depend on both \( \lambda \) and \( n \). For simplicity of notation, in the rest of the paper we shall omit indices that correspond to \( \lambda \) and \( n \).

### 4.2 QED

Here we explore types of constraints that give rise to the QED regime, discussed in Section 2.1. Recall that this regime allows to combine efficiency (high servers’ utilization) and service quality.

Assume that the patience density at the origin exists and is positive: \( G'(0) \Delta g_0 > 0 \). Define the following functions:

\[
P_w(\beta) \triangleq \left[ 1 + \frac{g_0}{\mu} \cdot \frac{h_\phi(\hat{\beta})}{h_\phi(-\hat{\beta})} \right]^{-1}, \quad -\infty < \beta < \infty, \quad (4.7)
\]

\[
P_\alpha(\beta) \triangleq \sqrt{g_0} \cdot (h_\phi(\hat{\beta}) - \hat{\beta}), \quad -\infty < \beta < \infty, \quad (4.8)
\]

\[
W(\beta, t) \triangleq \frac{\Phi(\hat{\beta} + \sqrt{g_0} \cdot t)}{\Phi(\hat{\beta})}, \quad -\infty < \beta < \infty, \quad t \geq 0, \quad (4.9)
\]

where

\[
\hat{\beta} \triangleq \beta \sqrt{\frac{\mu}{g_0}}, \quad (4.10)
\]

and

\[
h_\phi(x) = \frac{\phi(x)}{1 - \Phi(x)} = \frac{\phi(x)}{\Phi(x)}
\]

is the hazard rate of the standard normal distribution. (\( \Phi(x) \) is the standard normal cumulative distribution function, \( \Phi(x) = 1 - \Phi(x) \) is the survival function and \( \phi(x) = \Phi'(x) \) is the density.)

Theorem 4.1 from Zeltyn and Mandelbaum [40] implies that under the QED scaling, as defined by the square-root staffing rule (2.1) with \( \lambda \rightarrow \infty \) (or \( R = \lambda/\mu \rightarrow \infty \)), the following approximations prevail:

\[
P\{W > 0\} \sim P_w(\beta), \quad (4.11)
\]
\begin{align*}
\Pr \{\text{Ab} \} &= \frac{1}{\sqrt{\lambda}} \cdot P_a(\beta) P_w(\beta) + o\left(\frac{1}{\sqrt{\lambda}}\right), \\
\mathbb{E}[W] &= \frac{1}{\sqrt{\lambda}} \cdot \frac{1}{g_0} \cdot P_a(\beta) P_w(\beta) + o\left(\frac{1}{\sqrt{\lambda}}\right), \\
\Pr \left\{ W > \frac{t}{\sqrt{\lambda}} \right\} &\sim W(\beta, t) \cdot P_w(\beta), \quad t \geq 0.
\end{align*}
\hspace{1cm} (4.12, 4.13, 4.14)

(Here and later \( f \sim g \) stands for \( \lim_{\lambda \to \infty} f(\lambda)/g(\lambda) = 1 \).)

**Theorem 4.1 (QED)** Assume that the cost function (4.2) is given by
\begin{equation}
U(n, \lambda) = C_{ab} \cdot \sqrt{\lambda} \cdot \Pr \{\text{Ab} \} + C_w \cdot \sqrt{\lambda} \cdot \mathbb{E}[W] + C_b \cdot \Pr \left\{ W > \frac{t}{\sqrt{\lambda}} \right\},
\end{equation}
where the constants \( C_{ab}, C_w, C_b, t \) are non-negative. Assume that either \( C_{ab} > 0 \) or \( C_w > 0 \), or the cost constraint \( M < C_b \).

a. The optimal staffing level (4.4) satisfies
\begin{equation}
n^* = R + \beta^* \sqrt{R} + o(\sqrt{R}),
\end{equation}
where \( \beta^* \) is the unique solution of the following equation with respect to \( \beta \):
\begin{equation}
\left\{ \left( \frac{C_{ab} + C_w}{g_0} \right) \cdot P_a(\beta) + C_b \cdot W(\beta, t) \right\} \cdot P_w(\beta) = M.
\end{equation}

b. Introduce the staffing level
\begin{equation}
n_{QED}^* = \left\lceil R + \beta^* \sqrt{R} \right\rceil.
\end{equation}
Then the staffing level \( n_{QED}^* \) is asymptotically feasible (4.5) and asymptotically optimal (4.6) in the sense that
\begin{equation}
|n_{QED}^* - n^*| = o(\sqrt{R}).
\end{equation}

Theorem 4.1 is proved in Section 7.1.

**Remark 4.2 (On the solution of the equation (4.17))** All functions in (4.17) are monotone. (See the proof of Theorem 4.1 in Section 7.1 for an elaboration.) Therefore, its solution is tantamount to the calculation of the inverse of the left-hand side.

**Remark 4.3 (On \( \sqrt{\lambda} \) scaling of the probability to abandon and waiting time)** In Theorem 4.1 we scale some parameters of the problem by \( \sqrt{\lambda} \). In call center practice, on the other hand, one works with unscaled constraints. The translation between the two types of the constraints is straightforward. Assume, for example, that the service-level constraint is given by \( \Pr \{\text{Ab} \} \leq \epsilon \). According to (4.12), we should first solve
\begin{equation}
P_a(\beta) P_w(\beta) = \epsilon \sqrt{\lambda}
\end{equation}
with respect to \( \beta \) and then apply the staffing level (4.18) with the solution \( \beta^* \).
Remark 4.4 (Two versions of QED approximations) In Zeltyn and Mandelbaum [40], the asymptotic results (4.12)-(4.14) were formulated in a slightly different way using $1/\sqrt{n}$ scaling. For example, the approximation for $P\{Ab\}$ in [40] was given by

$$P\{Ab\} = \frac{1}{\sqrt{n}} \cdot \tilde{P}_a(\beta) P_w(\beta) + o\left(\frac{1}{\sqrt{n}}\right),$$

(4.19)

where

$$\tilde{P}_a(\beta) \triangleq \sqrt{\frac{g}{\mu}} \cdot (h_\phi(\hat{\beta}) - \hat{\beta}).$$

In the QED regime, $\lambda \sim n\mu$, therefore the two versions of approximations are asymptotically equivalent. However, the version presented in this paper is preferable over its [40] counterpart if we apply it in highly overloaded systems, conventionally described by the ED regime. See Section 6.2 for an elaboration.

4.3 ED

Here we study the ED regime defined in (2.3). It is characterized by significant understaffing that gives rise to very high servers’ utilization and only moderate service level.

Assume that the patience distribution function $G$ is strictly increasing for all $x$ such that $0 < G(x) < 1$. Define $H(x) = \int_0^x \bar{G}(u) du$, where $\bar{G}(\cdot) = 1 - G(\cdot)$ is the survival function of patience.

Theorem 6.1 from Zeltyn and Mandelbaum [40] states that in the ED operational regime, defined by (2.3) and $\lambda \to \infty$, the probability to abandon and average wait converge to $\gamma$ and $H(G^{-1}(\gamma))$, respectively.

Theorem 4.2 (ED) Assume that the cost function (4.2) is given by

$$U(n, \lambda) = C_{ab} \cdot P\{Ab\} + C_w \cdot E[W],$$

(4.20)

where the constants $C_{ab}, C_w$ are non-negative. Assume that the positive cost constraint $M$ is less than $C_{ab} + C_w \cdot E[\tau]$, where $E[\tau] = \infty$ is allowed. Then

a. The optimal staffing level (4.4) satisfies

$$n^* = (1 - \gamma^*) \cdot R + o(R),$$

where $\gamma^*$ is the unique solution of the following equation with respect to $\gamma$:

$$C_{ab} \cdot \gamma + C_w \cdot H(G^{-1}(\gamma)) = M.$$

(4.21)

b. Introduce the staffing level

$$n_{ED}^* = \lceil (1 - \gamma^*) \cdot R \rceil.$$
Then \( n_{ED}^* \) is asymptotically feasible (4.5) and asymptotically optimal (4.6) in the sense that

\[
|n_{ED}^* - n^*| = o(R) .
\]

See Section 7.1 for the proof of Theorem 4.2.

**Remark 4.5** With each one of the cost coefficients in (4.20) vanishing, we get two important special cases: a constraint on \( P\{Ab\} \) and a constraint on \( E[W] \). Note that, in the first case, \( \gamma = M/C_{ab} \) and the ED staffing \( n_{ED}^* \) does not depend on the patience distribution. In contrast, in the second case one can observe a very significant dependence of \( n_{ED}^* \) on \( G \); see Example 2.2 in the Introduction and Section 6.2.

### 4.4 ED+QED

Now we study the new operational regime, introduced in Section 2.3. This regime combines ED and QED staffing, which arises when one seeks to satisfy a constraint on the tail probability of wait, namely

\[
P\{W > T\} \leq \alpha . \tag{4.22}
\]

First, we recall from [40] that in the ED regime (2.3), the waiting time converges weakly to \( \min(\tau, G^{-1}(\gamma)) \), where \( \tau \) is the patience time and \( G \) is the patience distribution. This suggests the following approximation for the tail probability \( P\{W > T\} \):

\[
P\{W > T\} \approx \begin{cases} 
\bar{G}(T), & T < G^{-1}(\gamma), \\
0, & T > G^{-1}(\gamma).
\end{cases} \tag{4.23}
\]

In (4.23) we assume that \( \gamma \) is fixed and \( T \) is varied. But to identify the least staffing level that adheres to (4.22), we view (4.23) as a function of \( \gamma \):

\[
P\{W > T\} \approx \begin{cases} 
\bar{G}(T), & \gamma > G(T), \\
0, & \gamma < G(T),
\end{cases} \tag{4.24}
\]

which is too crude to capture \( \alpha \) in (4.22). Hence one must refine (4.24) around \( \gamma = G(T) \). To this end, introduce ED+QED staffing with the ED parameter \( \gamma = G(T) \) as follows:

\[
n = (1 - G(T)) \cdot R + \delta \sqrt{R} + o(\sqrt{R}) , \quad -\infty < \delta < \infty . \tag{4.25}
\]

The next theorem enables the calculation of \( \delta \) that corresponds to the target level \( \alpha \) of the tail probability. It also presents approximations for other key performance measures in the ED+QED regime. The theorem is formulated in the spirit of the M/M/n statement in Halfin and Whitt [23].
Theorem 4.3 (ED+QED performance measures) Consider a sequence of M/M/n+G queues indexed by $n = 1, 2, \ldots$, with fixed service rate $\mu$ and patience distribution $G$. Let $T$ and $\alpha$ be scalars such that $0 < T < \infty$, $0 < \alpha < \bar{G}(T)$ and the patience density $g(T) = G'(T) > 0$. Then the following four asymptotic statements are equivalent, as $n \to \infty$ (and hence $\lambda \to \infty$ and $R \to \infty$):

1. **Staffing level:** $n = (1 - \gamma)R + \delta \sqrt{R} + o(\sqrt{R})$;
2. **Tail probability:** $P\{W > T\} = \alpha + o(1)$;
3. **Probability to abandon:** $P\{\text{Ab}\} = \gamma - \delta \sqrt{R} + o(1)$;
4. **Average wait:** $E[W] = \int_0^T \bar{G}(u)du - \frac{\delta}{\sqrt{R}} \frac{1}{h_G(T)} + o(1)$.

Here $h_G(T) = g(T)/\bar{G}(T)$ is the hazard rate of the patience distribution $G$, $\gamma = \bar{G}(T)$ and

$$\delta = \Phi^{-1}\left(\frac{\alpha}{\bar{G}(T)}\right) \cdot \sqrt{\frac{g(T)}{\mu}}. \tag{4.26}$$

**Remark 4.6** Note that (4.26) and Statement 2 of Theorem 4.3 imply the following approximation for the delay probability under the ED+QED staffing level (4.25):

$$P\{W > T\} \sim \bar{G}(T) \cdot \Phi\left(\delta \sqrt{\frac{\mu}{g(T)}}\right).$$

**Remark 4.7** If the constraint parameter $\alpha \geq \bar{G}(T)$, then for any staffing level $n$

$$P\{W > T\} \leq \bar{G}(T) \leq \alpha, \tag{4.27}$$

since the waiting time $W$ does not exceed the patience time $\tau$. Hence, $\alpha \geq \bar{G}(T)$ cannot be attained as a limit in Part 2, and $\alpha = \bar{G}(T)$ can be attained even if $n = 0$, namely service is not provided at all.

Now we can formulate the constraint satisfaction result for the ED+QED regime.

**Theorem 4.4 (ED+QED constraint satisfaction)** Assume that the cost function (4.2) is given by

$$U(n, \lambda) = C_b \cdot P\{W > T\}, \quad C_b > 0, \quad T > 0, \tag{4.28}$$

and that the patience density at $T$ is positive: $g(T) > 0$. The optimization problem (4.4) then takes the form

$$n^* = \arg\min_n \{P\{W > T\} \leq \alpha\},$$
where $\alpha = M/C_b$. Assume that $\alpha < \bar{G}(T)$. Then

a. The optimal staffing level (4.4) satisfies
\[
n^* = \bar{G}(T) \cdot R + \delta^* \sqrt{R} + o(\sqrt{R}) ,
\]
where
\[
\delta^* = \Phi^{-1} \left( \frac{\alpha}{\bar{G}(T)} \right) \cdot \sqrt{\frac{g(T)}{\mu}} .
\] (4.30)

b. Introduce the staffing level
\[
n_{ED+QED}^* = \left\lceil \bar{G}(T) \cdot R + \delta^* \sqrt{R} \right\rceil .
\]
Then $n_{ED+QED}^*$ is asymptotically feasible (4.5) and asymptotically optimal (4.6) in the sense that
\[
|n_{ED+QED}^* - n^*| = o(\sqrt{R}) .
\]

Remark 4.8 In continuation to Remark 4.7, if the constraint parameter $\alpha \geq \bar{G}(T)$, then the optimal staffing $n^* = 0$.

5 Global constraint satisfaction

Here we study the global constraint satisfaction that was discussed in Section 2.5 in the Introduction. We introduce the same assumptions as in Sections 4.1 and 4.2: the service rate $\mu$ and patience distribution $G$ are fixed, with $G$ having a positive density $g_0$ at the origin. However, now we consider a set of $K$ time intervals that constitute a day of work. Arrivals to each interval are governed by a Poisson process with rates $r_i \lambda$, $1 \leq i \leq K$, $\sum_{i=1}^{K} r_i = 1$. In other words, $r_i$ are the fractions of daily arrival rate. The staffing costs at the intervals are equal to $c_i$, $1 \leq i \leq K$.

The vector of staffing levels is determined according to the overall arrival rate and is denoted by
\[
\bar{n}(\lambda) \triangleq (n_1(\lambda), \ldots, n_K(\lambda)) ,
\]
where $n_i(\lambda)$ is the staffing level during interval $i$. In other words, we shall let $\lambda \to \infty$ while keeping the fractions $(r_1, \ldots, r_K)$ fixed.

Modify definition (4.4) of the optimal staffing level to
\[
\bar{n}^*(\lambda) = \arg \min \sum_{i=1}^{k} c_i n_i(\lambda) ,
\]
subject to $U(\bar{n}, \lambda) \leq M$. 

Definition (4.5) of asymptotic feasibility is unchanged and, finally, the notion of asymptotic optimality is modified to

\[ \left| \sum_{i=1}^{K} c_i n_i(\lambda) - \sum_{i=1}^{K} c_i n^*_i(\lambda) \right| = o(f(\lambda)) = o(f(R)). \]

We analyze two special cases that give rise to the QED operational regime: constraint on the delay probability in Section 5.1 and a scaled constraint on the probability to abandon in Section 5.2. See Section 8 for suggestion on possible future research on alternative global constraints.

5.1 Global constraint on the delay probability

Define a global constraint on the delay probability by

\[ \Pr\{W > 0\} \leq \alpha. \] (5.1)

(This is equivalent to the disutility function \( U(n, \lambda) = c_b P_{n, \lambda}\{W > 0\} \).)

In order to present our asymptotically optimal solution, we must solve two optimization problems. First, introduce an optimization problem on all subsets of time intervals \( H \subseteq \{1, 2, \ldots, K\} \):

\[
\begin{cases}
\max_{H} \sum_{i \in H} c_i r_i, \\
\text{s.t.} \sum_{i \in H} r_i \leq \alpha,
\end{cases}
\] (5.2)

which is an example of the classical "knapsack problem".

Assume that (5.2) has a unique solution \( H^* \) and define

\[ \tilde{\alpha} \triangleq \alpha - \sum_{i \in H^*} r_i. \] (5.3)

Assume, in addition, that \( \tilde{\alpha} \) is positive.

Let \( \bar{H}^* \) be the complement of \( H^* \). Introduce a second non-linear optimization problem with respect to \( \bar{\beta} = \{\beta_i, \ i \in \bar{H}^*\} \):

\[
\begin{cases}
\min_{\bar{\beta}} \sum_{i \in \bar{H}^*} c_i \beta_i \sqrt{r_i}, \\
\text{s.t.} \sum_{i \in \bar{H}^*} r_i P_w(\beta_i) = \tilde{\alpha},
\end{cases}
\] (5.4)

where the function \( P_w(\cdot) \) was defined in (4.7). As discussed at the beginning of Section 7.1, \( P_w(\cdot) \) is a continuous function that decreases from one to zero. In addition, (5.3) implies that

\[ \sum_{i \in \bar{H}^*} r_i = 1 - \sum_{i \in H^*} r_i > \tilde{\alpha}. \]

Therefore, the problem (5.4) has at least one solution. Denote the value of the minimum in (5.4) by \( \delta^* \).
Theorem 5.1 (Global delay probability) Under the definitions and conditions presented above:

a. The optimal staffing level with respect to (5.1) satisfies

\[ n_i^* = o(\sqrt{R}), \quad i \in H^*, \]  \hspace{1cm} (5.5)

\[ n_i^* = R_i + O(\sqrt{R}), \quad i \in \bar{H}^*, \]  \hspace{1cm} (5.6)

\[ \sum_{i \in H^*} c_i n_i^* = \sum_{i \in \bar{H}^*} c_i R_i + \delta^* \sqrt{R} + o(\sqrt{R}), \]  \hspace{1cm} (5.7)

where \( R_i = (r_i \lambda) / \mu \) and \( R = \sum R_i = \lambda / \mu \).

In addition, if \( \{ \beta_i^*, i \in \bar{H}^* \} \), is the unique solution of the optimization problem (5.4), then

\[ n_i^* = R_i + \beta_i^* \sqrt{R_i} + o(\sqrt{R}), \quad i \in \bar{H}^*. \]  \hspace{1cm} (5.8)

b. Consider the staffing level \( \tilde{n}^* = (\tilde{n}_1^*, \ldots, \tilde{n}_K^*) \), given by

\[ \tilde{n}_i^* = 0, \quad i \in H^*, \]  \hspace{1cm} (5.9)

\[ \tilde{n}_i^* = \left\lceil R_i + \beta_i^* \cdot \sqrt{R_i} \right\rceil, \quad i \in \bar{H}^*; \]  \hspace{1cm} (5.10)

here \( \beta_i^*, i \in \bar{H}^* \), is some solution of (5.4). Then the staffing level (5.9)-(5.10) is asymptotically feasible and asymptotically optimal in the sense that:

\[ \sum_{i=1}^{K} c_i \tilde{n}_i^* - \sum_{i=1}^{K} c_i n_i^* = o(\sqrt{R}). \]  \hspace{1cm} \blacksquare

Remark 5.1 In words, Theorem 5.1 gives rise to the following structure of the optimal solution. First, approach the constraint (5.1) as close as possible by “closing the gate” at the intervals \( i \in H^* \), according to (5.9). Then satisfy the constraint by assigning QED staffing at the other intervals \( i \in \bar{H}^* \), according to (5.10). This solution, although valid theoretically, is not very appropriate as a practical recommendation. For example, customers that were turned away will possibly try to get service again at subsequent time intervals.

It is also possible to show that for the global constraint on the tail probability \( P\{W > T / \sqrt{X}\} \leq \alpha \) (scaled deadline), the solution will have the structure of Theorem 5.1 (QED on some intervals, no service at the others). Similarly, for the unscaled constraint \( P\{W > T\} \leq \alpha \), some intervals will have ED+QED staffing while others will be closed.

In the next section we present a setting that seems to us more practical.

5.2 Global constraint on the probability to abandon

We have seen in Section 5.1 that the setting with global constraint on the delay probability can imply a solution that is not practical: “closing the gate” at some intervals. Now we present a setting that implies QED staffing at all intervals.
Retain the assumption articulated in the beginning of Section 5.

**Theorem 5.2 (Global probability to abandon)** Consider the following constraint on the overall probability to abandon:

\[ P\{\text{Ab}\} \leq \frac{\alpha}{\sqrt{\lambda}}. \]  

(5.11)

Define the following optimization problem with respect to \( \bar{\beta} = \{\beta_i, \ 1 \leq i \leq K\} \):

\[
\begin{align*}
\min_{\bar{\beta}} & \sum_{i=1}^{K} c_i \beta_i \sqrt{r_i}, \\
\text{s.t.} & \sum_{i=1}^{K} r_i P_w(\beta_i) P_a(\beta_i) = \alpha.
\end{align*}
\]

(5.12)

Here \( P_a(\cdot) \) was defined in (4.8). Again, as in (5.4), at least one solution of (5.12) exists. Denote the minimal value of (5.12) by \( \delta^* \).

a. The optimal staffing level with respect to condition (5.11) satisfies

\[ n_i^* = R_i + O(\sqrt{R}), \quad 1 \leq i \leq K, \]

\[ \sum_{i=1}^{K} c_i n_i^* = \sum_{i=1}^{K} c_i \cdot R_i + \delta^* \sqrt{R} + o(\sqrt{R}). \]

If \( \{\beta_i^*\} \) is the unique solution of (5.12), then

\[ n_i^* = R_i + \beta_i^* \cdot \sqrt{R_i} + o(\sqrt{R}), \quad 1 \leq i \leq K. \]

b. Consider the staffing level

\[ \tilde{n}_i^* = \left[ R_i + \beta_i^* \cdot \sqrt{R_i} \right], \quad 1 \leq i \leq K, \]

(5.13)

where \( \{\beta_i^*\} \) is a solution of (5.12). Then the staffing level (5.13) is asymptotically feasible and asymptotically optimal in the sense that

\[ \left| \sum_{i=1}^{K} c_i \tilde{n}_i^* - \sum_{i=1}^{K} c_i n_i^* \right| = o(\sqrt{R}). \]

The proof of Theorem 5.2 is similar to that of Theorem 5.1.

**Remark 5.2** The intuition for the constraint in (5.12) is the following. In the QED regime, the probability to abandon at the \( i \)-th interval can be approximated, as \( \lambda \to \infty \), by

\[ P_i\{\text{Ab}\} \sim \frac{P_w(\beta_i)P_a(\beta_i)}{\sqrt{r_i \lambda}}. \]
Since $P\{\text{Ab}\} = \sum_{i=1}^{K} r_i P_i \{\text{Ab}\}$,
\[
\sum_{i=1}^{K} \sqrt{r_i} P(w(\beta_i) P_a(\beta_i)) \approx P\{\text{Ab}\} \sqrt{\lambda} \approx \alpha.
\]

**Remark 5.3** If the unscaled global constraint $P\{\text{Ab}\} \leq \alpha$ is considered, asymptotically optimal staffing is ED, possibly somewhat modified:

\[
n_i = (1 - \gamma_i) R_i + o(R_i), \quad 0 \leq \gamma_i \leq 1, \quad 1 \leq i \leq K.
\]

In order to calculate $\{\gamma_i\}$, the following linear programming problem should be solved:

\[
\begin{align*}
\max_{\{\gamma_i\}} & \sum_{i=1}^{K} c_i r_i \gamma_i, \\
\text{s.t.} & \sum_{i=1}^{K} \gamma_i r_i = \alpha, \quad 0 \leq \gamma_i \leq 1.
\end{align*}
\]

If all staffing costs $c_i$ are equal, then any $\{\gamma_i\}$ with $\sum \gamma_i r_i = \alpha$ is asymptotically optimal. Otherwise, assume, without loss of generality, that $c_1 \geq c_2 \geq \ldots \geq c_K$. Then an optimal solution is obtained recursively: $\gamma_i = \min(1, \alpha_i / r_i)$, $1 \leq i \leq K$, where $\alpha_1 = \alpha$ and $\alpha_i = \alpha_{i-1} - \gamma_i r_i - \gamma_i R_i$, $2 \leq i \leq K$.

**Conjecture.** Our numerical experiments with problem (5.12) revealed the following interesting phenomenon. Assume the case of equal costs $c_i$, $1 \leq i \leq K$, at all intervals. Then the problem (5.12) seems to have a unique solution with equal $\beta_1^* = \beta_2^* = \ldots = \beta_K^*$. This implies, in particular, higher service levels for intervals with larger arrival rates.

6 Numerical experiments

In this section, we validate the quality of our asymptotic solutions developed in Sections 4 and 5. We cover the single-interval QED, ED and ED+QED regimes in Sections 6.1-6.3, respectively. Our general approach is to compare the asymptotic staffing levels, derived in Theorems 4.1, 4.2 and 4.4, with the exact optima calculated via formulae in [40]. We also compare our approximations, checking how QED works in the parameter range of the other two regimes; to this end, in Section 6.4 we revisit Examples 2.1-2.3 from the Introduction. Finally, in Section 6.5 we provide some experiments on the global constraints.

6.1 QED approximations

In the QED regime, we perform an extensive numerical experiment within the following framework. Let the service rate $\mu=1$. (In other words, the average service time is our time unit.) We
consider arrival rates that vary from 10 to 100 by step 10, from 100 to 400 by step 20 and from 400 to 1,000 by step 40, for a total of 40 values of $\lambda$. In addition, six patience distributions are chosen:

- Two exponential distributions with means 2 and 0.5;
- Two uniform distributions on $[0,4]$ and $[0,1]$;
- Two hyperexponential distributions, both being a 50-50% mixture of two exponentials. The exponential means are 1 and 3 in the first case (mean patience equals to 2), and $1/4$ and $3/4$ in the second case (mean patience $1/2$).

Note that we consider three types of distributions and, for each type, choose two representatives: the first one with average patience longer than the average service time, and the second one with shorter patience.

For each combination of $\lambda$, patience distribution and a specific form of disutility function, we compare between the exact optimal and asymptotically optimal staffing levels $n^*$ and $n^*_{QED}$.

**Constraint on the delay probability.**

In Theorem 4.1 assume the cost coefficients $C_{ab} = 0$, $C_w = 0$, $C_b = 1$ and the deadline $t = 0$. In this case, the constraint (4.4) is equivalent to

$$P\{W > 0\} \leq M.$$ 

We take values of $M$ equal to 0.1–0.9 by step 0.1. Combining these values with the above-mentioned arrival rates and patience distributions, we get $40 \times 6 \times 9 = 2160$ special cases. The optimal staffing levels $n^*$, as in (4.4), vary from 5 to 1,045. For all these values, the asymptotically optimal $n^*_{QED}$ deviates from $n^*$ by no more than 2. Specifically, $n^*_{QED}$ matches $n^*$ exactly in 57% of the cases; it deviates from $n^*$ by 1 in 40% of the cases; and by 2 in the remaining 3%.

The exact optimal staffing levels are always larger or equal to the approximate ones and the differences become smaller as $M$ decreases. The fit for longer patience is, overall, better.

**Constraint on the probability to abandon.**

In Theorem 4.1 assume $C_{ab} = 1$, $C_w = 0$, $C_b = 0$. Then the constraint is equivalent to

$$P\{Ab\} \leq M/\sqrt{\lambda}.$$ 

We consider values of $M$ that vary from $0.03\cdot \sqrt{10}$ to $0.3\cdot \sqrt{10}$ by step $0.03\cdot \sqrt{10}$. (For example, $M = 0.03\cdot \sqrt{10}$ corresponds to 3% abandonment for $\lambda = 10$ and 0.3% for $\lambda = 1000$.)

Out of 2,400 experiments, we observe a perfect fit in 2,242 (93%). In all other cases, except for one, the difference is 1 (either negative or positive). For large negative QoS parameters (that
correspond to large values of $M$), the staffing levels for the different distributions are very close to each other. (See Section 6.2 for an explanation of this phenomenon.)

**Constraint on average wait.**

In Theorem 4.1 assume $C_{ab} = 0$, $C_w = 1$, $C_b = 0$. Then the constraint condition is given by

$$E_{n,\lambda}[W] \leq M/\sqrt{\lambda}.$$  

For the distributions with mean patience equal to 2, values of $M$ are chosen equal to $(1/12) \cdot \sqrt{10}, \ldots, \sqrt{10}$ by step $(1/12) \cdot \sqrt{10}$. (For example, $M = (1/3) \cdot \sqrt{10}$ implies maximal average wait of $1/3$ average service times for $\lambda = 10$ and $1/30$ average service times for $\lambda = 1000$.) For the three distributions with smaller patience $(1/2)$, $M$ varies from $(1/48) \cdot \sqrt{10}$ to $(1/4) \cdot \sqrt{10}$ by step $(1/48) \cdot \sqrt{10}$.

The fit is excellent again. Out of 2,880 experiments, a perfect fit in observed in 2,111 (73%) cases. Otherwise, the difference is equal to 1, except for a single case when it equals 2. We observe considerable staffing differences between distributions if relatively large wait is acceptable (large $M$).

**Conclusions.** The QED approximations are superb for an extensive set of parameters and patience distributions. We now proceed to check how these approximations work for very extreme values that arise if staffing is performed according to the other two operational regimes.

### 6.2 ED approximations

Here we check the quality of the approximations in Theorem 4.2, comparing $n^\ast$ and $n_{ED}^\ast$. In addition, for each numerical experiment we calculated an asymptotically optimal QED staffing level. Specifically, we assigned to the cost coefficients in Theorem 4.1 the values $C_{ab} = 1/\sqrt{\lambda}$, $C_w = 1/\sqrt{\lambda}$ and $C_b = 0$.

**Constraint on the probability to abandon.** The constraints on $P\{Ab\}$ are varied from 0.04 to 0.4 by step 0.04. We consider the 3 patience distributions from Section 6.1 with mean 1/2. (Results for the distributions with mean 2 are similar, but the differences between $n^\ast$, $n_{ED}^\ast$ and $n_{QED}^\ast$ are somewhat smaller.)

We consider 3 values of the offered load $R = \lambda/\mu$: 10 (small), 100 (moderate) and 1,000 (large). The results are displayed in Figures 4 and 5. Stars are for the exact optima, dashed lines are for the QED approximations and the solid line is for the ED approximation. (The ED approximation for $P\{Ab\}$ does not depend on the patience distribution.)
Already for $R = 10$ or $R = 100$, we observe an almost perfect fit between the exact optimal staffing and its QED approximation. For small values of the constraint (0.05-0.2), ED estimates imply significant understaffing.

In the case $R = 1000$, all the three types of lines merge into a single line. Both approximations are excellent.

**Constraint on average wait.** Constraints on $E[W]$ are varied from 1/15 to 2/3 by step 1/15. The three patience distribution from Section 6.1, with mean 2, are considered; offered load is
again chosen to be $R=10$, 100 and 1000. (The three distributions with mean 1/2 provide very similar results.)

Figure 6: **Constraint on average wait**

If $R = 10$ (Figure 6) the fit of the QED approximations is excellent; ED is a bit worse for small values of constraint. However, for $R = 100$ we observe a bias in the QED estimates for the case of non-exponential patience distributions (2-3 at the largest values of constraint).

Figure 7: **Constraint on average wait**

If $R = 1000$ (Figure 7) the fit of the QED approximations is excellent; ED is a bit worse for small values of constraint. However, for $R = 100$ we observe a bias in the QED estimates for the case of non-exponential patience distributions (2-3 at the largest values of constraint).
Finally, for $R = 1000$ the QED bias can be very significant (20-30 for the largest value of constraint). The ED approximation, on the other hand, is almost perfect.

**Constraint on the probability to abandon: conclusions and discussion.** The results for the ED approximations are to be expected: the fit is not that good for small (strict) values of the constraints. It improves for looser (larger) constraints and for larger arrival rates. The results for the QED approximations are astonishing. In addition to the experiments presented in Figures 4 and 5, we performed an extensive numerical study with the 6 patience distributions and 40 values of the offered load from Section 6.1, and 10 values of constraints. Exact fit has been observed in 2,346/2,400 (98%) cases, otherwise the difference is 1. Hence, QED estimates turn out to be excellent for all the special cases considered in Sections 6.1 and 6.2.

Why do we have a perfect fit for $P\{Ab\}$? Recall the QED approximation of the probability to abandon:

$$P\{Ab\} \approx \frac{1}{\sqrt{\lambda}} \sqrt{g_0} [h_\phi(\hat{\beta}) - \hat{\beta}].$$

For large negative $\beta$, the normal hazard rate $h_\phi(\cdot)$ is negligible. Using definition (4.10) of $\hat{\beta}$, we can easily deduce that the QED approximation is then close to $(R - n)/R$, namely the ED approximation.

**Constraint on average wait: conclusions and discussion.** The conclusions for the ED approximations of $E[W]$ are similar to the conclusions for $P\{Ab\}$: they work very well for large arrival rates (hundreds and more) and relatively loose values of constraints. However, the situation with the QED approximations is different.

The QED approximations provide an excellent fit for the exponential distribution. This can be explained by the high quality of $P\{Ab\}$ approximations and the relation $P\{Ab\} = \theta \cdot E[W]$, which prevails for both exact values and QED approximations. (Note that the exponential parameter $\theta = g_0$.)

However, if $\lambda$ and the constraint values are large, we observe a strong bias of the QED estimates for non-exponential distributions. Note that the relation $P\{Ab\} = g_0 \cdot E[W]$ prevails for the QED approximations but does not for ED, and the latter is very close to the exact values in this case. Therefore, the excellent QED fit for $P\{Ab\}$ is not always retained for $E[W]$.

**6.3 ED+QED approximation for the tail probability**

Now we check the quality of the ED+QED approximation from Theorem 4.4. Consider the three distributions from Section 6.2 with patience mean 2 and take the deadline equal to $1/3$. Constraints on the probability to exceed this deadline are varied from 0.05 to 0.5 by step 0.05. Finally, three values of the arrival rate (offered load) are considered: 10, 100 and 1,000.
We compare, in Figures 8 and 9, these approximation with exact optimal staffing and the QED approximation derived via Theorem 4.1.

Figure 8: **Constraint on the tail probability**

Offered load $R = 10$

Offered load $R = 100$

From (4.29), the approximate ED+QED staffing level is equal to

$$n^*_{ED+QED} = \left\lceil (1 - \gamma^*) \cdot R + \delta^* \sqrt{R} \right\rceil,$$

where $\gamma^* = G(T)$ and $\delta^*$ is defined by (4.30).

Figure 9: **Constraint on the tail probability**

Offered load $R = 1000$
Note that for every distribution in consideration, the ED coefficient $\gamma^*$ is constant for all experiments. Specifically, it is 0.154 for the exponential distribution, 0.083 for uniform and 0.194 for hyperexponential. (That explains why the uniform distribution requires a larger staffing level.) The QED coefficient $\delta^*$ provides fine tuning for different values of constraints.

In the case $R = 10$ (Figure 8) we observe a nearly perfect fit for QED staffing, which is preferable over ED+QED. If $R = 100$ (moderate number of agents) the fit for ED+QED is already slightly better. Finally, if $R = 1000$ (Figure 9) the fit for ED+QED is fine, while the QED estimate is strongly biased for exponential and hyperexponential distributions (6-10 and 15-23 servers, respectively). The bias takes place for very large negative QoS parameters (-5 \ldots -7). For the uniform distribution the QoS parameters are smaller (-2 \ldots -3), so the fit is considerably better.

The main conclusion, therefore, is that the ED+QED approximation is preferable over QED for moderate-to-large call centers and moderate-to-loose constraints on the service level.

6.4 Comparison between operational regimes

Recall that in Section 2.4 of the Introduction we discussed existence of a single operational regime that is preferable over the others. Here we re-visit Examples 2.1-2.3 from the Introduction in order to show that there is no such regime. Note that the same conclusion can be deduced if one summarizes numerical examples from Section 6.2 and 6.3.

Example 6.1 (Constraint satisfaction in a small call center) Consider the setting of Example 2.1 with hyperexponential patience, where the QED approximations provide us with a perfect fit. Apply the ED staffing (2.4) and the corresponding approximations. It is straightforward to check, via $P_{ED}\{Ab\} = \gamma$, that the ED recommendation for the constraint $P\{Ab\} \leq 2\%$ is $n^*_{ED} = 59$. This is very far from the exact optimum $n^* = 67$ and would lead to $P\{Ab\} = 6.7\%$ – more than three-fold worse than the service goal. Therefore, the ED recommendations should not be used for small call centers.

Now we check if the ED+QED regime is robust for a small call center. Applying it for Exp(mean=3), $U(0,6)$ and our hyperexponential distribution, we get respectively that $n^*_{ED+QED} = 63$, 64 and 61, while $n^* = n^*_{QED} = 64$, 66 and 61. We observe a perfect QED fit. Hence the ED+QED recommendations are not that bad but the QED ones, nevertheless, are preferable for small call centers.

Example 6.2 (Constraint satisfaction in a large efficiency-driven call center) Consider the large call center from Examples 2.2-2.3, where the ED and ED+QED approximations were found appropriate. We check if the QED approximations are robust in this case, considering the three patience distributions from the end of Example 6.1. First, consider the constraint $P\{Ab\}$
≤ 10% from Example 2.2. QED recommends 1,081 agents for all distributions, which coincides with the exact optima.

In contrast, QED staffing for the constraint “E[W] ≤ 20 seconds” is 1,067, 1,134 and 961 vs. the exact optima of 1,067, 1,132 and 972. We observe that the fit for our hyperexponential patience is relatively poor.

Considering the constraint on \( P\{ W > T \} \) from Example 2.3, we also get a poor fit of QED approximations, especially for the hyperexponential distribution: \( n_{ED+QED}^* = 1,000 \) vs. \( n^* = 1,021 \).

Hence, using QED approximations in large ED call centers can mislead if moderate-to-loose constraints on the waiting time are considered.

### 6.5 Experiments on global constraint

**Global constraint on the delay probability.** We present several examples where the QED regime is asymptotically optimal at all intervals. The Lagrange multipliers are used in order to solve the optimization problem (5.4). Fix the service rate \( \mu = 1 \).

**Example 6.3** Consider two time-intervals, assume overall arrival rate \( \lambda = 100, r_1 = 0.7, r_2 = 0.3 \), the constraint \( \alpha = 0.2 \) (80% of customers get service immediately) and six patience distributions considered in Section 6.1. Assume that the staffing costs at both intervals are equal: \( \bar{c} = (1, 1) \). Below we compare exact optimal staffing \( \bar{n}^* \) and the approximation \( \bar{n}_{QED}^* \) derived via Theorem 5.1. We observe an excellent fit between \( \bar{n}^* \) and \( \bar{n}_{QED}^* \).

<table>
<thead>
<tr>
<th>Patience distribution</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{n}^* )</td>
<td>(80,34)</td>
<td>(80,35)</td>
<td>(79,34)</td>
<td>(79,32)</td>
<td>(79,34)</td>
<td>(78,32)</td>
</tr>
<tr>
<td>( \bar{n}_{QED}^* )</td>
<td>(80,34)</td>
<td>(80,34)</td>
<td>(79,34)</td>
<td>(79,32)</td>
<td>(79,33)</td>
<td>(78,31)</td>
</tr>
</tbody>
</table>

**Example 6.4** Now assume staffing costs \( \bar{c} = (1, 1.8) \) and retain the other settings from Example 6.3. Below are the results of this numerical experiment.

<table>
<thead>
<tr>
<th>Patience distribution</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{n}^* )</td>
<td>(82,32)</td>
<td>(82,33)</td>
<td>(80,33)</td>
<td>(81,30)</td>
<td>(82,31)</td>
<td>(80,30)</td>
</tr>
<tr>
<td>( \bar{n}_{QED}^* )</td>
<td>(82,32)</td>
<td>(82,33)</td>
<td>(82,31)</td>
<td>(83,29)</td>
<td>(82,31)</td>
<td>(83,28)</td>
</tr>
</tbody>
</table>

The two optima are close again. Some differences can be due to the fact that \( \bar{n}_{QED}^* \) is derived by rounding a continuous solution of (5.4) that can be somewhat different from the optimal solution in integers. Comparing with Example 6.3, there are more agents at the first interval and less at the second one due to the change in staffing costs.
Is the solution of (5.4) always unique? In fact, the uniqueness depends on the convexity of the curve \( r_1 P_w(\beta_1) + r_2 P_w(\beta_2) = \tilde{\alpha} \). From [40] we deduce that if a counterexample exists, it should be searched for at large values of \( \tilde{\alpha} \) and impatient customers (large \( g_0 \)).

**Example 6.5** Let \( g_0 = 10, k = 2, r_1 = r_2 = 0.5 \) and \( \tilde{\alpha} = 0.45 \). Then \( \beta_1^* = -3.6671, \beta_2^* = 1.2389 \) and \( \beta_1^* = 1.2389, \beta_2^* = -3.6671 \) are two global minimums of (5.4). The point \( \beta_1^* = \beta_2^* = -1.0381 \) is a local maximum.

Overall, our numerical experiments show that, unless \( \tilde{\alpha} \) is large and customers are very impatient (as in Example 6.5), the solution of (5.4) is unique and larger QoS parameters correspond to larger arrival rates \( r_i \). The following example illustrates this phenomenon.

**Example 6.6** Let \( g_0 = 0.5, k = 5, \tilde{\tau} = (0.25, 0.15, 0.25, 0.15, 0.2) \) and \( \tilde{\alpha} = 0.1 \). The the optimal solution to (5.4) is \( \tilde{\beta}^* = (1.4132, 1.2306, 1.4132, 1.2306, 1.3363) \).

**Global constraint on the probability to abandon.** We have already presented a large numerical experiment in Section 2.5 of Introduction. The following two small experiments study the fit between approximate and exact optimal staffing levels. Consider the constraint “\( P\{\text{Ab}\} \leq 3\% \)” and retain other parameters from Examples 6.3 and 6.4 on the delay probability.

**Example 6.7** Assume equal staffing costs: \( \bar{c} = (1, 1) \).

<table>
<thead>
<tr>
<th>Patience distribution</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{n}^* )</td>
<td>(73,31)</td>
<td>(71,31)</td>
<td>(73,32)</td>
<td>(75,34)</td>
<td>(74,33)</td>
<td>(76,34)</td>
</tr>
<tr>
<td>( \bar{n}_{QED}^* )</td>
<td>(73,32)</td>
<td>(71,31)</td>
<td>(73,32)</td>
<td>(75,34)</td>
<td>(74,33)</td>
<td>(76,34)</td>
</tr>
</tbody>
</table>

**Example 6.8** Assume staffing costs \( \bar{c} = (1, 1.8) \).

<table>
<thead>
<tr>
<th>Patience distribution</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{n}^* )</td>
<td>(73,31)</td>
<td>(72,30)</td>
<td>(74,31)</td>
<td>(77,32)</td>
<td>(76,31)</td>
<td>(78,32)</td>
</tr>
<tr>
<td>( \bar{n}_{QED}^* )</td>
<td>(75,31)</td>
<td>(73,30)</td>
<td>(75,31)</td>
<td>(78,32)</td>
<td>(76,31)</td>
<td>(78,32)</td>
</tr>
</tbody>
</table>

In both experiments we observe a very good fit between \( \bar{n}^* \) and \( \bar{n}_{QED}^* \).

7 Proofs

7.1 Proofs of Theorems 4.1-4.4

**Monotonicity statements.** We start with verification of monotonicity in the number of servers for our performance measures and their QED approximations, which is needed for the proofs below. First, we need to prove that the functions (4.7)-(4.9) are strictly decreasing in \( \beta \), where \( P_w(\beta) \) and \( W(\beta, t), t > 0 \), decrease from one to zero and \( P_a(\beta) \) decreases from infinity to zero.
Second, given that \( \lambda, \mu, G \) and \( t \geq 0 \) are fixed, we show that the performance measures \( P\{Ab\} \), \( E[W] \) and \( P\{W > t\} \) are decreasing in \( n \).

Recall (see [40], for example) that the standard normal hazard rate \( h_\phi(\cdot) \) is a strictly increasing convex function and \( h_\phi(x) - x \downarrow 0 \), as \( x \to \infty \). This property implies monotonicity of (4.7)-(4.9). (Use the well-known relation \( \Phi(t) = \exp\left\{ -\int_{-\infty}^t h_\phi(x)dx \right\} \) in order to verify the statement for (4.9).)

Now we check monotonicity of performance measures. The result for \( P\{Ab\} \) was derived in Bhattacharya and Ephmerides [6].

For \( P\{W > t\} \) we use M/M/n+G formulae from Zeltyn and Mandelbaum [40]. The probability not to get service immediately is given by:

\[
P\{V > 0\} = \frac{\lambda J}{\mathcal{E} + \lambda J},
\]

where

\[
J \triangleq \int_0^\infty \exp\{\lambda H(x) - n\mu x\} dx, \quad (7.15)
\]

\[
\mathcal{E} \triangleq \int_0^\infty e^{-t}\left(1 + \frac{t\mu}{\lambda}\right)^{n-1} dt, \quad (7.16)
\]

\[
H(x) \triangleq \int_0^x \bar{G}(u)du. \quad (7.17)
\]

Since \( J \) decreases in \( n \) and \( \mathcal{E} \) increases in \( n \), \( P\{V > 0\} \) decreases in \( n \). Now note that

\[
P\{W > t\} = \bar{G}(t) \cdot P\{V > 0\} \cdot P\{V > t|V > 0\}.
\]

Hence, it is enough to prove monotonicity in \( n \) for

\[
P\{V > t|V > 0\} = \frac{\int_t^\infty \exp\{\lambda H(x) - n\mu x\} dx}{\int_0^\infty \exp\{\lambda H(x) - n\mu x\} dx},
\]

or, in other words, that for \( n_2 > n_1, t > 0 \)

\[
\frac{\int_t^\infty \exp\{\lambda H(x) - n_2\mu x\} dx}{\int_t^\infty \exp\{\lambda H(x) - n_1\mu x\} dx} \leq \frac{\int_0^\infty \exp\{\lambda H(x) - n_2\mu x\} dx}{\int_0^\infty \exp\{\lambda H(x) - n_1\mu x\} dx}. \quad (7.18)
\]

Formula (7.18) is implied by the following observation:

**Lemma.** Let functions \( f_1 \) and \( f_2 \) be defined on \([0, \infty)\). Assume that they are positive and \( f_2/f_1 \) decreases. Then

\[
\frac{\int_t^\infty f_2(y)dy}{\int_t^\infty f_1(y)dy} \leq \frac{\int_0^\infty f_2(y)dy}{\int_0^\infty f_1(y)dy},
\]

assuming that the right-hand side is well-defined.
The proof of the Lemma follows from the inequalities
\[
\frac{\int_{t}^{\infty} f_2(y)dy}{\int_{t}^{\infty} f_1(y)dy} \leq \frac{f_2(t)}{f_1(t)} \leq \frac{\int_{0}^{t} f_2(y)dy}{\int_{0}^{t} f_1(y)dy},
\]
which are easily verified. For example,
\[
\int_{t}^{\infty} f_2(y)dy = \int_{t}^{\infty} f_1(y) \cdot \frac{f_2(y)}{f_1(y)}dy \leq \frac{f_2(t)}{f_1(t)} \cdot \int_{t}^{\infty} f_1(y)dy.
\]
Finally, monotonicity in \( n \) of \( P\{W > t\} \) implies monotonicity of the average wait \( E[W] \). (The conventional stochastic order implies the same order between expectations.)

**Proof of Theorem 4.1 (QED).** First, note that the existence of a unique solution \( \beta^* \) of equation (4.17) follows from strict monotonicity of (4.7)-(4.9).

Then for some \( \epsilon > 0 \) define
\[
n_1 = \left\lceil R + (\beta^* - \epsilon)\sqrt{R} \right\rceil,
n_2 = \left\lfloor R + (\beta^* + \epsilon)\sqrt{R} \right\rfloor.
\]
According to relations (4.11)-(4.14), monotonicity of the functions (4.7)-(4.9) and the definition of \( \beta^* \), we get that \( U(n_1, \lambda) \to M + \delta_1 \) and \( U(n_2, \lambda) \to M - \delta_2 \), for some \( \delta_1, \delta_2 > 0 \), as \( \lambda \to \infty \). Therefore, for \( \lambda \) large enough, \( U(n_1, \lambda) > M + \delta_1/2 \) and \( U(n_2, \lambda) < M - \delta_2/2 \).

Since \( P\{Ab\}, E[W] \) and \( P\{W > t\} \) decrease in \( n \), the cost function (4.15) decreases too. Then definition (4.4) implies that, for \( \lambda \) large enough, \( n_1 < n^{*}_{\lambda} \leq n_2 \), which, since \( \epsilon > 0 \) is arbitrary, proves (4.16) and Part a.

If we define asymptotically optimal staffing by
\[
n^{*}_{\text{QED}} = \left\lceil R + \beta^*\sqrt{R} \right\rceil,
\]
Part b follows from (4.11)-(4.14) and (4.16).

**Proof of Theorem 4.2 (ED).** Theorem 6.1 from Zeltyn and Mandelbaum [40] implies that in the ED operational regime:
\[
n = (1 - \gamma) \cdot R + o(R), \quad \gamma > 0;
\]
the probability to abandon and average wait converge to \( \gamma \) and \( H(G^{-1}(\gamma)) \), respectively. The proof of Theorem 4.2 follows from this statement in the same way that Theorem 4.1 follows from (4.11)-(4.14).

**Remark 7.1 (On the \( o(R) \) term in (7.19))** Theorem 6.1 in [40] defined the ED regime with an \( o(\sqrt{R}) \) term. However, the convergence proofs for \( P\{Ab\} \) and \( E[W] \) remain unchanged if definition (7.19) is used. The \( o(\sqrt{R}) \) term is needed if one derives an exponential rate of convergence for the delay probability.
Proof of Theorem 4.3 (ED+QED performance measures).

1 ⇒ 2. We shall prove this statement using a continuous indexing of M/M/n+G queues by the arrival rate \( \lambda \), which is more general than indexing by \( n \). (We do this for a smoother transition into Theorem 4.4, where indexing by \( \lambda \) is required.) In this case, the ED+QED staffing level is given by

\[
    n = \bar{G}(T) \cdot \frac{\lambda}{\mu} + \delta \sqrt{\frac{\lambda}{\mu} + f(\lambda)}, \quad -\infty < \delta < \infty, \quad f(\lambda) = o(\sqrt{\lambda}).
\]  

(7.20)

From Zeltyn and Mandelbaum [40],

\[
P\{W > T\} = \frac{\lambda \bar{G}(T)J(T)}{\mathcal{E} + \lambda J}.
\]

(7.21)

where \( J \) and \( \mathcal{E} \) were defined in (7.15)-(7.16) and

\[
    J(T) = \int_{-T}^{T} \exp \{\lambda H(x) - n\mu x\} \, dx.
\]

In order to calculate the asymptotics for \( P\{W > T\} \), we start with the building blocks \( J \), \( J(T) \) and \( \mathcal{E} \), using the Laplace method for asymptotic calculation of integrals (see de Bruijn [12] or Sections 10 and 11 in Zeltyn and Mandelbaum [40]). We shall produce a detailed derivation of the approximation for \( J \); the derivations for \( J(T) \) and \( \mathcal{E} \) are similar.

Lemma 7.1 (Approximation for \( J \)) Under the staffing level \( n \) given by (7.20), and \( \lambda \to \infty \),

\[
    J \sim \exp \{\lambda H(T) - n\mu T\} \cdot \exp \left\{ \frac{\delta^2 \mu}{2g(T)} \right\} \cdot \sqrt{\frac{2\pi}{\lambda g(T)}}.
\]

(7.22)

Proof of Lemma 7.1. Performing the change-of-variables \( y = x - T \), we get

\[
    J = \exp \{\lambda H(T) - n\mu T\} \cdot \int_{-\infty}^{\infty} \exp \left\{ \lambda \int_{-T}^{T+y} \bar{G}(u) du - n\mu y \right\} \, dy,
\]

(7.23)

where, if \( y < 0 \), \( \int_{-T}^{T+y} \Delta = -\int_{T+y}^{T} \). The main idea of the proof, via the Laplace method, is to replace the internal integral in (7.23) by the first two terms of the Taylor expansion:

\[
    \int_{-T}^{T+y} \bar{G}(u) du \approx \bar{G}(T)y - \frac{1}{2}g(T)y^2 + O(y^3), \quad (y \to 0)
\]

(7.24)

and validate this substitution by showing that the value of the external integral in (7.23) depends mainly on the behavior of the integrand near the origin.

Specifically, define

\[
    J_{A,\epsilon} \Delta \equiv \exp \{\lambda H(T) - n\mu T\} \cdot \int_{-\infty}^{\infty} \exp \left\{ \lambda \bar{G}(y) - \frac{1}{2}\lambda(g(T) + \epsilon)y^2 - n\mu y \right\} \, dy
\]

(7.25)

\[
    = \exp \{\lambda H(T) - n\mu T\} \cdot \int_{-\infty}^{\infty} \exp \left\{ -\delta \sqrt{\lambda\mu y} - \frac{1}{2}\lambda(g(T) + \epsilon)y^2 - f(\lambda)y \right\} \, dy
\]

(7.26)

\[
    \sim \exp \{\lambda H(T) - n\mu T\} \cdot \exp \left\{ \frac{\delta^2 \mu}{2(g(T) + \epsilon)} \right\} \cdot \sqrt{\frac{2\pi}{\lambda(g(T) + \epsilon)}}.
\]

(7.27)
where (7.26) follows from (7.20), and (7.27) is derived via straightforward calculations. Now define $J_{A,\epsilon}^+$, replacing $(g(T) + \epsilon)$ by $(g(T) - \epsilon)$ in (7.25). In addition, define $J_{\xi}$, $J_{A,\epsilon,\xi}^-$ and $J_{A,\epsilon,\xi}^+$ replacing $\int_T^\infty$ by $\int_\xi^\infty$ in (7.23) and (7.25).

We need to prove that, for some $\nu > 0$,

$$|J - J_{\xi}| = \exp\{\lambda H(T) - n\mu T\} \cdot o(e^{-\nu\lambda}),$$  \hspace{1cm} (7.28)

and that the same exponential bound prevails also for $|J_{A,\epsilon,\xi}^- - J_{A,\epsilon}^-|$ and $|J_{A,\epsilon,\xi}^+ - J_{A,\epsilon}^-|$.

In order to prove (7.28), define

$$\zeta = \frac{\bar{G}(T) - \bar{G}(T + \xi)}{2},$$

where $\zeta > 0$ since $G$ has positive density in $T$. Then, for large $\lambda$,

$$\int_\xi^\infty \exp \left\{ \int_T^{T+y} \left( \lambda \bar{G}(u) - \lambda \bar{G}(T) - \delta \sqrt{\lambda \mu} - f(\lambda) \mu \right) du \right\} dy$$

$$= \int_\xi^\infty \exp \left\{ \int_T^{T+\zeta/2} \ldots + \int_T^{T+\xi/2} \ldots \right\} dy$$

$$\leq \int_\xi^\infty \exp \left\{ -(\zeta/2)(\delta \sqrt{\lambda \mu} + f(\lambda) \mu) - \lambda \zeta (y - \xi/2) \right\} dy$$

$$= \frac{1}{\lambda \zeta} \cdot \exp \left\{ -(\xi/2)(\delta \sqrt{\lambda \mu} + f(\lambda) \mu) \right\} \cdot \exp \left\{ -\frac{\lambda \xi \zeta}{2} \right\} = o(e^{-\nu\lambda}).$$

Other exponential bounds are derived in a similar manner.

Now the Taylor expansion (7.24) of $\int_T^{T+y} \bar{G}(u) du$ implies that $\forall \epsilon > 0 \ \exists \xi > 0$ such that for $y \in [-\xi, \xi]$:

$$\bar{G}(T)y - \frac{1}{2} (g(T) + \epsilon/4) y^2 < \int_T^{T+y} \bar{G}(u) du < \bar{G}(T)y - \frac{1}{2} (g(T) - \epsilon/4) y^2,$$

which implies, combined with (7.23) and (7.25), that

$$J_{A,\epsilon/4,\xi}^- < J_{\xi} < J_{A,\epsilon/4,\xi}^+.$$ 

The exponential bounds (7.28) imply that, for large $\lambda$,

$$J_{A,\epsilon/2}^- < J < J_{A,\epsilon/2}^+,$$

and (7.27) implies that, for large $\lambda$,

$$\exp\{\lambda H(T) - n\mu T\} \cdot \exp \left\{ \frac{\delta^2 \mu}{2(g(T) + \epsilon)} \right\} \cdot \sqrt{\frac{2\pi}{\lambda (g(T) + \epsilon)}} \leq J$$ \hspace{1cm} (7.29)

$$\leq \exp\{\lambda H(T) - n\mu T\} \cdot \exp \left\{ \frac{\delta^2 \mu}{2(g(T) - \epsilon)} \right\} \cdot \sqrt{\frac{2\pi}{\lambda (g(T) - \epsilon)}}.$$
Since $\epsilon$ is arbitrary, we proved Lemma 7.1.

In the same way, via the Laplace method,

$$J(T) \triangleq \int_T^\infty \left\{ \lambda \int_0^x \tilde{G}(u) du - n\mu x \right\} dx$$

$$\sim \exp\{\lambda H(T) - n\mu T\} \cdot \exp\left\{ \frac{\delta^2 \mu}{2g(T)} \right\} \cdot \sqrt{\frac{2\pi}{\lambda g(T)}} \cdot \Phi \left( \delta \sqrt{\frac{\mu}{g(T)}} \right).$$

(7.30)

Finally,

$$\mathcal{E} \triangleq \lambda \int_0^\infty e^{-\lambda x}(1 + \mu x)^{n-1} dx$$

$$= \lambda \int_0^\infty \exp \left\{ -\lambda x + \left[ \frac{\lambda}{\mu} \tilde{G}(T) + \delta \sqrt{\frac{\lambda}{\mu} + f(\lambda) - 1} \right] \log(1 + \mu x) \right\} dx.$$  

Using the Laplace method and the Taylor expansion for $\log(1 + \mu x)$, we get that

$$\mathcal{E} \sim \frac{1}{G(T)},$$

(7.31)

which is asymptotically negligible compared to the $\lambda J$ term in the denominator of (7.21). Substitution of (7.22), (7.30) and (7.31) into (7.21) implies the following approximation for the tail probability under the staffing (7.20):

$$P\{W > T\} \sim \tilde{G}(T) \cdot \Phi \left( \delta \sqrt{\frac{\mu}{g(T)}} \right).$$

(7.32)

Hence, if the QED parameter (7.20) is given by:

$$\delta^* = \Phi^{-1} \left( \frac{\alpha}{G(T)} \right) \cdot \sqrt{\frac{g(T)}{\mu}},$$

then

$$P\{W > T\} \sim \alpha.$$  

1 $\Rightarrow$ 3. This result is a straightforward consequence of the ED results from [40]. Specifically, the probability to abandon of delayed customers is given by

$$P\{\text{Ab} \mid V > 0\} = \frac{1 + (\lambda - n\mu)J}{\lambda J},$$

where $J \to \infty$, as $\lambda \to \infty$, at an exponential rate. Moreover, $P\{V > 0\}$ converges to 1 at an exponential rate. Now, the asymptotics for $P\{\text{Ab}\}$ can be derived by substituting the ED+QED staffing level in Part 1 of the theorem into the expression $\frac{\lambda - n\mu}{\lambda}$.

1 $\Rightarrow$ 4. From [40], the average wait of the delayed customers in M/M/n+G is given by

$$E[W \mid V > 0] = \frac{J_H}{J},$$

(7.33)
where

\[ J_H = \int_0^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\} \, dx, \]

and \( J, H \) were defined in (7.15) and (7.17), respectively. Since the delay probability \( P\{V > 0\} \) converges to 1 at an exponential rate, the asymptotics for \( \mathbb{E}[W] \) and \( \mathbb{E}[W|V > 0] \) coincide. Now we shall calculate the asymptotics of (7.33) under the ED+QED staffing (7.20).

The variable-change \( y = x - T \) transforms (7.33) to

\[
H(T) + \frac{\int_T^{T+y} G(u) \, du}{\int_{-T}^{\infty} \exp\{\lambda \int_T^{T+y} G(u) \, du - n\mu y\} \, dy} \cdot 
\int_{-T}^{\infty} \exp\{\lambda \int_T^{T+y} G(u) \, du - n\mu y\} \, dy
\]

If we substitute into (7.34) the Taylor approximation (7.24) and replace \(-T\) by \(-\infty\) in the integrals, we get

\[
H(T) + \frac{\int_{-\infty}^{T+y} G(T) y \cdot \exp\{\delta \sqrt{\lambda \mu y} - \frac{1}{2} \lambda g(T) y^2\} \, dy}{\int_{-\infty}^{\infty} \exp\{\delta \sqrt{\lambda \mu y} - \frac{1}{2} \lambda g(T) y^2\} \, dy}
\]

\[
= H(T) + \frac{\int_{-\infty}^{T+y} \left( \frac{z}{\sqrt{\lambda g(T)}} - \delta \sqrt{\frac{\mu}{\lambda g(T)}} \right) e^{-z^2/2} \, dz}{\int_{-\infty}^{\infty} e^{-z^2/2} \, dz}
\]

\[
\sim H(T) - G(T) \cdot \frac{1}{\sqrt{\lambda}} \cdot \frac{\delta}{\sqrt{1-\gamma}} \cdot \frac{1}{n g(T)}.
\]

For a rigorous validation of the transition from (7.34) to (7.35), asymptotic equivalence of the numerators of the second terms in these formulae must be shown. This equivalence can be checked along the same lines of Lemma 7.1, and is omitted for brevity.

2, 3, 4 \( \Rightarrow \) 1. These statements can be proved in the spirit of Theorem 4.1 above, using monotonicity of the corresponding performance measures.

**Proof of Theorem 4.4 (ED+QED: constraint satisfaction)** The proof of Theorem 4.4 is similar to that of Theorem 4.1 and uses the equivalence between Statements 1 and 2 in Theorem 4.3.

### 7.2 Proof of Theorem 5.1

The proof of a proceeds via the following statements.
Statement 1.
\[
\limsup_{\lambda \to \infty} \frac{\sum_{i=1}^{k} c_i n_i^*(\lambda) - \sum_{i \in \bar{H}^*} c_i R_i(\lambda)}{\delta^* \sqrt{R(\lambda)}} \leq 1. \tag{7.36}
\]

Statement 2. For all \( i \in H^* \), \( P_{\lambda}^i \{ W > 0 \} \to 1 \), as \( \lambda \to \infty \).

Statement 3.
\[
\liminf_{\lambda \to \infty} \frac{\sum_{i=1}^{k} c_i n_i^*(\lambda) - \sum_{i \in \bar{H}^*} c_i R_i(\lambda)}{\delta^* \sqrt{R(\lambda)}} \geq 1. \tag{7.37}
\]

Statement 4. For all \( i \in H^* \), \( n_i^*(\lambda) = o(\sqrt{\lambda}) \).

Assume that \( \{ \beta_i^*, i \in \bar{H}^* \} \) is a solution (maybe not unique) of the optimization problem (5.4). Define the staffing level:
\[
\tilde{n}_i(\lambda) = 0, \quad i \in H^*, \tag{7.38}
\]
\[
\tilde{n}_i(\lambda) = \left[ R_i(\lambda) + (\beta_i^* + \epsilon_i) \cdot \sqrt{R_i(\lambda)} \right], \quad i \in \bar{H}^*, \tag{7.39}
\]
where \( \epsilon_i \) are non-negative and, at least, one of them is positive.

It follows from (4.11), (5.4) and monotonicity properties from the proof of Theorem 4.4 that this staffing is feasible for large \( \lambda \). Since \( \epsilon_i \) can be arbitrarily small, feasibility of (7.38)-(7.39) implies Statement 1.

Part a of Theorem 4.1 from Zeltyn and Mandelbaum [40] implies that if \( P^i \{ W > 0 \} \to c, 0 < c < 1 \), then there exists finite \( \beta \) such that
\[
n_i(\lambda) = R_i(\lambda) + \beta \sqrt{R_i(\lambda)} + o(\sqrt{\lambda}). \tag{7.40}
\]
Since the number of time intervals is finite, there exists a sequence \( \{ \lambda_k \} \to \infty \), such that for \( 1 \leq i \leq K \),
\[
P_{\lambda_k}^i \{ W > 0 \} \to \xi_i, \quad k \to \infty, \tag{7.41}
\]
where \( 0 \leq \xi_i \leq 1 \).

Define by \( F^* \) the set of states with \( \xi_i = 1 \) and by \( \bar{F}^* \) its complement. Due to (7.40),
\[
\liminf_{\lambda \to \infty} \frac{\sum_{i=1}^{K} n_i^*(\lambda_k)}{\sum_{i \in F^*} R_i(\lambda_k)} \geq 1. \tag{7.42}
\]
According to (7.36) and the optimality of \( \{ n_i^* \} \),
\[
\sum_{i \in F^*} r_i \leq \sum_{i \in H^*} r_i, \tag{7.43}
\]
and feasibility demands
\[
\sum_{i \in F^*} r_i \leq \alpha. \tag{7.44}
\]
Formulae (7.42)-(7.44), combined with the definition of \( \bar{H}^* \), imply \( F^* = H^* \) and \( \bar{F}^* = \bar{H}^* \).
Now it is straightforward to prove Statement 2. Assume that, for some \( i \in H^* \), \( P^i \{ W > 0 \} \rightarrow c \neq 1 \). Define a sub-subsequence \( \lambda_{k_i} \) such that the delay probabilities for all intervals converge along it and get a contradiction.

We proceed to Statement 3. Assume that there exists a subsequence \( \{ \lambda_k \} \rightarrow \infty \) such that

\[
\lim_{k \rightarrow \infty} \frac{\sum_{i=1}^k c_i n_i^*(\lambda_k) - \sum_{i \in H^*} c_i R_i(\lambda_k)}{\delta^* \sqrt{R(\lambda_k)}} < 1. \tag{7.45}
\]

Define the QED parameters for \( i \in \bar{H}^* \) via

\[
\beta_i(\lambda) = \frac{n_i^*(\lambda) - R_i(\lambda)}{\sqrt{R_i(\lambda)}}. \tag{7.46}
\]

There exists a converging sub-subsequence \( \{ \lambda_{k_l} \} \) such that

\[
\beta_i(\lambda_{k_l}) \rightarrow \tilde{\beta}_i, \quad \text{as } l \rightarrow \infty. \tag{7.47}
\]

If all limits in (7.47) are finite then, combining (7.45) and (7.46),

\[
\sum_{i \in H^*} c_i \tilde{\beta}_i \sqrt{\lambda_{k_l}} < \delta^*,
\]

which combined with (5.4) contradicts feasibility.

If, for some \( i \in \bar{H}^* \), \( \tilde{\beta}_i = -\infty \) then \( P^i \{ W > 0 \} \rightarrow 1 \), which, combined with Statement 2, contradicts feasibility. If, for some \( i \in \bar{H}^* \), \( \tilde{\beta}_i = +\infty \) then, combining (7.45) and (7.46), there should be \( \tilde{\beta}_j = -\infty \) for some \( j \in \bar{H}^* \). We got contradictions in all the special cases and proved Statement 3.

Now assume that there exists \( i \in H^* \), \( \epsilon > 0 \) and a sequence \( \lambda_k \rightarrow \infty \) such that

\[
n_i^*(\lambda_k) > \epsilon \sqrt{\lambda_k}, \quad \epsilon > 0.
\]

Construct converging subsequences \( \beta_i(\lambda_{k_l}) \) for \( i \in \bar{H}^* \). Consider the staffing level

\[
\bar{n}_i(\lambda_{k_l}) = \begin{cases} 0, & i \in H^*, \\ R_i(\lambda_{k_l}) + \beta_i \sqrt{R_i(\lambda_{k_l})} + \frac{\epsilon}{2K} \sqrt{\lambda_{k_l}}, & i \in \bar{H}^*. \end{cases}
\]

It is easy to check, via (4.11), that this staffing level is feasible for large \( k \) and

\[
\sum_{i=1}^K c_i \bar{n}_i(\lambda_{k_l}) < \sum_{i=1}^K c_i n_i^*(\lambda_{k_l}),
\]

which contradicts the optimality of \( n_i^*(\lambda_k) \). Hence, Statement 4 is proven.

In order to complete Part a, we must prove (5.6) and (5.8). The proof of (5.8) is similar to the proof of Statement 3. We should show that the QoS parameters cannot have any other limiting point, except the unique solution. In order to prove (5.6), we must show that along any subsequence of arrival rates, the QoS parameters for \( i \in \bar{H}^* \) cannot converge to \( \pm \infty \). The proof, again, is similar to Statement 3.

The proof of Part b directly follows from (4.11) and Part a.
8 Possible future research

To conclude, we outline several types of problems that we propose for future research.

- **Revenue/cost optimization.** As already discussed in the Introduction, optimization of revenues and/or costs constitutes an alternative to the approach of the present paper. The ongoing research [27] is dedicated to this problem for the M/M/n+G queue, continuing the work of Borst et al. [7] on Erlang-C.

- **Additional research on global constraint satisfaction.** Section 5 of the present paper gives rise to interesting research problems. For example, one could try to verify the conjecture at the end of Subsection 5.2. It would be also interesting to study the staffing level for several joint constraints, for example \( P\{\text{Ab}\} \) and \( P\{W > T\} \) (or rather \( P\{W > T, \text{Sr}\} \)). We believe that, in that case, unlike the single-interval problem, several constraints could be binding for an asymptotic solution. This will give rise to new asymptotic results that could be relevant in practice.

- **Time-inhomogeneous arrival rate.** Such queues are prevalent in practice and their time-varying analysis poses a challenge. A common approach is to approximate the time-varying arrival-rate by a piecewise-constant function, and then apply steady-state results during periods when the arrival rate is assumed constant. An implicit assumption is that the arrival rate is slow-varying with respect to the durations of services. Recently, Feldman et al. [15] developed an alternative simulation-based algorithm for staffing time-varying queues with abandonment in order to achieve a constant delay probability. We believe that a similar approach can be applied to other constraint satisfaction problems such as those analyzed in the present paper.

- **Generally distributed service times.** The M/M/n+G model assumes exponential services. However, this assumption does not apply for many call centers. For example, in several application (e.g. [11]) we encountered a lognormal distribution of service times. Therefore, it is very important to study the M/G/n+G model with a general service distribution. Whitt [37] suggests to approximate M/G/n+G by M/M/n+G with the same service mean. We believe that additional research is worthy in this direction.

- **Random arrival rate.** In Brown et al. [11] and Weinberg et al. [32] it was shown that Poisson arrival rates in two different call centers vary from day to day and the prediction of arrival rates raises statistical and practical challenges. Therefore, it is very important to study queueing models where the arrival rate \( \Lambda \) of a homogeneous Poisson arrival process is in fact a random variable. We expect that both the QED and ED regimes (and, maybe,
some new regimes) can be relevant in this case, depending on the order of the variation of \( \Lambda \). See Whitt [36], and Bassamboo, Harrison and Zeevi [4] for the “cruder” ED case.

References


44


[28] Palm C. (1957) Research on telephone traffic carried by full availability groups. Tele, vol.1, 107 pp. (English translation of results first published in 1946 in Swedish in the same journal, which was then entitled *Tekniska Meddelanden fran Kungl. Telegrafstyrelsen*.) 1.1, 3


47