Call centers with impatient customers: exact analysis and many-server asymptotics of the M/M/n+G queue

Sergey Zeltyn

Call centers with impatient customers: exact analysis and many-server asymptotics of the M/M/n+G queue

Research Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Sergey Zeltyn

Submitted to the Senate of the Technion - Israel Institute of Technology
Tishrei, 5765 Haifa October, 2004

The Research Thesis was done under the supervision of
Professor Avishai Mandelbaum
in the Faculty of Industrial Engineering and Management

The generous financial help of the Technion, Davidson Fund, Israel Science Foundation, Niedersachsen Fund, U.S.-Israel Binational Foundation and Wharton Financial Institutions Center is gratefully acknowledged

Contents

List of Figures

List of Tables

Abstract

The subject of the present research is the M/M/n+G queue. This queue is characterized by Poisson arrivals at rate λ, exponential service times at rate μ, n service agents and generally distributed patience times of customers. The model is applied in the call center environment, as it captures the tradeoff between operational efficiency (staffing cost minimization) and service quality (accessibility of agents).

First, we provide an extensive background on the M/M/n+G model and its most important special case: the Erlang-A queue with exponential patience times. In particular, we present an extensive list of formulae, many of which are new, for M/M/n+G performance measures.

The next part of our research is motivated by a phenomenon that has been observed in call center data: a clear linear relation between the probability to abandon P{Ab} and average waiting time E[W]. Such a relation is theoretically justifiable when customers' patience is exponential, but it lacks an explanation in general. We thus analyze its robustness within the framework of the M/M/n+G queue, which gives rise to further theory and empirically-driven experiments.

From the theoretical point of view, we establish order relations for performance measures of the M/M/n+G queues, and some light-traffic results. In particular, we prove that, with λ, μ, n and *average patience time fixed*, deterministic patience minimizes the probability to abandon and maximizes the average wait in queue.

In our experiments, we describe the behavior of M/M/n+G performance measures for different patience distributions. The findings are then related to our theoretical results and some observed real-data phenomena. In particular, clear non-linear relations (convex, concave and mixed) emerge between the probability to abandon and average wait. However, when restricted over low to moderate abandonment rates, approximate linearity prevails, as observed in practice.

In the last part of the research, three asymptotic operational regimes for medium to large call centers are introduced and studied. These regimes correspond to the following three staffing rules, as λ and n increase indefinitely and μ held fixed:
Efficiency-Driven (ED): $n \approx (\lambda/\mu)\cdot(1-\gamma)\,, \quad \gamma > 0,$
Quality-Driven (QD): $n \approx (\lambda/\mu)\cdot(1+\gamma)\,, \quad \gamma > 0,$ and
Quality and Efficiency Driven (QED): $n \approx \lambda/\mu + \beta\sqrt{\lambda/\mu}\,, \quad -\infty < \beta < \infty.$

In the ED regime, the probability to abandon and average wait converge to constants. In the QD regime, we observe a very high performance level at the cost of possible overstaffing. Finally, the QED regime carefully balances quality and efficiency: agents are highly utilized, but the probability to abandon and the average wait are small (converge to zero at rate $1/\sqrt{n}$). In addition, in the QED and ED regimes, we establish an asymptotic linear P{Ab}/E[W] relation.

Numerical experiments demonstrate that for a wide set of system parameters, the QED formulae provide excellent approximation for exact M/M/n+G performance measures. In turn, the much simpler ED approximations are very useful for overloaded queueing systems.

Finally, our theoretical results are applied to call-by-call data of a large bank. Several interesting phenomena concerning applications of the QED approximations and linear P{Ab}/E[W] relation are observed and studied.

List of Symbols

$G(\cdot)$ distribution of patience time

$\bar{G}(\cdot)$ survival function of patience time $(\bar{G} = 1 - G)$

$g(\cdot)$ density of patience time $(g = G')$

g_0 value of patience density at the origin $(g_0 = g(0))$

g_{0k} k-th derivative of patience density at the origin

$h(\cdot)$ hazard rate of the standard normal distribution

n number of agents

Q queue length

R offered load $(R = \lambda/\mu)$

S service time

V offered waiting time

$v(\cdot)$ density of offered waiting time

W actual waiting time

β service grade in the QED operational regime

γ service grade in the ED and QD operational regimes

Δ safety-staffing level

θ individual abandonment rate

λ arrival rate

μ service rate

ρ offered load per agent $(\rho = \lambda/(n\mu))$

τ patience time

$\bar{\Phi}(\cdot)$ survival function of the standard normal distribution

P{Ab} probability to abandon

P{Blk} balking probability

P{Sr} probability to get service

$f(\lambda) \sim g(\lambda) \qquad \lim_{\lambda \to \infty} f(\lambda)/g(\lambda) = 1$

List of Acronyms

A/S/n/m+P Notation for describing queueing models where

- **A** indicates interarrival distribution
- **S** indicates service-time distribution
- **n** indicates number of agents
- **m** indicates system capacity (maximal number in queue plus service)
- **P** indicates patience distribution

ACD Automatic Call Distributor

AHT Average Handling Time

ASA Average Speed of Answer

CTI Computer-Telephone Integration

ED Efficiency-Driven

EOS Economies Of Scale

FCFS First Come First Served

HR Human Resources

KM Kaplan-Meyer

MLE Maximum-Likelihood Estimator

PASTA Poisson Arrivals See Time Averages

QD Quality-Driven

QED Quality and Efficiency Driven

SBR Skill-Based Routing

TSF Total Service Factor

Part I
Introduction

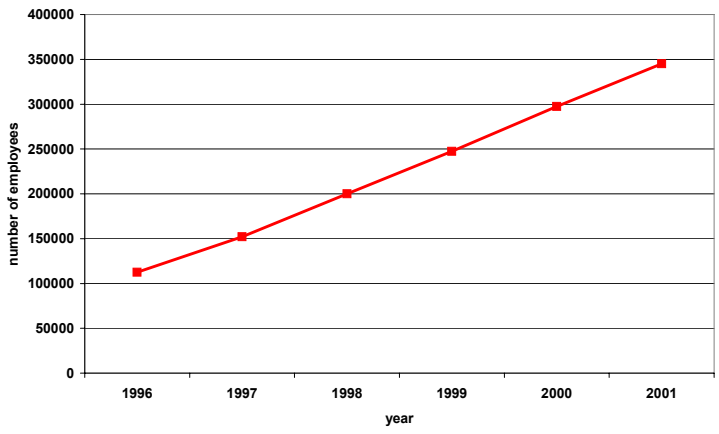
1 Some facts about call centers

1.1 The call center industry: quality/efficiency tradeoff

During the last two decades, there has been an explosive growth in the number of companies that provide services via the telephone, as well as in the variety of telephone services provided. One study [4] estimates the number of call center employees in 1999-2000 at between four to almost nine millions in the USA (3-6% of the total workforce), 600,000 in the UK (2.3%) and 200,000 in Holland (almost 3%). Gans, Koole and Mandelbaum [27], citing [19, 66], report a more conservative estimate for the USA: 1.55 million for the private sector in 1999.

Recent research of Rafaeli [59] reports that in Israel there are currently approximately 500 call centers. The number of employees in the 200 largest call centers is near 11,000. (Which relatively to the total population, is less than in the three countries referred to above.) Most Israeli call centers are operating in banking, medical care, insurance, communication, tourism, transport, emergency services and the food industry.

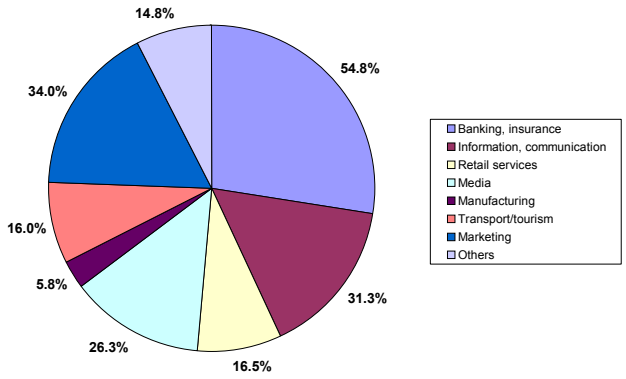
Figure 1: **Number of call center employees in Germany**



As an additional representative example, consider the dynamics of call center employment in Germany (Figure 1). (The averages of upper and lower estimates, provided in [8], are shown.) Here we observe more than three-fold increase over five years.

Figure 2 [8] shows a sectoral distribution of call centers in Germany. (The numbers do not sum up to 100% since a single call center can provide several types of service.) Note a large share of banking call centers in the industry, which is the sector that our practical examples are taken from.

Figure 2: **Sectoral distribution of call centers in Germany**



Technological progress has significantly affected the development of the call-center industry. Computer-Telephone Integration (CTI) provides numerous opportunities for combining telephone service with e-mail and Internet services. Consequently, many call center evolve to so-called *contact centers*. Automated Speech Recognition techniques help extend the number of tasks that the Voice Response Units (VRU), traditionally touch tone, have been able to perform. (Eventually, speech recognition and/or Internet will probably annihilate the trend, shown in Figure 1, in the same way as the invention of the automatic call distributor dramatically decreased the number of employees in the telephone industry [69].)

A central characteristic of a call center is whether it handles *inbound* or *outbound* traffic. Inbound call centers handle incoming calls that are initiated by customers calling

in to a center. For example, these types of centers provide customer support, help desk services, reservation and sales support for airlines and hotels, and order-taking functions for catalog and web-based merchants. Outbound call centers handle outgoing calls, calls that are initiated from within a center. These types of operations have been traditionally associated with tele-marketing and survey businesses.

In this research, we shall focus on models that are more appropriate for inbound call centers. We shall not explore VRU models (that environment rarely gives rise to queueing).

A central challenge in designing and managing a service operation in general, and telephone-based services in particular, is to achieve a desired balance between *operational efficiency* and *service quality*. Note that these concepts are multi-dimensional. Below we explain which aspects of quality and efficiency are of most interest for our research.

Although telephone call centers are technology-intensive operations, often 70% or more of their operating costs are devoted to Human Resources (HR). Hence, effective HR management is crucial in the call center environment. HR management challenges usually belong to one of the following two types:

- In the yearly and monthly time scale, one should use effective *hiring* and *training* policy [29]. Indeed, call centers are characterized by very high turnover rates. Bordoloi [7] reports that the turnover rate in the US call centers for the year 2000 was approximately 50%. This is consistent with our experience for Israeli call centers. According to [7], the average cost to recruit and train a new agent is more than \$6000. Therefore, call center managers should carefully account for turnover and long-term demand trends.

- At the weekly and daily levels, *queueing* and *scheduling* models should be used. Queueing models are used to determine how many agents must be available to serve calls over a given half-hour or hour; scheduling models determine time periods during the day, week or month, when each agent will work.

Our research deals with queueing models. See [27] for a review on scheduling and hiring models.

Now we explain the aspects of service quality that are of most interest in this study. Service challenges can be classified, in large, to three main categories.

- *Accessibility* of agents. Here typical questions are: "How long did customers have to wait (if at all) to speak to an agent? How many abandoned the tele-queue before being served?"

- *Effectiveness* of service. The key question is "Has the customer's problem been resolved?"

- The characteristics of agents' *interactions* with customers. (Important factors are agent's politeness, number of transfers encountered by a customer, etc.)

In our research, we consider the first aspect of service quality, which can be called *operational quality*. We thus focus on the phenomena of waiting in queues and abandonment, which are common in call centers and can be relatively easily formalized via mathematical models.

Summarizing, we treat the following aspect of the *quality/efficiency tradeoff*: having the right number of agents in place at the right times. "The right number" means not too many, saving operating costs, and not too few, avoiding excessive customers' wait and abandonment.

The quality and efficiency levels of a well-run modern call center can be extraordinary. In a large performance-leader enterprize, thousands of agents could serve many thousands of calling customers per hour; agents' utilization levels exceed 90%, yet about 50% of the customers are answered immediately upon calling; callers who are delayed demand a response within seconds, the vast majority gets it, and scarcely few of the rest, say 1% of those calling, abandon during peak-congestion due to impatience. But most call centers (in particular Israeli call centers) are far from achieving such levels of performance. To these, scientific models are prerequisites for climbing the performance ladder, and such models are the subject of the present thesis.

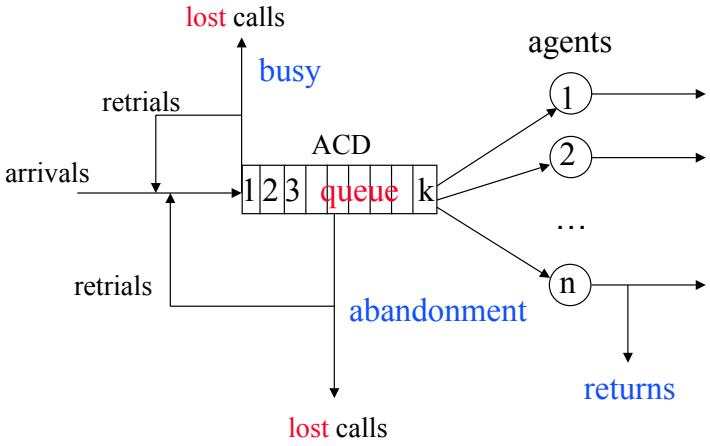
1.2 Call centers as queueing systems

In light of the quality/efficiency tradeoff challenge, as posed above, it is natural to model call centers by queueing systems. Note that unlike many other queues, call center queues are *invisible*: callers cannot observe how long a queue is and their progress in it. Conse-

quently, the abandonment behavior of customers in tele-queues is different from face-to-face queues.

For example, customers in a bank face-to-face service are probably unhappy at the start of their wait, but become more and more satisfied approaching the teller. In addition, it is easier for face-to-face customers to make an *informed choice*: is it worth to join the queue? In contrast, the irritation of tele-customers often increases with time spent in queue. In fact, informing customers about their expected wait is the way many call centers choose to make their tele-queue partly visible. (But see the discussion on the second plot of our subsequent Figure 5.)

Figure 3: **Schematic representation of a telephone call center**



Modelling a Call Center. A simplified representation of traffic flows in a call center is given in Figure 3. Incoming calls form a single queue, waiting for service from one of n statistically identical agents. There are $k+n$ telephone trunk-lines. These are connected to an Automatic Call Distributor (ACD) which manages the queue, connects customers to available agents, and also archives operational data. Customers arriving when all lines are occupied encounter a busy signal. Such customers might try again later ("retrial") or give up ("lost call"). Customers who succeed in getting through at a time when all agents are busy (that is, when there are at least n but fewer than $k+n$ customers within the call center), are placed in the queue. If waiting customers run out of patience before their

service begins, they hang up ("abandon"). After abandoning, customers might try calling again later while others are lost. After service, there are "positive" returns of satisfied customers, or "negative" returns due to complaints.

Note that the model in Figure 3 ignores multiple service types and skilled-based routing that are present in many modern call centers. However, a lot of questions still remain open even for models with homogeneous servers.

In basic models, the already simple representation in Figure 3 is simplified even further. Specifically, in this study we assume that there are enough trunk-lines to avoid busy signals ($k = \infty$). This assumption materializes in many of today's call centers. In addition, we assume out retrials and return calls, which corresponds to absorbing them within the arrivals. This is reasonable to do when retrials/returns are not too immediate. However, and unlike most models used in practice, we do acknowledge and accommodate abandonment. The reasons for this will become clear momentarily.

2 Patience in invisible queues

A classical M/M/n[1] queueing model, also called the **Erlang-C** model, is the one most frequently used in *workforce management* of call centers. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed service times with a rate μ, and n independent statistically-identical agents. (Time-varying arrival rates are accommodated via piecewise constant approximations.) But Erlang-C does not allow abandonment. This, as will now be argued, is a significant deficiency: customer abandonment is not a minor, let alone a negligible, aspect of call center operations.

According to the "Help Desk and Customer Support Practices Report, 1997" [36], more than 40% of call centers set a target for fraction of abandonment, but in most cases this target is not achieved. Moreover, the lack of understanding of the abandonment phenomenon and the scarcity of models that acknowledge it, has lead practitioners to ignore it altogether. (For example, this lead [16] to conclude that abandonment is "not a good indicator of call center performance".) But models which ignore abandonment either distort or fail to provide information that is important to call center managers. We now support this last statement, first qualitatively and then quantitatively.

[1] See List of Acronyms for a review of conventional notation in queueing theory.

- Abandonment statistics constitute the only ACD measurement that is customer-subjective: those who abandon declare that the service offered is not worth its wait. (Other ACD data, such as average waiting times, are "objective"; also, ACD reports do not include the only other customer-subjective operational measures, namely retrial/return statistics.)

- Some call centers focus on the average waits of only those who get served, which does not acknowledge abandoning customers. But under such circumstances, the service-order that optimizes performance is LIFO = Last-In-First-Out! [28] which clearly suggests that a distorted focus has been chosen.

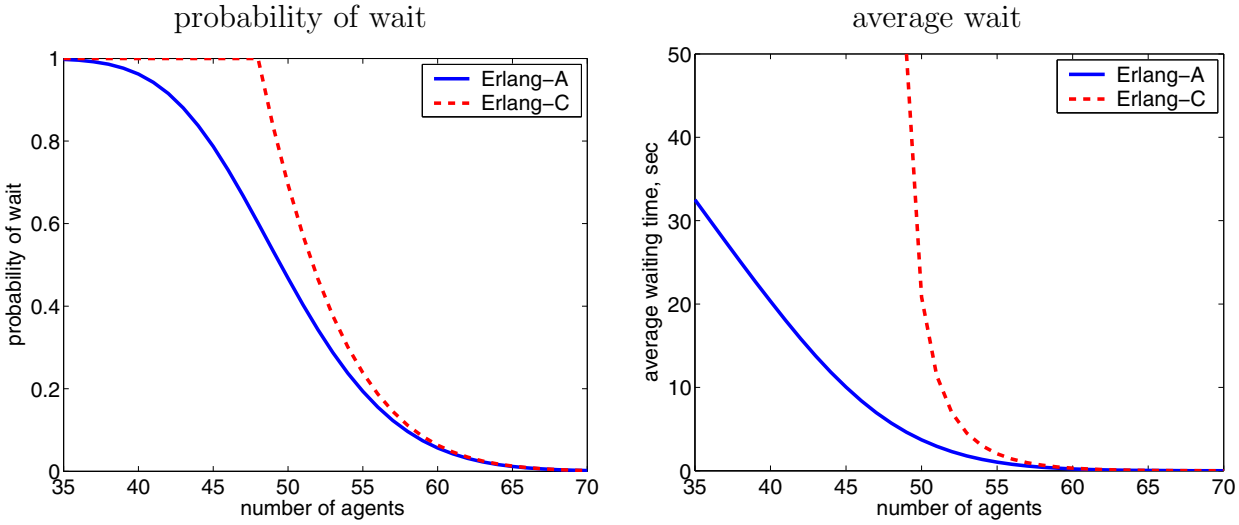
- Ignoring abandonment can cause either under- or over-staffing: On the one hand, if service level is measured only for those customers who reach service, the result is unjustly optimistic since the effect of an abandonment is less delay for those further back in line, as well as for future arrivals. This would lead to under-staffing. On the other hand, using workforce management tools that ignore abandonment would result in over-staffing since, with abandonment acknowledged, fewer agents would be needed in order to meet most abandonment-ignorant service goals.

Erlang-A: The simplest (most tractable) way to model abandonment, is to enrich Erlang-C (M/M/n) in the following manner. Associated with each arriving caller there is an exponentially distributed *patience time* with mean θ^{-1}. An arriving customer encounters an *offered waiting time*, which is defined as the time that this customer would have to wait given that her or his patience is infinite. If the offered wait exceeds the customer's patience time, the call is then abandoned, otherwise the customer awaits service. The patience parameter θ will be referred to as the *individual abandonment rate*. (We shall omit "individual", when obvious.) Assuming independence among inter-arrival times, individual patience times and service times yields a Markov model (Birth & Death Process) which we denote by M/M/n+M, or **Erlang-A** (**A** for **A**bandonment, and for the fact that it interpolates between Erlang-C and Erlang-B, the latter being M/M/n/n).

With Erlang-A, the quantitative significance of abandonment can be demonstrated through simple numerical examples. We start with Figure 4, which shows the fraction of delayed customers and the average wait in an M/M/n and a corresponding M/M/n+M

model. (The first plot of Figure 4 is adapted from Garnett et al. [29].) In both models, the arrival rate is 48 calls per minutes, the average service time equals 1 minute, and the number of agents is varied from 35 to 70. Average patience is taken to be 2 minutes for the Erlang-A model. Clearly, the two curves convey a rather different picture of what is happening in the system they depict, especially within the range of 40 to 50 agents: as shown below, Erlang-C is stable only with 49 or more agents, while Erlang-A is always stable. (See Subsection 6.1.1.)

Figure 4: **Comparison between Erlang-A and Erlang-C**
48 calls per min., 1 min. average service time, 2 min. average patience



The above M/M/n and M/M/n+M models are further compared in Table 1, using the 4CallCenters software [22][2].

Note that exponential patience with an average of 2 minutes gives rise to 3.1% abandonment. Then note that the average wait and queue length are both strikingly shorter, and this with only 3.1% abandonment taking place. (Significantly, this high-level performance is *not* achieved if the arrival rate to the M/M/n system is merely decreased by 3.1%; for example, the "average speed of answer" in such a case is 8.8 seconds, compared with 3.7.) Finally, note that system performance in such heavy traffic is very sensitive to staffing levels - adding 3 or 4 agents to M/M/n would result in M/M/n+M performance, as emerging from the horizontal distance between the graphs in Figure 4. Nonetheless,

[2]We shall demonstrate several applications of this useful program in Subsections 6.1.2 and 6.1.4.

Table 1: **Comparing models with/without abandonment**
50 agents, 48 calls per min., 1 min. average service time, 2 min. average patience

	M/M/n	M/M/n+M	M/M/n, $\lambda \downarrow 3.1\%$
Fraction abandoning	–	3.1%	-
Average waiting time	20.8 sec	3.7 sec	8.8 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec	28.2 sec
Average queue length	17	3	7
Agents' utilization	96%	93%	93%

since personnel costs are the major operational costs of running call centers (as already mentioned, prevalent estimates run at about 60-70% of the total), even a 6%-8% reduction in personnel is economically significant (and much more so for large call centers that employ thousands of agents).

M/M/n+G: The Erlang-A model assumes exponentially distributed patience times. Is this assumption valid in practice?

Figure 5: **Bank data: hazard rates of patience times**

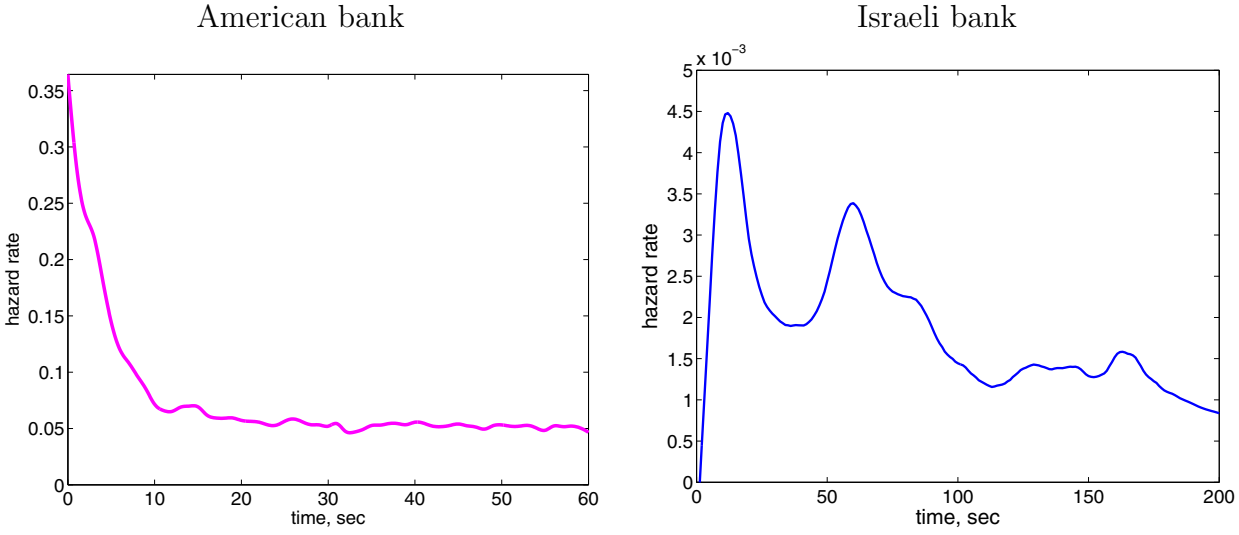
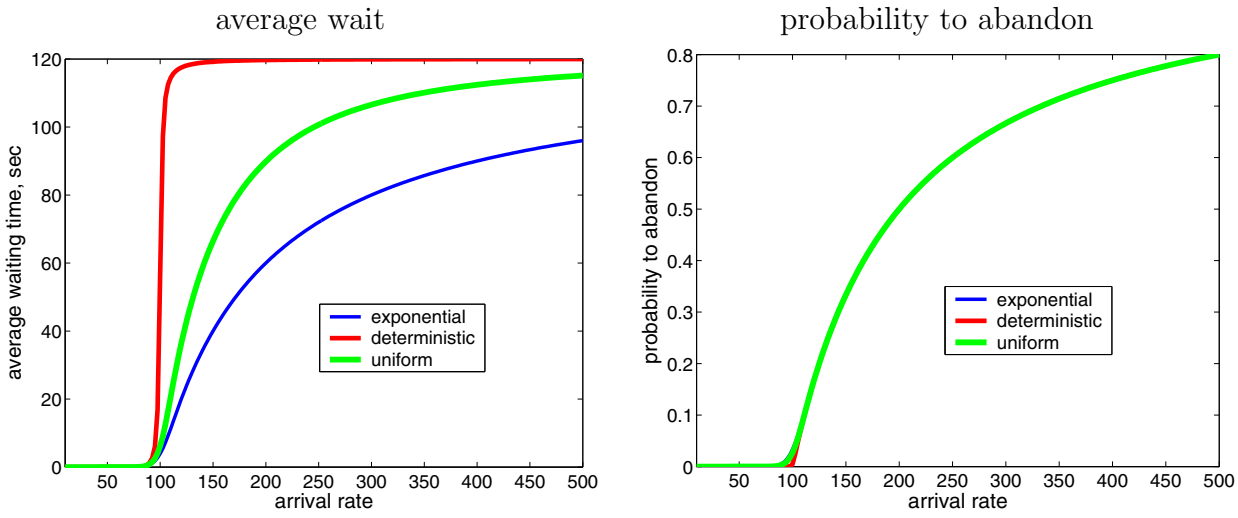


Figure 5 provides two examples of negative answers to the last question. Here we display the hazard-rate estimates of customers' patience for two banks: a large American bank (the related data will be studied in Section 19) and a small Israeli one. Recall that the hazard rate of an exponential random variable is a constant.

In the two cases we observe different, but clearly non-exponential patterns. American customers are very impatient at the beginning of their wait, but their patience stabilizes after approximately 10 seconds. In contrast, Israeli customers have two clear peaks of abandonment: approximately at 15 and at 60 seconds. (It turns out that these two surges of abandonment take place immediately after two recorded messages to which customers are exposed: the first one when they enter the queue and the second after approximately 1 minute.)

Figure 6: **Dependence of performance on patience distribution**



Therefore, at least in some applications (according to our experience, in most), customers' patience times are non-exponential. In this research, we study the **M/M/n+G** model, which allows a general distribution of patience times G (rather than exponential in M/M/n+M).

We now show that non-exponential distribution of patience can significantly affect system performance, when compared to the Erlang-A system with the same average patience. Figure 6 is related to the research reported in Section 12. We consider M/M/n+G queues with $n = 100$ and service rate $\mu = 1$. Three patience distributions with the same average patience are compared: exponential with average 2, constant 2 and uniform on [0,4]. We varied the arrival rate λ from 10 to 500, in step 2.5, calculating two performance measures: the probability to abandon and average wait.

Observe that the two plots of Figure 6 are not similar. The three average-wait curves are very different, except for small values of λ when the wait is negligible. In contrast, the probability-to-abandon curves seem almost identical. Only if we zoom around $\lambda = 100$, there is a noticeable difference between deterministic patience and the two other distributions.

This example shows that

- Patience-distribution can significantly affect performance of the M/M/n+G queue;
- The effect of patience distribution strongly depends on the performance measure we consider (average wait and probability to abandon, in our example);
- The effect of patience distribution depends on the load the system is working under. Specifically, it is very important if the *offered load per agent* $\frac{\lambda}{n\mu}$ is significantly below 1, around 1, or significantly above 1.

Figure 6 gives rise to questions that will be answered in subsequent parts of our research. For example:

- Why are the three curves in Figure 6 ordered in that specific way? (See Lemma 11.1 and Theorem 11.1 in Section 11.)
- Why the average waits for highly-loaded system ($\lambda >> 100$) are very different and the probabilities-to-abandon are almost the same? (See Section 17.)

3 Operational measures of performance

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 (thousands of agents), the implementation of these calculations must be both scalable and numerically stable.

3.1 Practical measures: Waiting Time

The most popular measure of operational (positive) performance is $\mathrm{P}\{W \leq T, \mathrm{Sr}\}$, where W is the waiting time, $\{\mathrm{Sr}\}$ is the event "customer gets service" and T is a target time

that is determined by Management/Marketing. For example, in a call center that caters to emergency calls, $T = 0$ (or T very small) would be appropriate. A common rule of thumb (without any theoretical backing, as far as we know), is the goal that at least 80% of the customers be served within 20 seconds; formally, $P\{W \leq 20, \text{Sr}\} \geq 0.8$. To this, one sometimes adds E[W], or $\text{E}[W|W > 0]$, as some measure of an average (negative) experience for those who waited.

An important measure that is rarely used in practice is $P\{W > 0\}$, the fraction of customers who encounter a delay. This is a useful stable measure of congestion. Its importance stems from the fact that it identifies an organization's operational focus, in the following sense:

- $P\{W > 0\}$ close to 0 indicates a Quality-Driven operation, where the focus is on *service quality*;
- $P\{W > 0\}$ close to 1 indicates an Efficiency-Driven operation, where the focus is on *servers' efficiency* (in the sense of high servers' utilization);
- $P\{W > 0\}$ strictly between 0 and 1 (for example 0.5) indicates a careful *balance* between Quality and Efficiency, which we abbreviate to **QED = Quality & Efficiency Driven** operational regime.

The above three-regime dichotomy is rather delicate. For example, consider a system in which customers' average patience is close to the average service duration (for example, let both be equal to one minute), and assume that its offered load λ/μ is 100 Erlangs. Then, staffing of 100 servers would lead to the QED regime, with high levels of both service and efficiency that are balanced as follows: about 50% of the customers are served immediately upon arrival, the average wait is 2.3 seconds, 4% of the customers abandon due to their impatience, and servers' utilization levels are 96%. The QED regime still prevails at staffing levels between 95 and 105. With 90 servers, the system is efficiency-driven: 11% of the customers abandon, only 15% served immediately, and utilization is over 99%. With 110 agents, it is quality-driven: abandonment is less than 1%, and 83% are served immediately.

Significant part of our research (Part IV) will explore the three operational regimes, introduced above.

3.2 Practical measures: accounting for Abandonment

In a quality-driven service, $P\{W > 0\}$ seems the "right" measure of operational performance. Thus, consider hereafter an operation in which $P\{W > 0\}$ is not close to vanishing.

As explained before, performance measures must take into account those customers who abandon. Indeed, if forced into choosing a *single* number as a proxy for operational performance, we recommend the probability to abandon $\mathbf{P\{Ab\}}$, the fraction of customers who explicitly declare that the service offered is not worth its wait. Some managers actually opt for the refinement $\mathrm{P}\{W > \epsilon; \mathbf{Ab}\}$, for some small $\epsilon > 0$, for example $\epsilon = 3$ seconds. The justification is that those who abandon within 3 seconds can not be characterized as poorly served. There is also a practical rational that arises from physical limitations, specifically that such "immediate" abandonment could in fact be a malfunction or an inaccuracy of the measurement devices.

The single abandonment measure $\mathrm{P}\{\mathrm{Ab}\}$ can be refined to account explicitly for those customers who were in fact well-served. Thus, we propose:

- $P\{W \leq T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{\mathrm{Ab}\}$ - fraction of poorly-served.

A further refinement, that yields a four-dimensional service measure, could be:

- $P\{W \leq T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{W > T; \mathrm{Sr}\}$ - fraction of served, with a potential for improvement (say, a higher priority on their next visit);
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served;
- $P\{W \leq \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined - see above for an elaboration.

4 Research motivation and objectives

4.1 Impact of customers' patience on delay and abandonment

In Section 2 we presented a hierarchy of queueing models: starting from the classical (and inadequate for the call center environment) Erlang-C, continuing to the simplest abandonment model M/M/n+M and ending up with M/M/n+G, the model with general customers' patience.

Garnett et al. [29] analyzed the M/M/n+M (*Erlang-A*) model, in which customers' patience is exponentially distributed. "Rules of thumb" for the design and staffing of medium to large call centers were then derived. In Brown et al. [13] the following noteworthy facts were established: patience patterns could be far from exponential, yet, in many aspects, the Erlang-A formulae provide useful accurate approximations for observed performance and data characteristics. (See Subsection 6.1.3.)

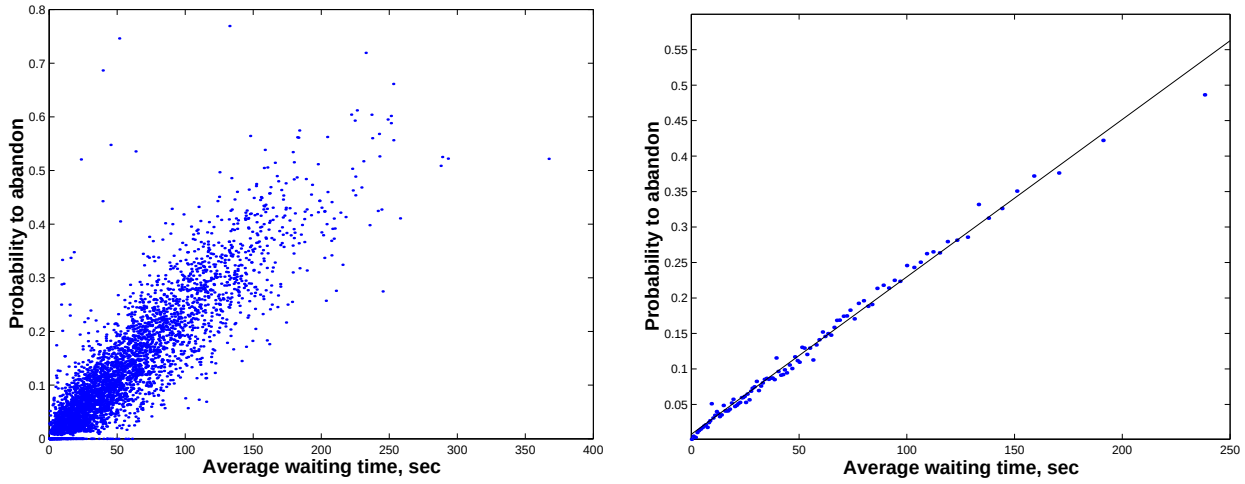
An important question thus arises: how robust is the M/M/n+M model with respect to deviations in its characteristics? In Part III of the thesis we answer it for customers (im)patience, which is natural to pursue within the M/M/n+G framework. This model assumes that customers arrive to the queueing system equipped with *patience times* τ that are G-distributed, iid across customers. A customer that has to wait for service more than τ abandons.

Above we underlined the significance of the probability to abandon as an operational measure. It is thus theoretically important and practically useful to identify functional relations between the probability of abandonment and other performance measures. Such relations could be used, for example, in predicting some performance characteristics from knowledge of others.

The following example gives a flavor of the problems, practical and theoretical, that motivate Part III of our research. Its objective is to relate two performance measures in steady state: the probability to abandon P{Ab} and the average waiting time in queue E[W]. (Here and in the sequel, E[W] is the average wait of *all* customers, either served or abandoning.)

Figure 7 displays an empirical relationship between the two measures. It was plotted using the yearly data of the Israeli bank call center [14], analyzed in Brown et al. [13] and Mandelbaum et al. [49]. First, the probability to abandon and average wait were

Figure 7: **Probability to abandon vs. average waiting time**



computed for the 4158 hour intervals that constitute the year 1999. The left plot of Figure 7 presents the resulting "cloud" of points, as they scatter on the plane. For the right plot, we are using an aggregation procedure that is designed to emphasize dominating patterns. Specifically, the 4158 intervals were ordered according to their average waiting times, and adjacent groups of 40 points were aggregated (further averaged): this forms the 104 points of the second plot in Figure 7. (The last point of the aggregated plot is an average of only 38 hour intervals.) We checked that, in fact, the regression lines for the two plots in Figure 7 are nearly identical.

The linear fit that emerges from the graphs is remarkable. And indeed, if W denotes waiting time and τ patience time, the law

$$\mathrm{P}\{\mathrm{Ab}\} = \frac{\mathrm{E}[W]}{\mathrm{E}[\tau]} \tag{4.1}$$

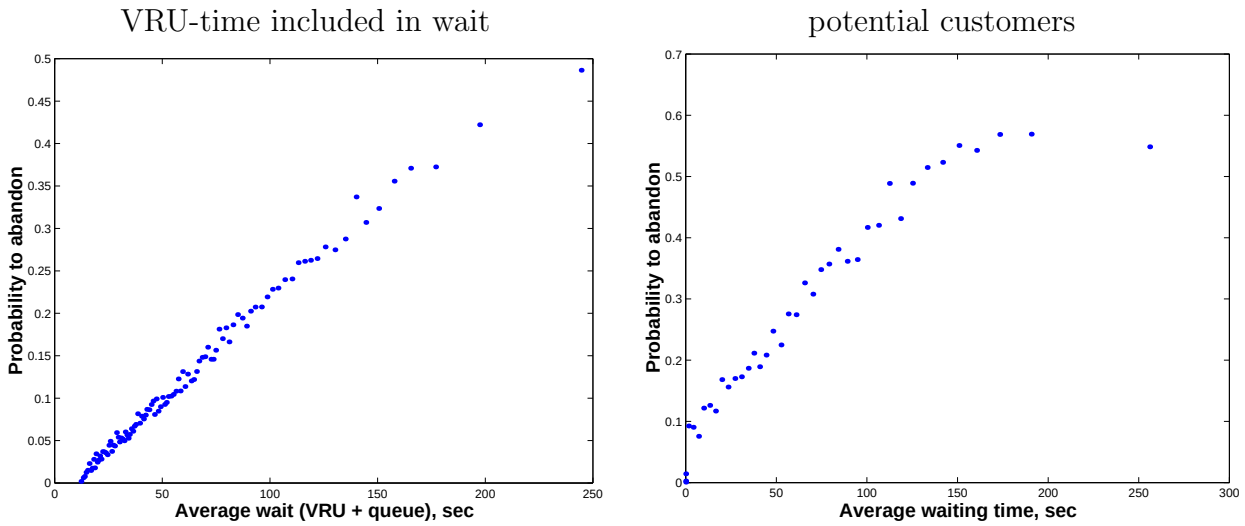
is provable for models with *exponential* patience (see Subsection 6.1.2). But, as will now be recalled from Section 2, this obviously is *not* our case: the hazard rate of patience is far from being constant, as it should have been if τ was exponential.

The second plot of Figure 5 from Section 2 shows the estimate of the hazard rate for patience of regular customers (approximately 70% of all customers; other types of customers exhibit similar patterns). The Kaplan-Meier estimate (for example, see [17] or Subsection 8.4) and a smoothing algorithm [26] were performed in order to produce that curve. The hazard pattern is clearly nonlinear, hence the pattern of the patience

distribution is far from exponential.

Two other examples of a linear relation between the probability to abandon and average wait are presented in Figure 8. Both are based on the same yearly call center data, referred to above. The first plot takes into account all customers. It differs from Figure 7 by its definition of waiting time: here it includes also the time spent by customers in the VRU (Voice Response Unit). The second plot is for a specific type of customer: potential customers asking for information on available services (approximately 15% of phone calls over the year). Both plots aggregate data along the guidelines of the second graph of Figure 7.

Figure 8: **Probability to abandon vs. average waiting time**



The points on the first plot are approximated by a straight line that intersects the x-axis around 10 seconds (average time spent in the VRU). The pattern of the second plot is close to a straight line with a positive intercept of the y axis, approximately at 0.1 (except for a couple of points near the origin and the point with the largest wait). Theoretically, a similar relation arises in the case of exponential abandonment with *balking*: customers' patience is equal to zero with probability p (immediate abandonment if wait is encountered) and it is exponential with probability $(1-p)$. The relation between the theoretical balking model and the second empirical plot in Figure 8 can be partially clarified using [49]. It was shown there that potential customers are less patient than customers overall, and they often abandon during their first seconds of wait (prior to the

first peak in Figure 5). This explains balking but does not fully account for the linear pattern.

The above examples give rise to a more general question: how do patience patterns affect queueing system performance? In particular, it is important to understand the circumstances under which one can practically use simple relations that, theoretically, apply perhaps only to models with exponential patience. This issue will be studied in our research (Part III and, to some extent, Part IV), both theoretically and via numerical experiments.

4.2 Asymptotic operational regimes for the M/M/n+G queue

Consider Table 2, which displays a typical daily ACD report of a moderate-to-large call center in the US, from the Health Insurance industry. For every half-hour interval, the report depicts the number of incoming calls, abandonment fraction, the Average Speed of Answer (ASA), the Average Handling Time (AHT), the agents' occupancy and the average number of agents over the interval in consideration.

We observe that performance level, presented by ASA and Abn%, varies significantly over the day. We shall concentrate on three time intervals, highlighted in bold: 13:30, 14:30 and 17:00.

The first interval is characterized by 100% occupancy, relatively high abandonment rate (9.4%) and considerable ASA (more than 1 minute). During this half-hour, the call center is working in the *Efficiency-Driven (ED) regime*, in the sense that the emphasis is on agents' utilization, or efficiency. (Note that the number of agents is smaller than in the adjacent intervals. A probable cause could be lunch break.)

The interval that starts at 17:00 presents a contrasting service pattern. There is no abandonment and the average wait is negligible (2 sec). The agents' occupancy is far below 100% (83%). Such a service regime will be called *Quality-Driven (QD)*, in the sense that the emphasis is on customers' service quality.

Finally, the last interval (14:30) demonstrates an intermediate service regime. Although utilization is high (96.6%), it is not 100%. Abandonment and waiting are neither negligible nor high. Since in this half-hour, high efficiency and service level are achieved simultaneously, this operational regime has been called *QED (Quality and Efficiency-*

Driven).

Table 2: **Example of Half-Hour ACD Report**

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
Total	20,577	19,860	3.5%	30	307	95.1%	
8:00	332	308	7.2%	27	302	87.1%	59.3
8:30	653	615	5.8%	58	293	96.1%	104.1
9:00	866	796	8.1%	63	308	97.1%	140.4
9:30	1,152	1,138	1.2%	28	303	90.8%	211.1
10:00	1,330	1,286	3.3%	22	307	98.4%	223.1
10:30	1,364	1,338	1.9%	33	296	99.0%	222.5
11:00	1,380	1,280	7.2%	34	306	98.2%	222.0
11:30	1,272	1,247	2.0%	44	298	94.6%	218.0
12:00	1,179	1,177	0.2%	1	306	91.6%	218.3
12:30	1,174	1,160	1.2%	10	302	95.5%	203.8
13:00	1,018	999	1.9%	9	314	95.4%	182.9
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**
14:00	1,173	1,082	7.8%	78	313	99.5%	188.9
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**
15:00	1,137	1,122	1.3%	15	320	96.9%	205.8
15:30	1,169	1,137	2.7%	17	311	97.1%	202.2
16:00	1,107	1,059	4.3%	46	315	99.2%	187.1
16:30	914	892	2.4%	22	307	95.2%	160.0
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**
17:30	420	420	0.0%	0	328	73.8%	103.5
18:00	49	49	0.0%	14	180	84.2%	5.8

The ACD report in Table 2 does not include a fundamental performance measure – the fraction of customers that encounter delay, namely the probability of wait. (Recall the discussion in Subsection 3.1.) It is natural to assume that during the ED-interval essentially all customers were delayed. Due to a negligible average wait, we expect that the probability of wait was very small during the QD-interval. During the QED interval, in contrast and as will be explained and demonstrated later, the fraction of delayed customers is expected to be neither close to zero nor to one.

Now, using the example above, we shall present formal definitions of the three operational regimes. We start with calculating the *offered load* $R = \dfrac{\lambda}{\mu}$ for the three intervals. We get

$$R_{ED} \;=\; 1061 : \frac{1800}{306} \;=\; 180.37$$

for the ED-interval (1800 is the number of seconds in an interval),

$$R_{QD} \;=\; 615 : \frac{1800}{328} \;=\; 112.07$$

for the QD-interval, and, finally,

$$R_{QED} \;=\; 1212 : \frac{1800}{304} \;=\; 204.69$$

for the QED-interval.

In the ED regime, we observe that the offered load R_{ED} (180.37) is considerably larger than the number of agents n (163.4). This implies that agents could not have coped with the offered load unless abandonment took place. The formal definition of the ED regime is in terms of the following relationship between n and R_{ED}:

$$n \;=\; R_{ED} \cdot (1 - \gamma)\,, \tag{4.2}$$

where the constant $\gamma > 0$ is interpreted as a *service grade*: larger γ will imply larger wait and abandonment. In our example,

$$\gamma \;=\; 1 - \frac{n}{R_{ED}} \;=\; 1 - \frac{163.4}{180.37} \;=\; 0.094\,.$$

Note that γ is equal to %Abn. This is not a coincidence, and an explanatory asymptotic statement will be proved in Theorem 17.1.

For the QD-regime, we have $R_{QD} = 112.07$ and $n = 135$. The definition of this regime is

$$n \;=\; R_{QD} \cdot (1 + \gamma)\,, \qquad \gamma > 0\,. \tag{4.3}$$

In our case, the service grade γ is

$$\gamma \;=\; \frac{n}{R_{QD}} - 1 \;=\; \frac{135}{112.07} - 1 \;=\; 0.205\,.$$

Now we proceed to, what we believe is, the most important operational regime: QED at 14:30. In this regime, the difference between n (206.1) and R_{QED} (204.69) is relatively

small and cannot be measured in units of R, as in (4.2) and (4.3). Furthermore, this difference can be either positive or negative. The definition we use is

$$n \;=\; R_{QED} + \beta\sqrt{R_{QED}}\,, \qquad -\infty < \beta < \infty\,, \tag{4.4}$$

where, in our example, the service grade

$$\beta \;=\; \frac{n - R_{QED}}{\sqrt{R_{QED}}} = \frac{206.1 - 204.69}{\sqrt{204.69}} \;=\; 0.10\,.$$

The *square-root staffing rule* (4.4) was already described by Erlang [23], as early as in 1924. He reports that it had 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 [34]. In that paper, the authors establish an important relation: as R (or λ) increase indefinitely, sustaining the QED operational regime (4.4) with fixed $\beta > 0$ is equivalent to holding the probability of wait at a fixed level α, $0 < \alpha < 1$. A one-to-one relation between α and β exists; see Subsection 7.2 for precise details. (Note that for Erlang-C [34], $\beta > 0$ must prevail.)

Garnett, Mandelbaum and Reiman [29] studied the QED regime for the Erlang-A model with exponential abandonment, establishing results that are analogous to [34]. In this case, the service grade β can go negative, as in (4.4).

In Section 2 we established the need to study models with general patience of customers. In concert with this, Part IV will be dedicated to the operational regimes (4.2)-(4.4) in the M/M/n+G queue framework. Our goal is, first, to develop asymptotic formulae for performance measures, given the arrival rate $\lambda \to \infty$. Next, we compare the derived approximations, check them against exact M/M/n+G formulae and, finally, proceed with their validation, based on real data.

Remark 4.1 As already mentioned for the QED example, the number of agents ($n = 206.1$) and the offered load ($R_{QED} = 204.69$) are very close. In the absence of abandonment (Erlang-C model), that would lead to low performance level. Indeed, given the characteristics of this time interval, Erlang-C predicts ASA that is equal to 208 seconds and 99.4% agents' occupancy. Therefore, we observe that a small fraction of abandonment (2.7%) drastically improves the service level (23 seconds vs. 208 previously).

Remark 4.2 Formulae (4.2)-(4.4) do not take into account that, in most applications, the number of servers n must be an integer. Subsequently, two modifications of these formulae will be used. Taking (4.4) as an example, the staffing regime could be either:

$$n \;=\; [R+\beta\sqrt{R}]\,,$$

where $[\ldots]$ denotes the nearest integer, or

$$n \;=\; R+\beta\sqrt{R}+\epsilon(R)\,,$$

where $\epsilon(R)$ is an "asymptotically small" term (e.g. $o(\sqrt{R})$), that depends on a specific problem.

4.3 Call center data

Most modern call centers are equipped with an ACD: this is the switch that routes calls to agents, while tracing and capturing the history of each call as it flows through the call center. ACD data include each call's arrival-time, waiting-time in the tele-queue, service duration etc.

ACD data is typically used through *aggregated* reports (recall Table 2). These consist of counts and averages over 15/30/60 minutes periods at the lowest level, and daily/weekly/yearly periods at higher levels.

Recently the need to proceed beyond aggregated data, exploring call-by-call data, has been more and more appreciated. The first comprehensive research of this kind was performed by Brown et al. [13], using data from a small Israeli bank. Currently, other data sets are analyzed by various researchers. Section 19, that is based on data from a large US bank, is a part of efforts in this direction.

5 Summary of results and structure of the thesis

5.1 Theoretical background and literature review

Since the central model in our research is the M/M/n+G queue, a queueing-theory background deserves and receives a major attention in part II.

Section 6 summarizes relevant exact results on queueing models with multi-servers and impatient customers. In addition to a general overview (Subsection 6.2), we elaborate on

four issues. Subsection 6.1 includes a comprehensive discussion on the very important Erlang-A model. Subsections 6.3 and 6.4 are dedicated to the two main sources for exact M/M/n+G formulae: Baccelli and Hebuterne [3] and Brandt and Brandt [11, 12], respectively. Finally, Subsection 6.5 contains a list of formulae, many of which are new, for M/M/n+G performance measures.

Section 7 is dedicated to asymptotic results, paying special attention to steady-state research in the three operational regimes, ED, QD and QED, which are then studied in Part IV.

Section 8 provides background for estimation of the parameters of M/M/n+G. The issue of inference of the patience distribution is presented in detail.

The last two sections of the Background include some support material for Part IV, on asymptotic results. Specifically, Section 9 contains formulations and proofs of several theorems on asymptotic behavior of integrals (the Laplace method [20]). Finally, Section 10 includes relevant information on the hazard rate of the standard normal distribution

5.2 Impact of customers' patience on delay and abandonment

The issues that arose in Subsection 4.1 will be explored within the framework of the M/M/n+G queueing system: under fixed exponential service rate μ and number of agents n (*internal* parameters of a call center), we explore system performance for a variety of patience distributions over different arrival rates.

We start with some M/M/n+G theory (Subsection 11.1). A non-obvious stochastic order relation for patience-time distributions G_1 and G_2 is presented (Lemma 11.1) that implies order relations between some performance measures of the corresponding M/M/n+G_1 and M/M/n+G_2 (probability to abandon and probability of wait). Then we verify (Theorem 11.1) that, for a fixed average patience $\bar{\tau}$, the deterministic patience (all customers are willing to wait *exactly* $\bar{\tau}$) maximizes some performance measures of M/M/n+G (average wait, average queue, probability of wait) while it minimizes the probability to abandon.

We continue with some *light-traffic* results (Subsection 11.2). Under the assumption that the arrival rate converges to zero, we compute the asymptotic ratio between the probability to abandon and average wait. In addition, we derive natural expressions for

the limits of the probability to abandon (11.8) and average wait (11.9), both conditional on positive wait.

Then we proceed with some theory-driven experiments (Section 12). Fixing the number of agents and the service rate, we vary the arrival rate of M/M/n+G from the very low to very high loads, while calculating steady-state performance characteristics for different patience distributions. In some sense, we are filling the gap between the light-traffic relations of Subsection 11.2 and heavy-traffic operating regimes, analyzed in Sections 15 and 17. In particular, we check for patience patterns that imply the relations in Figures 7 and 8. We verify that some non-exponential distributions (uniform, hyperexponential) give rise to a linear P{Ab} / E[W] relation (Figures 18-21) for loads that are not especially high (roughly, less than a third of the customers abandon). Then we consider the empirical patience-distribution that corresponds to the second plot of Figure 5 and observe a relatively linear pattern, thus supporting Figure 7. In addition, simulation is used in order to check influence of lognormal service times on system performance (vs. exponential services). In our experiments, we observe negligible differences for the probability to abandon and a very small one for the average wait (Figure 22).

On the other hand, some distributions (e.g. deterministic) imply strictly non-linear patterns for the above relation (Figure 23). We also connect the *abandonment rates*, introduced in Brandt and Brandt [12] with linearity or non-linearity of the P{Ab} / E[W] relation (Figure 24). Finally, some theoretical results from Section 11 and Brandt and Brandt [12] are validated via numerical experiments.

Summarizing, the structure of Part III is as follows. Section 11 provides formulations of theoretical results. In Section 12 we describe the theory-driven experiments mentioned above and Section 13 has our conclusions. Finally, proofs of the theoretical results are presented in Section 14.

5.3 Asymptotic operational regimes for the M/M/n+G queue

In Part IV we present a systematic treatment of the three operational regimes, described in Subsection 4.2. First, Sections 15, 16 and 17 are dedicated to the QED, Quality-Driven (QD) and Efficiency-Driven (ED) regimes, respectively.

Then Section 18 explores the Economies-Of-Scale (EOS) problem for the three regimes.

Specifically, assuming that the arrival rate increases by a factor $m > 1$, we apply the corresponding operational regime and check how the most important performance measures change in these circumstances.

In Section 19 our models are applied to call center data of a large bank in the USA. We start with a general description of the database, and explain how M/M/n+G primitives and main performance measures are calculated or estimated. Then we check if the linear relation P{Ab}/E[W] prevails. Finally, QED approximations are fitted to the data. In some of our experiments, we observe a good fit to the theoretical models, but others pose interesting challenges for future research.

Presentation and methodology. Sections 15-17 are each divided to three subsections. Those subsections cover, respectively, formulations of results with comments, numerical experiments and proofs.

The framework of our numerical experiments is as follows. For each case, we choose several test distributions. Overall, uniform, hyperexponential, Erlang and delayed exponential distributions are studied. Then we consider several values of the service grades β and γ from (4.4), (4.2) and (4.3). For each service grade, we vary the offered load per agent from small (20) to large (1000 or 2000) values.

In the QED regime we simply compare the exact values of different performance measures and their approximations. In the two other regimes, we also add comparisons with the QED approximations.

Finally, several words about our proofs. Most of the proofs combine two elements. First, we use the framework for exact M/M/n+G calculations, inspired by Baccelli and Hebuterne [3] and developed in Subsection 6.5. We show there that all the essential performance measures can be calculated via several building blocks defined in formulae (6.71)-(6.77). These building blocks have an integral form. Hence, for asymptotic analysis, we need a technique for asymptotic calculations of integrals. Here the *Laplace method* is helpful, and the necessary background on it is developed in Section 9.

We now briefly summarize our main results for the three operational regimes.

5.3.1 QED regime: summary of results

Recall that the QED operational regime is characterized by (4.4). It turns out that different patience distributions give rise to different asymptotic behavior of performance measures. Therefore, several special cases are considered in Theorems 15.1-15.7.

Main case: positive density at the origin. Assume the density of patience exists at the origin and denote its value by g_0. In most applications that we have encountered, a non-negligible abandonment rate during the first seconds of wait was observed. (See Section 19, for example.) Hence, there is a practical motivation to treat, as the main case, patience distributions with a positive density at the origin: $g_0 > 0$. In addition, there are significant theoretical reasons for this emphasis. It turns out that, in this case, performance measures behave similarly to Erlang-A, as analyzed in Garnett et al. [29]: the probability of wait converges to a constant, and the probability to abandon and average wait decrease at rate $1/\sqrt{n}$. If wait is measured in units of average-service-time, convergence rates depend only on the ratio g_0/μ and the service grade β. In fact, in order to get the right asymptotic expressions, when $g_0 > 0$, one simply replaces the exponential abandonment rate θ in the formulae of [29] by g_0, the patience density at the origin.

In addition, we establish an asymptotic linear relation between the probability to abandon and average wait. The exact relation

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[W]\,,$$

that holds for models with exponential patience, is replaced in Theorem 15.1 by

$$\mathrm{P}\{\mathrm{Ab}\} \;\approx\; g_0 \cdot \mathrm{E}[W]\,.$$

Although the last relation is approximate, our numerical experiments (Subsection 15.2) show that it provides an excellent approximation for a wide range of model parameters.

Since the case $g_0 > 0$ is the most important, we compute for it more performance measures than in the other special cases: for example, asymptotics for $\mathrm{E}[W|\mathrm{Ab}]$, $\mathrm{E}[W|W > t]$ etc. are derived only for this case; see Theorem 15.1.

Density vanishing near the origin. We would like to cover models where customers are going through several stages of (im)patience before reneging. (See, for example,

Issaev [38] or Baccelli and Hebuterne [3] who fit an Erlang distribution with 3 phases to patience in real data.) In such models, we cannot expect significant abandonment near the origin, which suggests patience distributions with density vanishing near the origin. These distributions are analyzed in Theorem 15.2. Specifically, assume that the k-th derivative of the density is positive and the first $(k-1)$ derivatives are zero. Then, in contrast to Theorem 15.1, positive, negative or zero values of the service grade β give rise to different performance regimes.

- If $\beta > 0$, the wait characteristics behave similar to the Erlang-C queue, described in Halfin and Whitt [34]. The probability to abandon decreases at $n^{-(k+1)/2}$ rate, i.e. faster than in the main case. Both statements above are connected: since abandonment is negligible, the system behaves like Erlang-C

- If $\beta < 0$, almost all customers are delayed and the average wait decreases to zero slowly (at rate $n^{-1/(2k+2)}$). The probability to abandon is asymptotically $-\beta/\sqrt{n}$, which is the minimal abandonment that is required to avoid queue explosion.

- The case $\beta = 0$ implies some intermediate behavior (e.g. the average wait decreases at rate $n^{-1/(k+2)}$).

An important special case of distributions, covered by Theorem 15.2, is *phase-type* (see Issaev [38] and references therein for their importance in Queueing Theory). Theorem 15.3 and formula (15.47) below show how to calculate the first non-zero derivative at the origin for these distributions.

Delayed distributions. Assume that, up to a fixed time $c > 0$, customers do not abandon. For example, customers could be listening to an announcement. Such situations inspire us to consider delayed distributions of patience, which can be represented by $c+\tau$, where τ represents (im)patience as before. (Recall the first plot of Figure 8.) The case of *deterministic patience* is important as well. As examples, one can consider overflowing[3], or Internet applications, where the waiting of jobs in queue is usually bounded.

Theorem 15.4 and 15.5 cover the two cases: delayed distributions and deterministic. Overall, our main conclusions are similar to Theorem 15.2; positive, negative and zero

[3]Customers that do not get service within a deterministic target time are sent to another call center or to the VRU.

values of β should be treated separately again. However, if $\beta \leq 0$, the average wait does not converge to zero. For negative service grades, it converges to the delay constant c, and if $\beta = 0$, to some number within the interval $(0, c)$, which we identify in Theorem 15.4.

Balking. Let customers that do not get service immediately balk with probability P{Blk}. From a practical point of view, this means that some customers do not agree to wait at all. (The reader surely recalls such a situation from personal experience.)

In this case (Theorem 15.6), the QED operational regime implies performance characteristics that are sometimes reminiscent of the Erlang-B analysis in Jagerman [44]. The probability of wait decreases at rate $1/\sqrt{n}$ and the average wait of delayed customers decreases at rate $1/n$. Hence, the unconditional average wait changes at rate $n^{-3/2}$.

If a customer is not served immediately, the probability to abandon converges to P{Blk}. (So abandonments after positive wait are rare.) Finally, the most surprising result arises for the unconditional probability to abandon. Asymptotically, it decreases at rate $1/\sqrt{n}$ and is equal to the blocking probability in Erlang-B, derived in [44].

Scaled balking In our last special case, we assume that the balking probability is scaled by $p_b/\sqrt{n}$, where n is the number of agents. It is designed to study queues with large number of servers and small, but non-negligible, balking probability. The findings, summarized in Theorem 15.7, are similar to the main case (Theorem 15.1).

Numerical experiments. The quality of the approximations from Theorems 15.1, 15.2 and 15.4 is checked in Subsection 15.2. Four patience distributions, seven values of the service grade and moderate-to-large values of λ and n were used in our experiments. If the offered load is larger than 100, we observe a good-to-excellent fit between approximations and exact values for absolute majority of special cases considered. A good fit for smaller values of the offered load (we started with 10) is also common.

5.3.2 QD regime: summary of results

The quality-driven operational regime was characterized by (4.3). Theorem 16.1 presents the asymptotic behavior of several important performance measures in this regime. It

covers patience distributions with positive density at the origin. The conclusions are as follows:

- The probability of wait decreases exponentially in n.
- The probability to abandon and the average wait of delayed customers decrease at rate $1/n$. From the previous statement, the unconditional performance measures decrease exponentially in n.
- The linear relation
$$\mathrm{P}\{\mathrm{Ab}\} \;\approx\; g_0 \cdot \mathrm{E}[W]\,, \tag{5.5}$$
that was established for the QED regime, prevails in the QD regime as well.
- The wait distribution of delayed customers is approximately exponential.

Numerical experiments. We consider two patience distributions: uniform and hyperexponential, comparing QD approximations with exact values and QED. It turns out that, in most cases, QED formulae are preferable over QD. However, if the probability of wait is very small (less than 0.1 or 0.05, for example), QD approximations should be used. Finally, we established again an excellent linear fit between $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$, as in (5.5).

5.3.3 ED regime: summary of results

Theorem 17.1 explores the ED operational regime, defined by (4.2). Here, in contrast to the other two regimes, patience behavior near the origin does not determine the values of asymptotic performance measures. The main assumptions are that the equation

$$G(x) \;=\; \gamma$$

has a unique solution x^* and that the patience density at x^* is positive. Then the probability to get service immediately decreases exponentially in n. The probability to abandon, average wait and average offered wait converge to constants (fluid limits); see Theorem 17.1.

Numerical experiments. The ED approximations for main performance measures are sometimes preferable over QED formulae. (This applies if $\rho \geq 1.2$ or for smaller ρ and very large values of λ.) The ED approximation for the probability of wait is less accurate than QED overall, but it does work well for large ρ.

Part II
Theoretical background and literature review

6 Queueing theory: relevant exact results

6.1 Review of the Erlang-A queue

In Section 2 of the Introduction we compared three queueing models: Erlang-C, Erlang-A and M/M/n+G. Roughly, Erlang-C is the "yesterday" of call-center practice, Erlang-A is "today" (most well-run modern call centers use this model or variations of Erlang-C that takes abandonment into account) and M/M/n+G is an example of a model that, we hope, will be in use "tomorrow".

Therefore, the Erlang-A model is very important both for M/M/n+G research (as the most well-known special case of this model) and for applications. Erlang-A results are scattered in Palm [56], Riordan [60], Garnett et al. [29], Whitt [74] and other books and papers. But we are unfamiliar with any published source that provides a relatively complete and application-oriented coverage of this model. Therefore, we believe that it is appropriate to provide a detailed discussion on Erlang-A in our Background. The treatment is based on the Teaching Note [52].

6.1.1 Birth-and-death process representation; Steady-state

Recall that the Erlang-A model is characterized by four parameters:

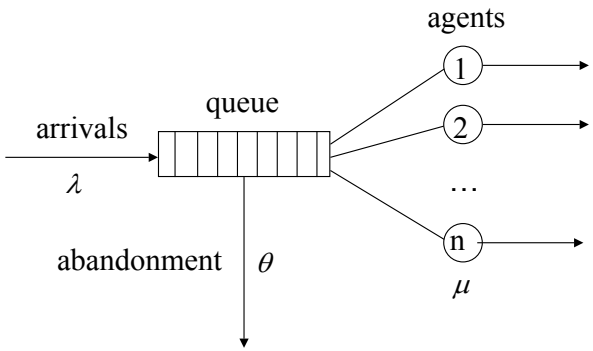
- λ – Poisson arrival rate ($\lambda > 0$);
- μ – individual service rate ($\mu > 0$);
- n – number of agents ($n = 1, 2, \ldots$);
- θ – individual abandonment rate ($\theta > 0$).

Figure 9 provides a representation of the traffic flows in Erlang-A, and a comparison with Figure 3 reveals its severe limitations. (Nevertheless, and as we hope to demonstrate, Erlang-A still turns out very useful and insightful, both theoretically and practically.)

More formally, in the Erlang-A model customers arrive to the queueing system according to a Poisson(λ) process. Customers are equipped with *patience times* τ that are exp(θ), i.i.d. across customers. And service times are i.i.d. exp(μ). Finally, the processes of arrivals, patience and service are mutually independent.

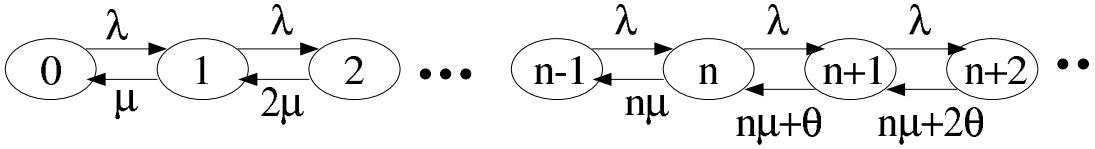
For a given customer, τ is the time that the customer is *willing* to wait for service - a wait that reaches τ results in an abandonment. Let V denote the *offered waiting time* - the time a customer, equipped with infinite patience, *must* wait in order to get service. The actual waiting/queueing time then equals $W = \min\{V, \tau\}$.

Figure 9: **Schematic representation of the Erlang-A model**



Denote by $L(t)$ the number-in-system at time t. Then $L = \{L(t), t \geq 0\}$ is a Markov birth-and-death process, with the following transition-rate diagram:

Figure 10: **Transition-rate diagram of the Erlang-A model**



Remark. Let d_j stand for the death-rate in state j, $0 \leq j < \infty$. Then

$$j \cdot \min\{\mu, \theta\} \leq d_j \leq j \cdot \max\{\mu, \theta\}\,. \tag{6.1}$$

The bounds on the left-hand and right-hand sides of (6.1) correspond to death-rates of an M/M/∞ queue with service rates $\min(\mu,\theta)$ and $\max(\mu,\theta)$, respectively. In some sense, which can be made precise via *stochastic orders* between distributions, these two M/M/∞ queues provide lower and upper (stochastic) bounds for the Erlang-A system. (The lower bound will be used later to prove that Erlang-A *always* reaches steady-state.) In the special case of equal service and abandonment rates ($\mu = \theta$), the steady-state distributions of Erlang-A and M/M/∞ coincide. ■

As customary, define the limit distribution of L by:

$$\pi_j \;=\; \lim_{t\to\infty} \mathrm{P}\{L(t)=j\}, \qquad j \geq 0\,.$$

When existing, the limit distribution is also a stationary distribution, which is calculated via the following version of the steady-state equations:

$$\left\{ \begin{array}{rcll} \lambda\pi_j &=& (j+1)\cdot\mu\pi_{j+1}, & 0\leq j\leq n-1 \\ \lambda\pi_j &=& (n\mu+(j+1-n)\theta)\cdot\pi_{j+1}, & j\geq n. \end{array} \right. \tag{6.2}$$

It is now straightforward to derive the "recipe" solution:

$$\pi_j \;=\; \left\{ \begin{array}{ll} \dfrac{(\lambda/\mu)^j}{j!}\pi_0\,, & 0\leq j\leq n \\ \displaystyle\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\pi_0\,, & j\geq n+1, \end{array} \right. \tag{6.3}$$

where

$$\pi_0 \;=\; \left[\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty}\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\right]^{-1}. \tag{6.4}$$

Our solution makes sense - equivalently the Markov process L is ergodic - if the infinite sum in (6.4) converges. This is a consequence of the lower bound in (6.1):

$$\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty}\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!} \;\leq\; \sum_{j=0}^{\infty}\frac{(\lambda/\min(\mu,\theta))^j}{j!} \;=\; e^{-\lambda/\min(\mu,\theta)}\,.$$

Formulae (6.3) and (6.4) include infinite sums, which can cause numerical problems. To overcome these, Palm [56] represented the Erlang-A steady-state distribution, and some of its important performance measures, in terms of special functions. To this end, define the *Gamma function*

$$\Gamma(x) \;\triangleq\; \int_0^\infty t^{x-1}e^{-t}dt\,, \qquad x>0, \tag{6.5}$$

and the *incomplete Gamma function*

$$\gamma(x,y) \triangleq \int_0^y t^{x-1}e^{-t}dt\,, \qquad x>0,\ y\geq 0. \tag{6.6}$$

(Both functions can be calculated in Matlab.) Let

$$A(x,y) \triangleq \frac{xe^y}{y^x}\cdot\gamma(x,y) = 1+\sum_{j=1}^{\infty}\frac{y^j}{\prod_{k=1}^{j}(x+k)}\,, \qquad x>0,\ y\geq 0. \tag{6.7}$$

(The second equality is taken from Palm [56].)

In addition, let $E_{1,n}$ denote the *blocking probability* in the M/M/n/n (Erlang-B) system, and recall the classic Erlang-B formula:

$$E_{1,n} = \frac{\frac{(\lambda/\mu)^n}{n!}}{\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!}}\,. \tag{6.8}$$

A simple way for calculating $E_{1,n}$ is the recursion

$$E_{1,0} = 0; \qquad E_{1,n} = \frac{\rho E_{1,n-1}}{1+\rho E_{1,n-1}}\,, \qquad n\geq 1,$$

in which ρ is the offered load per agent, namely

$$\rho \triangleq \frac{\lambda}{n\mu}\,.$$

In Subsection 6.1.5, we deduce from (6.7) the following solution for the limiting/stationary distribution:

$$\pi_j = \begin{cases} \pi_n\cdot\dfrac{n!}{j!\cdot\left(\frac{\lambda}{\mu}\right)^{n-j}}\,, & 0\leq j\leq n\,, \\ \pi_n\cdot\dfrac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta}+k\right)}\,, & j\geq n+1\,, \end{cases} \tag{6.9}$$

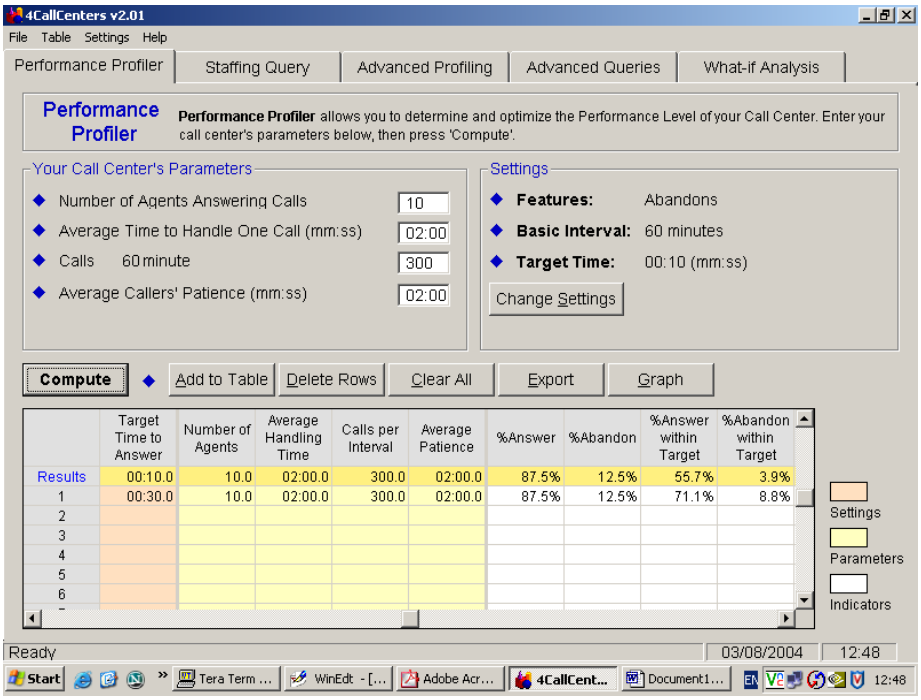
where

$$\pi_n = \frac{E_{1,n}}{1+\left[A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)-1\right]\cdot E_{1,n}}\,. \tag{6.10}$$

6.1.2 Operational measures of performance

Calculations: the 4CallCenters software

Figure 11: **4Callcenters. Example of output.**



Black-box Erlang-A calculations, as well as many other useful features, are provided by the free-to-use software 4CallCenters [22]. The calculation methods are described in Appendix B of [29]. They were developed in the Technion's M.Sc. thesis of the first author, Ofer Garnett.

These calculations are in fact for measures of the form $\mathrm{E}[f(V,\tau)]$, for various functions f (Table 3 in [29]). For example,

$$\mathrm{E}[W] = \mathrm{E}\left[\min\{V,\tau\}\right] \ , \qquad \mathrm{P}\{\text{Abandon}\} = \mathrm{E}\left[1_{\{\tau<V\}}\right] \ .$$

Figure 11 displays a 4CallCenters output and demonstrates how to calculate the four-dimensional service measure, introduced in Section 3.2.

The values of four Erlang-A parameters are displayed in the upper part of the screen. They are here as follows: $n = 10$, $1/\mu = 2$ min, $\lambda = 300$ per hour, and $1/\theta = 2$ min. Let

$T = 30$ seconds and $\epsilon = 10$ seconds. Then one should perform computations twice: with *Target Time* 10 and 30 seconds, respectively. We get:

- $P\{W \le T; \mathrm{Sr}\}$ - fraction of well-served is equal to 71.1%;
- $P\{W > T; \mathrm{Sr}\}$ - fraction of served, with a potential for improvement, is 16.4% (87.5% - 71.1%);
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served is 8.6% (12.5% - 3.9%);
- $P\{W \le \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined is 3.9%.

Note that the 4CallCenters output includes many more performance measures than those displayed in Figure 11: one could scroll the screen to display values of agents' occupancy, average waiting time, average queue length etc.

In Subsection 6.1.4 we describe several examples of the advanced capabilities of 4CallCenters.

A general approach for computing operational performance measures. Expressions for several performance measures of Erlang-A are derived in Riordan [60]. However, we recommend the use of more general M/M/n+G formulae, as the main alternative to the 4CallCenters software.

The Erlang-A queue is a special case of the M/M/n+G queue, in which patience times are generally distributed. A comprehensive list of M/M/n+G formulae will be provided in Subsection 6.5. We also explain there how to adapt those formulae to Erlang-A, in which patience is exponentially distributed:

$$G(x) = 1 - e^{-\theta x}, \qquad \theta > 0.$$

In addition to this general approach, we perform now several representative insightful calculations of key performance measures, based on conditioning and the incomplete gamma function introduced above. (See Subsection 6.1.5 for proofs.)

P$\{W > 0\}$. We start with the *delay probability* P$\{W > 0\}$, which represents the fraction of customers who are forced to actually wait for service. (The others are served

immediately upon calling.) Recall that this measure identifies operational regimes of performance.

Following Palm [56], we show in Subsection 6.1.5 that the representations (6.7) and (6.9) immediately imply

$$\mathrm{P}\{W>0\} \;=\; \sum_{j=n}^{\infty}\pi_j \;=\; \frac{A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)\cdot E_{1,n}}{1+\left(A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)-1\right)\cdot E_{1,n}}\,; \tag{6.11}$$

here, the first equality in (6.11) follows from PASTA.

P{Ab}. We proceed with calculating the probability to abandon, which represents the fraction abandoning. Define $\mathrm{P}_j\{\mathrm{Sr}\}$ to be the probability of ultimately getting served, for a customer that encounters j customers in queue upon arrival (equivalently, $n+j$ in the system). "Competition among exponentials" now implies that

$$\mathrm{P}_0\{\mathrm{Sr}\}=\frac{n\mu}{n\mu+\theta}\;.$$

Then,

$$\mathrm{P}_1\{\mathrm{Sr}\}=\frac{n\mu+\theta}{n\mu+2\theta}\cdot\mathrm{P}_0\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+2\theta}\,,$$

where we conditioned on the first event, after an arrival that encounters all servers busy and a single customer in queue; this event is either a service completion (with probability $\frac{n\mu+\theta}{n\mu+2\theta}$) or an abandonment. More generally, via induction:

$$\mathrm{P}_j\{\mathrm{Sr}\}=\frac{n\mu+j\theta}{n\mu+(j+1)\theta}\cdot\mathrm{P}_{j-1}\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+(j+1)\theta}\,,\qquad j\geq 1\,.$$

The probability to abandon service, given j customers in the queue upon arrival, finally equals

$$\mathrm{P}_j\{\mathrm{Ab}\} \;=\; 1-\mathrm{P}_j\{\mathrm{Sr}\} \;=\; \frac{(j+1)\theta}{n\mu+(j+1)\theta}\,,\qquad j\geq 0\,. \tag{6.12}$$

It follows that

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\sum_{j=n}^{\infty}\pi_j\mathrm{P}_{j-n}\{\mathrm{Ab}\}/\;\mathrm{P}\{W>0\} \;=\; \frac{1}{\rho A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)}+1-\frac{1}{\rho}\,. \tag{6.13}$$

The first equality in (6.13) is a consequence of PASTA, and the second is derived in Subsection 6.1.5. The fraction abandoning, $\mathrm{P}\{\mathrm{Ab}\}$, is simply the product $\mathrm{P}[\mathrm{Ab}|W>0]\times\mathrm{P}\{W>0\}$.

Theoretical relations among P{Ab}, E[W], E[Q]. A remarkable property of Erlang-A, which in fact generalizes to any model with patience that is $exp(\theta)$, is that

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[W]\,. \tag{6.14}$$

Proof: The proof is based on the balance equation

$$\theta \cdot \mathrm{E}[Q] \;=\; \lambda \cdot \mathrm{P}\{\mathrm{Ab}\}\,,$$

and on Little's formula

$$\mathrm{E}[Q] \;=\; \lambda \cdot \mathrm{E}[W]\,, \tag{6.15}$$

where Q is the steady-state queue length.

This is a steady-state equality between the rate that customers abandon the queue (left hand side) and the rate that abandoning customers (i.e. - customers who eventually abandon) enter the system. Applying Little's theorem (6.15) yields formula (6.14). ■

Observe that (6.14) is equivalent to

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\; \theta \cdot \mathrm{E}[W|W>0]\,. \tag{6.16}$$

Then, the average waiting time of delayed customers is computed via (6.13) and (6.16):

$$\mathrm{E}[W|W>0] \;=\; \frac{1}{\theta} \cdot \left[\frac{1}{\rho A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho}\right]\,. \tag{6.17}$$

The unconditional average wait E[W] equals the product of (6.11) with (6.17).

Sensitivity of performance measures to changes in model parameters. In the real world, one never knows the exact values of the four Erlang-A parameters. Therefore, it is essential to study the sensitivity of performance measures. In a recent paper [74], Whitt calculates *elasticities* in Erlang-A, which measure the percentage change of a performance measure caused by a small percentage change in a parameter. Both exact numerical algorithm and several types of approximations are used. It turns out that Erlang-A performance is quite sensitive to small changes in the arrival rate, service rate, or number of agents, but relatively insensitive to small changes in the abandonment rate.

6.1.3 Applications in call centers

Erlang-A performance measures: comparison against real data. We now validate the Erlang-A model against hourly data for an Israeli bank call center, which has been already used in Subsection 4.1. Three performance measures are considered: probability to abandon, average waiting time and probability of wait. Their values are calculated for the hourly intervals using exact Erlang-A formulae, and the results are then aggregated as in Figure 7. The resulting 86 points are compared against the line $y = x$: the closer the line is to the points, the better is the fit of Erlang-A to reality.

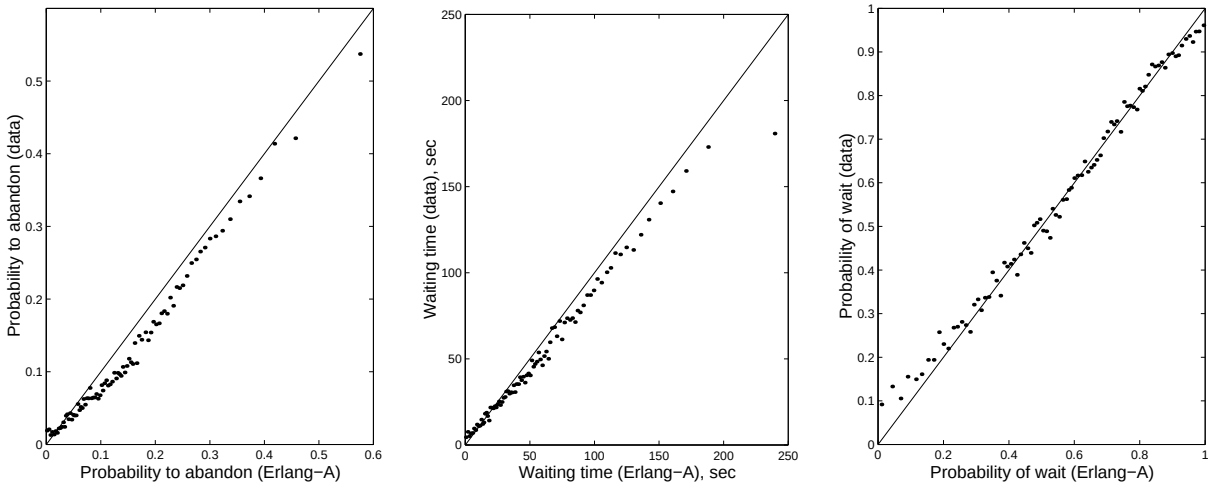
Computation of the Erlang-A parameters. The parameters λ and μ are calculated for every hourly interval. We also estimate each hour's average number of agents n. (See Section 19 for details.) Because the resulting n's need not be integral, we apply a continuous extrapolation of the Erlang-A formulae, obtained from relationships developed in Palm [56].

For θ, we use formula (6.14), valid for exponential patience, in order to compute 17 hourly estimates of $1/\theta = E(\tau)$ (for the 17 one-hour intervals 7am-8am, 8am-9am, ..., 11pm-12pm). The values for $\mathrm{E}(\tau)$ ranged from 5.1 min (8am-9am) to 8.6 min (11pm-12pm). We judged this to be better than estimating θ individually for each hour (which would be very unreliable) or, at the other extreme, using a single value for all intervals (which would ignore possible variations in customers' patience over the time of day; see [76]).

The results are displayed in Figure 12. The figure's two left-hand graphs exhibit a relatively small yet consistent overestimation with respect to empirical values, for moderately and highly loaded hours. The right-hand graph shows a very good fit almost everywhere, except for very lightly and very heavily loaded hours. The underestimation for small values of $\mathrm{P}\{W > 0\}$ can be probably attributed to violations of work conservation (idle agents do not always answer a call immediately). Summarizing, it seems that these Erlang-A estimates can be used as reasonable *upper bounds* for the main performance characteristics of our call center.

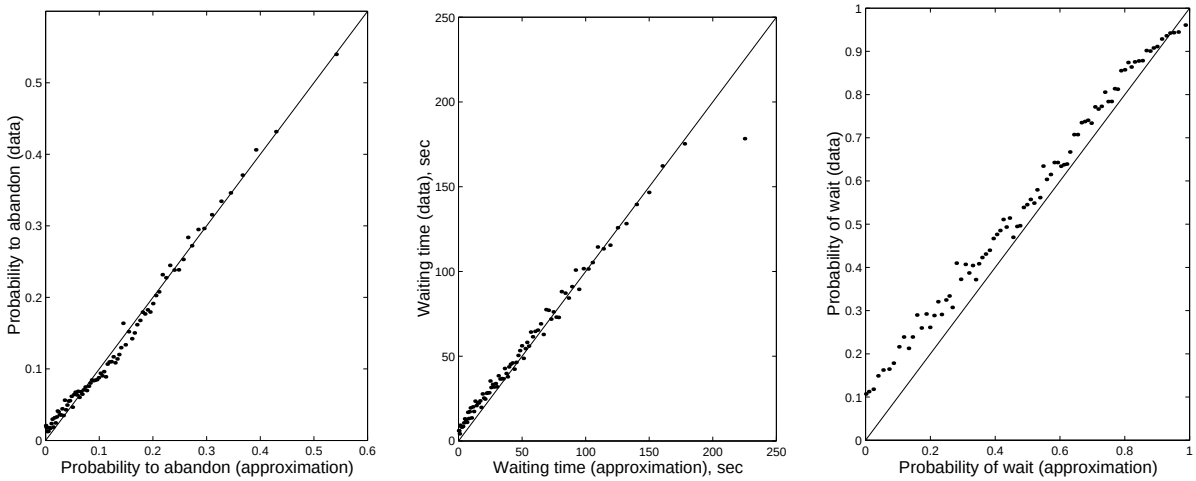
Erlang-A approximations: comparison against real data In Section 15 we shall discuss approximations of various performance measures for the Erlang-A (M/M/N+M) model. Such approximations require significantly less computational effort than exact

Figure 12: **Erlang-A formulas vs. data averages**



Erlang-A formulae, but they are at least as accurate; see Figure 13, based on the same data as Figure 12, demonstrates a good fit between data averages and the approximations.

Figure 13: **Erlang-A approximations vs. data averages**



The empirical fit of the simple Erlang-A model and its approximation turns out to be very (perhaps surprisingly) accurate. Thus, for the call center under consideration—and those like it – use of Erlang-A for capacity-planning purposes could and should improve

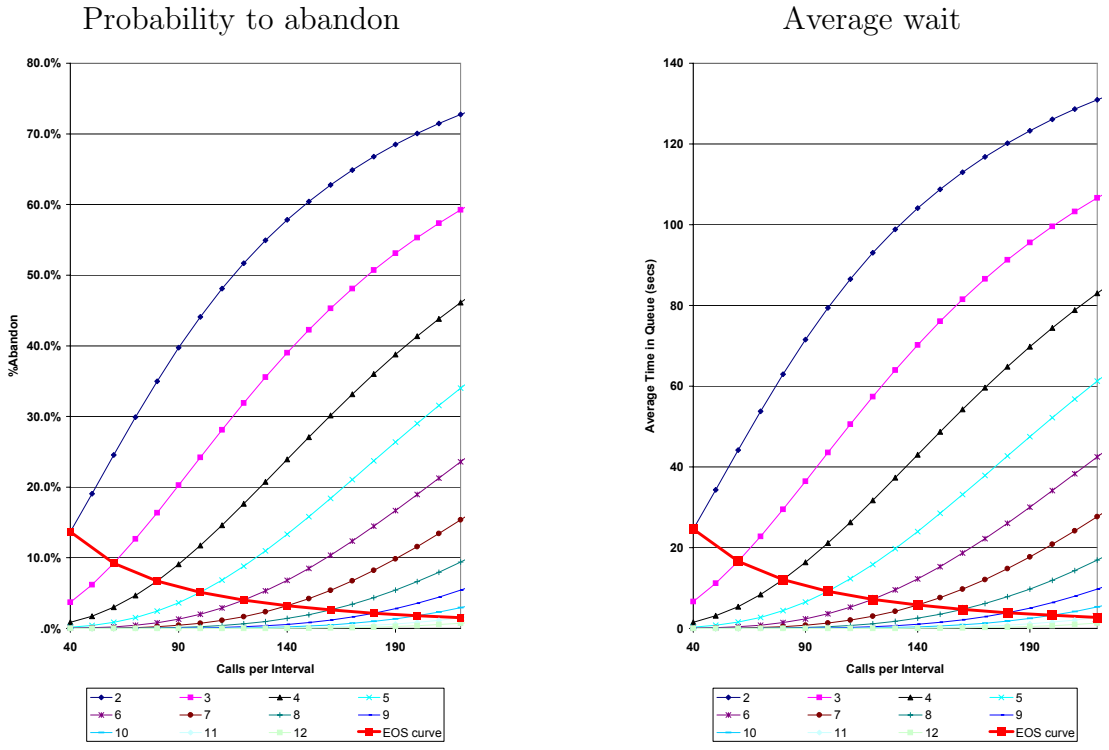
operational performance.

6.1.4 Advanced features of 4CallCenters

The 4CallCenters software [22] provides a valuable tool for implementing Erlang-A calculations. Its basic feature is "Performance Profiler" that enables calculation of all the useful performance measures, given the four Erlang-A parameters as input. In addition, 4CallCenters allows many advanced options: staffing queries, graphs, export and import of data etc.

Here we demonstrate, as an example, two advanced capabilities of 4CallCenters.

Figure 14: **4CallCenters. Advanced profiling.**



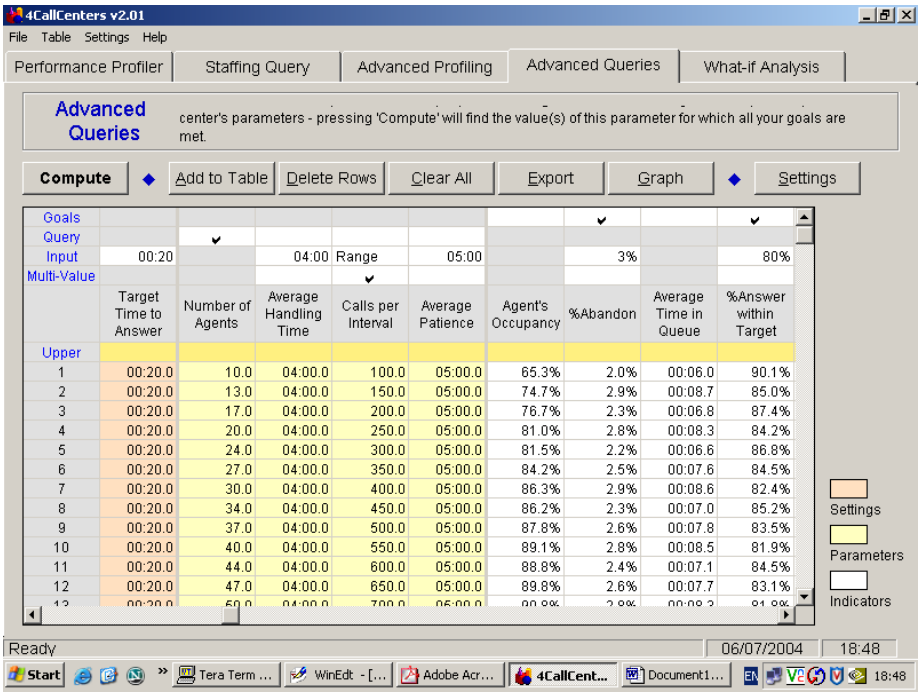
Example 1. Advanced profiling. One can vary any input parameters of the Erlang-A queue and display the model output (performance measures) either in a table or graphically. For example, let the average service time equal to 2 minutes, and average patience 3 minutes. Let the arrival rate vary from 40 to 230 calls per hour, in steps of 10, and the

number of agents from 2 to 12. Then one can immediately produce a table that contains values of different performance measures for all combinations of the two input parameters.

Figure 14 shows the dependence of the probability to abandon and average wait on the different number of agents. Note that the two plots look identical: the reason is relation (6.14). In addition, the red curves on both plots in Figure 14 illustrate Economies of Scale (EOS): while offered load per server remains constant along the curve $\left(\frac{\lambda}{n\mu} = \frac{2}{3}\right)$, performance significantly improves as the number of agents increases. (See Table 4 in Section 18.)

Example 2. Advanced staffing queries.

Figure 15: **4Callcenters. Advanced staffing queries.**



4CallCenters enables staffing queries with several performance goals. For example, assume that the average service time is equal to 4 minutes, and average patience is 5 minutes. We need to calculate appropriate staffing levels for arrival-rate values that vary from 100 to 1200, in steps of 50. Our performance goals are:

- Probability to abandon less than 3%,
- 80% of customers served within 20 seconds.

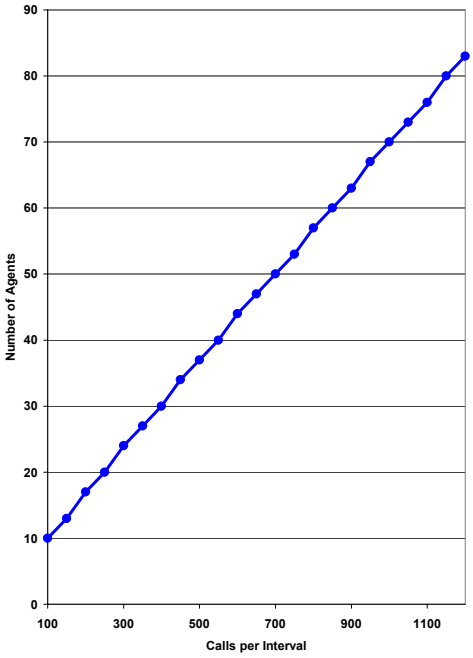
Figure 15 presents the screen output of 4CallCenters.

The first plot of Figure 16 displays the minimal staffing level that adheres to both goals. The EOS phenomenon is observed here as well: 10 agents are needed for 100 calls per hour but only 83 (rather than $10 \cdot 12 = 120$) for 1200 calls per hour. (Despite its look, the curve in the first plot is not a straight line.) The second plot displays the values of the two target performance measures. (This plot, unlike the first one, is not an immediate output of 4CallCenters.)

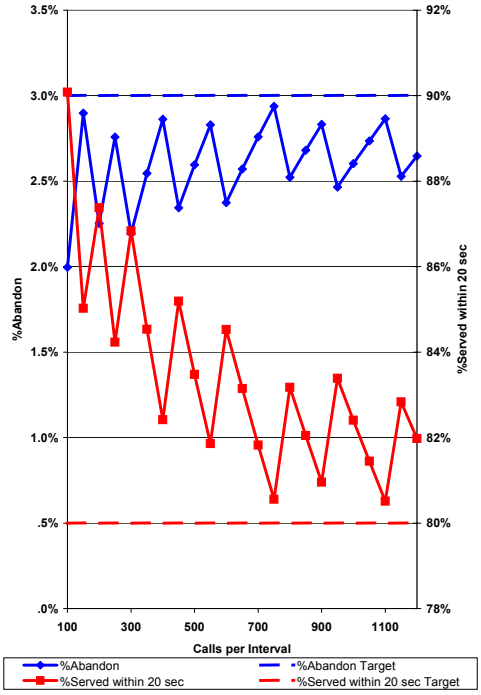
Remark 6.1 Since the number of agents must be integer, we observe performance "zigzags" in the right plot of Figure 16.

Figure 16: **4CallCenters. Staffing according to target performance values.**

Recommended staffing level



Target performance measures



6.1.5 Erlang-A: derivation of some performance measures

Steady-state distribution. Using formulae (6.4),(6.7) and definition (6.8), we have

$$\pi_0^{-1} = \sum_{j=0}^{n} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^n}{n!} \cdot \sum_{j=n+1}^{\infty} \prod_{k=n+1}^{j} \left(\frac{\lambda}{n\mu + (k-n)\theta} \right)$$

$$= \frac{(\lambda/\mu)^n}{n!} \cdot \left[\frac{1}{E_{1,n}} + \sum_{j=1}^{\infty} \frac{(\lambda/\theta)^j}{\prod_{k=1}^{j} \left(\frac{n\mu}{\theta} + k \right)} \right] = \frac{(\lambda/\mu)^n}{n!} \cdot \left[\frac{1}{E_{1,n}} + A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right].$$

Hence

$$\pi_0 = \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right] \cdot E_{1,n}} \cdot \frac{n!}{(\lambda/\mu)^n}. \tag{6.18}$$

For $1 \le j \le n$,

$$\pi_j \;=\; \pi_0 \cdot \frac{(\lambda/\mu)^j}{j!} \;=\; \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}} \cdot \frac{n!}{j! \cdot (\lambda/\mu)^{n-j}}\,. \tag{6.19}$$

Specifically,

$$\pi_n \;=\; \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}}\,. \tag{6.20}$$

Finally, for $j > n$

$$\pi_j \;=\; \pi_n \cdot \frac{\lambda^{j-n}}{\prod_{k=1}^{j-n}(n\mu + k\theta)} \;=\; \frac{E_{1,n}}{1 + \left(A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right) \cdot E_{1,n}} \cdot \frac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta} + k\right)}\,. \tag{6.21}$$

Probability of wait. From PASTA, (6.20) and (6.21), the delay probability is equal to

$$\begin{aligned} \mathrm{P}\{W > 0\} \;&=\; \sum_{j=n}^{\infty} \pi_j \;=\; \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}} \cdot \left[1 + \sum_{j=n+1}^{\infty} \frac{(\lambda/\theta)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta} + k\right)}\right] \\ &=\; \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}}\,. \end{aligned} \tag{6.22}$$

Probability to abandon. First, we need to perform some preliminary calculations. Differentiating (6.7), we get

$$\frac{\partial}{\partial y} A(x,y) \;=\; \frac{\partial}{\partial y}\left[\frac{xe^y}{y^x}\gamma(x,y)\right] \;=\; \frac{x}{y} + \left(1 - \frac{x}{y}\right) \cdot A(x,y)\,.$$

Then, for $x > 0$, $y > 0$,

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{(j+1)y^j}{\prod_{k=1}^{j+1}(x+k)} \;&=\; \frac{\partial}{\partial y}\left[\sum_{j=1}^{\infty} \frac{y^j}{\prod_{k=1}^{j}(x+k)}\right] \\ =\; \frac{\partial}{\partial y}[A(x,y) - 1] \;&=\; \frac{\partial}{\partial y} A(x,y) \;=\; \frac{x}{y} + \left(1 - \frac{x}{y}\right) \cdot A(x,y)\,. \end{aligned} \tag{6.23}$$

Using (6.22) and (6.12), the conditional probability to abandon is equal to

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;=\; \frac{\sum_{j=n}^{\infty} \pi_j \cdot \mathrm{P}_{j-n}\{\mathrm{Ab}\}}{\mathrm{P}\{W > 0\}}$$

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \sum_{j=n}^{\infty} \frac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta}+k\right)} \cdot \frac{\theta(j+1-n)}{n\mu+\theta(j+1-n)}$$

(where, by convention, $\prod_{k=1}^{0}\left(\frac{n\mu}{\theta}+k\right) \triangleq 1$)

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda}{\theta}\right)^{j} \cdot (j+1)}{\prod_{k=1}^{j+1}\left(\frac{n\mu}{\theta}+k\right)} = \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \frac{\partial}{\partial y}\left[A\left(\frac{n\mu}{\theta}, y\right)\right]_{y=\lambda/\theta}$$

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \left[\frac{n\mu}{\lambda} + \left(1-\frac{n\mu}{\lambda}\right) A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)\right] = \frac{1}{\rho A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho},$$

where the last line follows from (6.23).

6.2 General review of the M/M/n+G queue

A formal definition of the M/M/n+G queue requires a Poisson arrival rate λ, exponential service rate μ, number of agents n and a general patience distribution $G(\cdot)$. Let $\bar{G} = \{\bar{G}(x),\ x \geq 0\}$ denote the survival function of the patience time τ: $\bar{G}(x) = \mathrm{P}\{\tau > x\},\ x \geq 0$. We assume that an arriving customer encounters, in steady-state, an offered waiting time V (the time that a customer with infinite patience would have to wait). Then the actual queueing time of a steady-state customer equals $W = \min(V, \tau)$.

The seminal work on queueing systems with impatient customers is Palm [55, 56]. These articles have inspired the main directions of research on the topic: theoretical analysis of queueing systems, studying customers' impatience in the real-world and constructing mathematical models of impatience.

More specifically, Palm introduced the basic Erlang-A (M/M/n+M) queueing system with exponential patience times that was covered in Subsection 6.1.

Gnedenko and Kovalenko [30] analyzed the M/M/n+D queueing system (deterministic patience times). Jurkevic [44] applied their methods to the general M/M/n+G system. Independently, the M/M/n+G queue was analyzed by Baccelli and Hebuterne [3] and Haugen and Skogan [35]. Boxma and de Waal [10] developed several approximations for the probability to abandon in the M/M/n+G queue and tested them via simulation. Choi, Kim and Zhu [15] extended the analysis of deterministic patience time to MAP/M/n+D – a queueing system with a Markovian Arrival Process.

The derivation of M/M/n+G performance measures continued in Brandt and Brandt [11, 12]. They considered the more general M(k)/M(k)/n+G system where arrival and service rates are allowed to depend on the number k of calls in the system. (The service rate is assumed to remain constant for $k > n$.) However, some of the results in [11, 12] (for example, the distribution of total number-in-system) are new also for the M/M/n+G queue.

Another important branch of research is the estimation of the patience distribution in real tele-queue systems. Palm [55] introduced a mathematical model for irritation which postulated a Weibull distribution of patience times. Then he presented some real data that confirmed his hypothesis. Kort [48] also used the Weibull distribution to model patience while waiting for a dial tone. Baccelli and Hebuterne [3], using data from Roberts [61], fit it to an Erlang distribution with 3 phases. Brown et al. [13], in research on a bank call center, observed the patience times in the second plot of Figure 5. Finally, Daley and Servi [18] estimate Erlang-A parameters and performance characteristics (in particular, probability to abandon) given incomplete empirical data. They establish a useful relation between the Erlang-A queue and M/M/n with balking: it turns out that estimation of customers' loss rate is very similar for the two models.

Concerning models of customers' impatience, readers are referred to the papers of Mandelbaum, Shimkin and Zohar [50, 76, 62] where it is assumed that customers adapt their patience to the waiting patterns they expect to encounter. For further references and a more complete survey see Gans, Koole and Mandelbaum [27].

Below we elaborate on some theoretical results that are the most relevant for our research. We start with a detailed review of M/M/n+G results from Baccelli and Hebuterne [3] (Subsection 6.3) and Brandt and Brandt [11, 12] (Subsection 6.4). Later, in Section 7, we summarize relevant asymptotic results for Erlang-C, Erlang-A and other queueing systems.

6.3 Detailed description of the results of Baccelli and Hebuterne on M/M/n+G

Baccelli and Hebuterne [3] assume that arriving customers can calculate their offered wait V already at their arrival epochs. Thus, if the offered wait exceeds their patience time,

customers abandon immediately and do not join the queue. However, this model coincides with the classical M/M/n+G model as far as abandonment probability, offered wait and other performance measures that are not functions of queue-length are concerned.

The analysis in [3] is based on a Markov process $\{(N(t), \eta(t)),\ t \geq 0\}$, where $N(t)$ is the number of busy agents and $\eta(t)$ is the virtual offered waiting time (the offered wait of a virtual customer that arrives at time t). Then the steady-state characteristics are defined by:

$$\begin{cases} v(x) & \triangleq & \lim\limits_{t\to\infty}\lim\limits_{\epsilon\to 0}\dfrac{\mathrm{P}\{N(t)=n,\ x<\eta(t)\leq x+\epsilon\}}{\epsilon}, & x\geq 0 \\ \pi_j & \triangleq & \lim\limits_{t\to\infty}\mathrm{P}\{N(t)=j,\ \eta(t)=0\}, & 0\leq j\leq n-1 \end{cases} \tag{6.24}$$

Here $v(x)$ is the density of the virtual offered waiting time. The unique solution of the steady-state equations is given by

$$\pi_j \ = \ \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\,\pi_0\,, \quad 0\leq j\leq n-1 \tag{6.25}$$

$$v(x) \ = \ \lambda\,\pi_{n-1}\exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}, \tag{6.26}$$

where

$$\pi_0 \ = \ \left[1+\frac{\lambda}{\mu}+\cdots+\left(\frac{\lambda}{\mu}\right)^{n-1}\frac{1}{(n-1)!}(1+\lambda J)\right]^{-1}, \tag{6.27}$$

$$J \ \triangleq \ \int_0^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}dx\,. \tag{6.28}$$

Moreover, probability to abandon can be calculated by

$$\mathrm{P}\{\mathrm{Ab}\} \ = \ \left(1-\frac{n\mu}{\lambda}\right)\left(1-\sum_{j=0}^{n-1}\pi_j\right)+\pi_{n-1}\,. \tag{6.29}$$

6.3.1 Proofs of the results of Baccelli and Hebuterne

First we show that the steady-state equations can be presented in the following form:

$$\lambda\pi_j \ = \ \mu(j+1)\pi_{j+1}, \quad 0\leq j\leq n-2\,, \tag{6.30}$$

$$\lambda\pi_{n-1} \ = \ v(0)\,, \tag{6.31}$$

$$v(x) \ = \ \lambda\pi_{n-1}\exp\{-n\mu x\}+\lambda\int_0^x \bar{G}(u)v(u)\exp\{-n\mu(x-u)\}du\,. \tag{6.32}$$

Equations (6.30) follow from the Kolmogorov equations for the M/M/n queue:

$$\lambda\pi_0 \;=\; \mu\pi_1 \quad \text{and} \quad (\lambda+\mu j)\pi_j \;=\; \lambda\pi_{j-1}+\mu(j+1)\pi_{j+1}, \qquad 1\le j\le n-2$$

(In absence of a queue, M/M/n+G system behaves like M/M/n.)
In order to derive the Kolmogorov equation for the state $(n-1)$, note that

$$\begin{aligned}\mathrm{P}\{N(t)=n-1\} \;&=\; \mathrm{P}\{N(t-\epsilon)=n-1\}\cdot(1-(\lambda+\mu(n-1))\epsilon)\\ &+\; \mathrm{P}\{N(t-\epsilon)=n-2\}\cdot\lambda\epsilon+\mathrm{P}\{N(t-\epsilon)=n, 0<\eta(t)\le\epsilon\}+o(\epsilon)\,.\end{aligned}$$

The last equation and the second definition of (6.24) implies that in steady-state:

$$(\lambda+\mu(n-1))\pi_{n-1}=\lambda\pi_{n-2}+v(0)\,,$$

which, combined with $\lambda\pi_{n-2}=\mu(n-1)\pi_{n-1}$, implies (6.31).
In order to validate equation (6.32), note that for $x\ge 0$, $t>0$, $h>0$

$$\begin{aligned}\mathrm{P}\{\eta(t+h)>x\} \;&=\; \mathrm{P}\{\eta(t)>x+h\} \;+\; \mathrm{P}\{\eta(t+h)>x;\,\eta(t)=0\}\\ &+\; P\{\eta(t+h)>x;\,0<\eta(t)\le x+h\}\,.\end{aligned}$$

In steady-state,

$$\begin{aligned}\int_x^\infty v(y)dy \;&=\; \int_{x+h}^\infty v(y)dy \;+\; \lambda h\exp\{-n\mu x\}\pi_{n-1}\\ &+\; \int_0^x \lambda h\exp\{-n\mu(x-u)\}v(u)\bar{G}(u)du \;+\; o(h)\,.\end{aligned} \tag{6.33}$$

The second term of (6.33) corresponds to an arrival to the system with $(n-1)$ busy and one idle server. This arrival will increase the virtual offered wait by more than x with probability $\exp\{-n\mu x\}$. (If all servers are busy the time intervals between service terminations are distributed $\exp(n\mu)$.)

Finally, the third integral term describes an arrival to the system with the virtual offered wait u. In this case, the customer will get service (and affect the offered wait) with probability $\bar{G}(u)$.

Now, differentiating (6.33) with respect to h and taking $h\to 0$ we get (6.32).

If formula (6.32) prevails, the function

$$H(x)=\exp\{n\mu x\}\cdot v(x) \tag{6.34}$$

solves the integral equation

$$H(x) \;=\; \lambda\pi_{n-1} + \lambda \int_0^x \bar{G}(u)H(u)du\,. \tag{6.35}$$

On the other hand, the solution of (6.35) is equal to

$$H(x) \;=\; \lambda\pi_{n-1} \exp\left\{\lambda \int_0^x \bar{G}(u)du\right\}\,. \tag{6.36}$$

(Substitute formula (6.36) to (6.35) and compare derivatives and values at $x = 0$ of the right-hand and left-hand sides.)

Combining formulae (6.34) and (6.36), we get (6.26).

Formula (6.25) follows from equations (6.30). Equation (6.27) can be derived using the normalizing condition:

$$\sum_{j=0}^{n-1} \pi_j + \mathrm{P}\{V > 0\} \;=\; 1\,,$$

and

$$\mathrm{P}\{V > 0\} \;=\; 1 - \sum_{j=0}^{n-1} \pi_j \;=\; \int_0^\infty v(x)dx \;=\; \lambda\pi_{n-1}J\,, \tag{6.37}$$

where the last equality of (6.37) follows from (6.26) and (6.28).

In order to derive formula (6.29) for the probability to abandon, note that the agents' occupancy ρ can be alternatively computed using:

$$\rho \;=\; \frac{\lambda}{n\mu} \cdot (1 - \mathrm{P}\{\mathrm{Ab}\})$$

or

$$\rho \;=\; \sum_{j=0}^{n-1} \frac{j\,\pi_j}{n} + \left(1 - \sum_{j=0}^{n-1} \pi_j\right)\,. \tag{6.38}$$

Now it is easy to get (6.29) by substituting (6.38) into

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; 1 - \frac{n\mu}{\lambda} \cdot \rho\,.$$

6.4 Detailed description of the results of Brandt and Brandt on M/M/n+G

Brandt and Brandt [11, 12] consider the M(k)/M(k)/n+G queueing system, where arrival and service rates can depend on the number of customers k in the system. Unlike Baccelli and Hebuterne [3], they assume that customers abandon at the end of their patience

times. We provide here an M/M/n+G version of [11, 12], adapting their notation to our needs. Since the steady-state distribution of the number-in-system is calculated, the basic Markov process in [11, 12] is more complicated than in Baccelli and Hebuterne [3].

Model. Brandt and Brandt assume that patience time has a survival function $\bar{G}$ with a continuous density g. (We believe that many results from [11, 12] remain true also for a general patience distribution but we do not pursue this here.) Define
$N(t) \stackrel{\Delta}{=}$ number of customers in the system at time t.
$L(t) = (N(t) - n)_+ \stackrel{\Delta}{=}$ queue length.
$(X_1(t), \ldots, X_{L(t)}(t)) \stackrel{\Delta}{=}$ residual patience times of waiting customers, ordered according to their position in queue;
$(I_1(t), \ldots, I_{L(t)}(t)) \stackrel{\Delta}{=}$ original patience times of waiting customers, ordered according to their position in queue;
$\pi_k = P\{N(t) = k\} \stackrel{\Delta}{=}$ stationary distribution of the number-in-system.
The Markov process which is analyzed in [11, 12] is:

$$(N(t); X_1(t), \ldots, X_{L(t)}(t);\, U_1(t), \ldots, U_{L(t)}(t)). \tag{6.39}$$

Its stationary distribution is denoted by

$$P_k(x_1, \ldots, x_l;\, u_1, \ldots, u_l) \stackrel{\Delta}{=} \tag{6.40}$$

$$\stackrel{\Delta}{=} \lim_{t\to\infty} \mathrm{P}\{N(t) = n + l; X_1(t) \le x_1, \ldots, X_l(t) \le x_l; U_1(t) \le u_1, \ldots, U_l(t) \le u_l\},$$

where l is the queue length. Due to the FCFS (First Come First Served) service discipline, the support of distribution (6.40) is contained in

$$\Omega_l \stackrel{\Delta}{=} \{(x_1, \ldots, x_l; u_1, \ldots, u_l) \in R_{2l}^+ : u_1 - x_1 \ge \ldots \ge u_l - x_l \ge 0\}. \tag{6.41}$$

Define the stationary density (which exists, as proved in [11])

$$\pi_k(x_1, \ldots, x_l; u_1, \ldots, u_l) \stackrel{\Delta}{=} \frac{\partial^{2l}}{\partial x_1 \ldots \partial x_l \partial u_1 \ldots \partial u_l} P_k(x_1, \ldots, x_l; u_1, \ldots, u_l).$$

Summary of results. Now we cite results from [11, 12] that are relevant for our research. First, introduce the function

$$F(\xi) \stackrel{\Delta}{=} \int_0^{\xi/(n\mu)} \bar{G}(\eta) d\eta, \quad \xi > 0$$

and the constants

$$F_j \triangleq \frac{1}{j!}\int_0^\infty F(\xi)^j e^{-\xi} d\xi\,. \tag{6.42}$$

As common in the analysis of Markov chains, one must compute a normalization constant, say ω, which is here given by

$$\omega^{-1} = \sum_{j=0}^{n-1} \frac{n!\cdot\lambda^j\mu^{n-j}}{j!} + \sum_{j=0}^{\infty}\lambda^{n+j}F_j\,. \tag{6.43}$$

Then the steady-state distribution is

$$\pi_k \triangleq \lim_{t\to\infty}\mathrm{P}\{N(t)=k\} = \omega\frac{n!\cdot\lambda^k\mu^{n-k}}{k!}, \quad 0\le k\le n\,, \tag{6.44}$$

$$\pi_k = \omega\lambda^k F_{k-n}, \quad k>n\,, \tag{6.45}$$

$$\pi_k(x_1,\dots,x_l;u_1,\dots,u_l) = I\{(x_1,\dots,x_l;u_1,\dots,u_l)\in\Omega_l\}\cdot\omega\lambda^k\cdot\prod_{i=1}^{l} g(u_l)\cdot e^{-n\mu(u_1-x_1)} \tag{6.46}$$

According to Little's formula, the mean waiting time is

$$\mathrm{E}[W] = \frac{\sum_{k=n+1}^{\infty}(k-n)\pi_k}{\lambda}\,.$$

A formula for the waiting-time distribution was also derived in [11] and is omitted here.

The paper [12] contains, in particular, an expression for abandonment rates α_l given l customers in queue. Formally,

$$\alpha_l \triangleq \frac{1}{\pi_{n+l}}\sum_{i=1}^{l}\int_{R_+^{2l-1}}\pi_{n+l}(x_1,\dots,x_{i-1},0,x_{i+1},\dots,x_l;\,u_1,\dots,u_l)dx_1\dots dx_{i-1}dx_{i+1}\dots dx_l$$

$$du_1\dots du_l = \frac{F_{l-1}}{F_l} - n\mu\,, \tag{6.47}$$

where F_l are defined by (6.42).

The probability to abandon can be represented, alternatively to (6.29), by

$$\mathrm{P}\{\mathrm{Ab}\} = \frac{\sum_{l=n+1}^{\infty}\alpha_{l-n}\pi_l}{\lambda}\,. \tag{6.48}$$

The abandonment rates are asymptotically increasing according to the rate:

$$\lim_{l\to\infty}\frac{\alpha_l}{l} = \frac{1}{\mathrm{E}[\tau]} \tag{6.49}$$

If the m-th moment of the patience-time distribution is finite for $m>2$, then

$$\lim_{l\to\infty}(\alpha_l-\alpha_{l-1}) = \frac{1}{\mathrm{E}[\tau]} \tag{6.50}$$

Note that for exponential patience times, the last two equalities prevail exactly.

6.4.1 Derivation of the number-in-system distribution

The steady-state equations for the Markov process (6.39) are given by:

$$\lambda\pi_k = (k+1)\mu\pi_{k+1}, \quad 0 \le k \le n-1 \tag{6.51}$$

$$\lambda\pi_n = \int_{R_+} \pi_{n+1}(0,u)du + n\mu \int_{R_+^2} \pi_{n+1}(x,u)dxdu \tag{6.52}$$

$$\begin{aligned}
&\pi_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) \\
&\quad = \pi_k(x_1+h,\ldots,x_l+h;\, u_1,\ldots,u_l)\cdot(1-\lambda h - n\mu h) \\
&\quad + h\cdot\sum_{i=1}^{l+1}\int_{R_+} \pi_{k+1}(x_1,\ldots,x_{i-1},0,x_i,\ldots,x_l;\, u_1,\ldots,u_{i-1},u,u_i,\ldots,u_l)du \\
&\quad + hn\mu\cdot\int_{R_+^2} \pi_{k+1}(x,x_1,\ldots,x_l;\, u,u_1,\ldots,u_l)dxdu + o(h)
\end{aligned} \tag{6.53}$$

$$\begin{aligned}
&\pi_k(x_1,\ldots,x_{l-1},u_l;\, u_1,\ldots,u_{l-1},u_l) \\
&\quad = \lambda\pi_{k-1}(x_1,\ldots,x_{l-1};\, u_1,\ldots,u_{l-1})\cdot g(u_l)
\end{aligned} \tag{6.54}$$

Here equations (6.51) coincide with the M/M/n steady-state equations. The first integral term of (6.52) corresponds to abandonment and the second term to service termination. The three terms of (6.53) correspond to absence of events in the system during a time interval of length h, to abandonment of a queueing customer and to service termination, respectively. Finally, the boundary condition (6.54) corresponds to a new arrival. (That is the reason why the residual patience and the original patience of the l-th customer are equal.)

Introduce the function

$$\begin{aligned}
\varphi(t) &\triangleq \pi_k(x_1+t,\ldots,x_l+t;\, u_1,\ldots,u_l)e^{-n\mu t} \\
&+ \lambda\int_t^\infty \pi_k(x_1+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_l)e^{-n\mu\xi}d\xi \\
&- \sum_{i=1}^{l+1}\int_t^\infty\int_{R_+} \pi_{k+1}(x_1+\xi,\ldots,x_{i-1}+\xi,0,x_i+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_{i-1},u,u_i,\ldots,u_l)e^{-n\mu\xi}dud\xi \\
&- n\mu\int_t^\infty\int_{R_+^2} \pi_{k+1}(x,x_1+\xi,\ldots,x_l+\xi;\, u,u_1,\ldots,u_l)e^{-n\mu\xi}dxdud\xi
\end{aligned} \tag{6.55}$$

Taking into account that

$$e^{-n\mu(t+h)} - e^{-n\mu t} = -e^{-n\mu t}\cdot n\mu h + o(h)$$

we compute

$$\varphi(t+h) - \varphi(t) =$$

$$
\begin{aligned}
&= e^{-n\mu t}\cdot[\pi_k(x_1+t+h,\dots,x_l+t+h;\,u_1,\dots,u_l)-\pi_k(x_1+t,\dots,x_l+t;\,u_1,\dots,u_l)]\\
&- e^{-n\mu t}\cdot n\mu h\cdot\pi_k(x_1+t+h,\dots,x_l+t+h;\,u_1,\dots,u_l)\\
&- \lambda h e^{-n\mu t}\pi_k(x_1+t,\dots,x_l+t;\,u_1,\dots,u_l)\\
&+ he^{-n\mu t}\cdot\sum_{i=1}^{l+1}\int_{R_+}\pi_{k+1}(x_1+t,\dots,x_{i-1}+t,0,x_i+t,\dots,x_l+t;\,u_1,\dots,u_{i-1},u,u_i,\dots,u_l)du\\
&+ n\mu he^{-n\mu t}\cdot\int_{R_+^2}\pi_{k+1}(x,x_1+t,\dots,x_l+t;\,u,u_1,\dots,u_l)dxdu+o(h) \qquad (6.56)
\end{aligned}
$$

If we equate the right part (without $o(h)$) to zero, we get the steady-state equation (6.53) at the point $(x_1+t,\dots,x_l+t;\,u_1,\dots,u_l)$. Hence, given that (6.53) prevails, $\varphi(t+h)-\varphi(t)=o(h)$. Since the function φ is continuous, $\varphi(t)\equiv$ const. In particular,

$$
\varphi(0) \;=\; \varphi(u_l-x_l). \qquad (6.57)
$$

From (6.55),

$$
\varphi(u_l-x_l) \;=\; \pi_k(x_1+u_l-x_l,\dots,x_{l-1}+u_l-x_l,u_l;\,u_1,\dots,u_l)\cdot e^{-n\mu(u_l-x_l)} \qquad (6.58)
$$

(namely, all integrals from (6.55) are zero, since the functions under the integral are zero for $x_l+\xi>u_l$).
Equation (6.54) implies that

$$
\varphi(u_l-x_l) \;=\; \lambda\pi_{k-1}(x_1+u_l-x_l,\dots,x_{l-1}+u_l-x_l,u_l;\,u_1,\dots,u_{l-1})g(u_l)e^{-n\mu(u_l-x_l)}\,.
$$

Note that

$$
\begin{aligned}
\varphi(0) \;&=\; \pi_k(x_1,\dots,x_l;\,u_1,\dots,u_l)\\
&+ \lambda\int_{R_+}\pi_k(x_1+\xi,\dots,x_l+\xi;\,u_1,\dots,u_l)e^{-n\mu\xi}d\xi\\
&- \sum_{i=1}^{l+1}\int_{R_+^2}\pi_{k+1}(x_1+\xi,\dots,x_{i-1}+\xi,0,x_i+\xi,\dots,x_l+\xi;\,u_1,\dots,u_{i-1},u,u_i,\dots,u_l)e^{-n\mu\xi}dud\xi\\
&- n\mu\int_{R_+^3}\pi_{k+1}(x,x_1+\xi,\dots,x_l+\xi;\,u,u_1,\dots,u_l)e^{-n\mu\xi}dxdud\xi \qquad (6.59)
\end{aligned}
$$

Now (6.57), (6.58) and (6.59) imply the integral equation

$$
\begin{aligned}
&\pi_k(x_1,\dots,x_l;\,u_1,\dots,u_l)+\lambda\int_{R_+}\pi_k(x_1+\xi,\dots,x_l+\xi;\,u_1,\dots,u_l)e^{-n\mu\xi}d\xi\\
&= \lambda\pi_{k-1}(x_1+u_l-x_l,\dots,x_{l-1}+u_l-x_l;\,u_1,\dots,u_{l-1})g(u_l)e^{-n\mu(u_l-x_l)}
\end{aligned}
$$

$$+ \quad \sum_{i=1}^{l+1} \int_{R_+^2} \pi_{k+1}(x_1+\xi,\ldots,x_{i-1}+\xi,0,x_i+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_{i-1},u,u_i,\ldots,u_l)e^{-n\mu\xi}dud\xi$$
$$+ \quad n\mu \int_{R_+^3} \pi_{k+1}(x,x_1+\xi,\ldots,x_l+\xi;\, u,u_1,\ldots,u_l)e^{-n\mu\xi}dxdud\xi \tag{6.60}$$

On the other hand, it can be shown that the above formula implies the balance equations (6.53) and (6.54). Specifically, if $x_l = u_l$, (6.54) follows from the definition of Ω_l in (6.41). In order to derive (6.53), one should calculate

$$\pi_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) - \pi_k(x_1+h,\ldots,x_l+h;\, u_1,\ldots,u_l)$$

for small h, using (6.60).

Therefore, we must solve (6.51), (6.52) and (6.60). First,

$$\pi_k \;=\; \omega\lambda^k\mu^{n-k}\frac{n!}{k!}, \qquad k \le n\,, \tag{6.61}$$

where ω is a normalizing constant. We shall search for $\pi_k(\cdot),\ k \ge n$, of the form

$$\pi_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) = \omega\lambda^k \cdot q_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l)\,. \tag{6.62}$$

From (6.61), $\pi_n = \omega\lambda^n$, therefore, $q_n = 1$. Now (6.52) and (6.60) can be rewritten as

$$\int_{R_+} q_{n+1}(0,u)du + n\mu \int_{R_+^2} q_{n+1}(x,u)dxdu \;=\; 1\,, \tag{6.63}$$

and

$$q_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) + \lambda\int_{R_+} q_k(x_1+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_l)e^{-n\mu\xi}d\xi$$
$$= \quad q_{k-1}(x_1+u_l-x_l,\ldots,x_{l-1}+u_l-x_l;\, u_1,\ldots,u_{l-1})g(u_l)e^{-n\mu(u_l-x_l)}$$
$$+ \quad \lambda\sum_{i=1}^{l+1} \int_{R_+^2} q_{k+1}(x_1+\xi,\ldots,x_{i-1}+\xi,0,x_i+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_{i-1},u,u_i,\ldots,u_l)\cdot e^{-n\mu\xi}dud\xi$$
$$+ \quad \lambda n\mu \int_{R_+^3} q_{k+1}(x,x_1+\xi,\ldots,x_l+\xi;\, u,u_1,\ldots,u_l)e^{-n\mu\xi}dxdud\xi\,. \tag{6.64}$$

We now separate (6.64) to two equations and show that they both have the same solution, independent of λ. The two equations are:

$$q_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) \;=\; q_{k-1}(x_1+u_l-x_l,\ldots,x_{l-1}+u_l-x_l;\, u_1,\ldots,u_{l-1})g(u_l)e^{-n\mu(u_l-x_l)} \tag{6.65}$$

and

$$\int_{R_+} q_k(x_1+\xi,\ldots,x_l+\xi;\,u_1,\ldots,u_l)e^{-n\mu\xi}d\xi$$
$$= \sum_{i=1}^{l+1}\int_{R_+^2} q_{k+1}(x_1+\xi,\ldots,x_{i-1}+\xi,0,x_i+\xi,\ldots,x_l+\xi;\,u_1,\ldots,u_{i-1},u,u_i,\ldots,u_l)\cdot e^{-n\mu\xi}dud\xi$$
$$+ \quad n\mu\int_{R_+^3} q_{k+1}(x,x_1+\xi,\ldots,x_l+\xi;\,u,u_1,\ldots,u_l)e^{-n\mu\xi}dxdud\xi\,. \tag{6.66}$$

Equation (6.65) is solved by

$$q_k(x_1,\ldots,x_l;\,u_1,\ldots,u_l) \;=\; I\{(x_1,\ldots,x_l;\,u_1,\ldots,u_l)\in\Omega_l\}\cdot\prod_{i=1}^{l} g(u_i)\cdot e^{-n\mu(u_1-x_1)}\,. \tag{6.67}$$

Substituting (6.67) to the left side of (6.63):

$$\int_0^\infty g(u)e^{-n\mu u}du + n\mu\int_0^\infty\int_0^y g(u)e^{-n\mu(u-x)}dxdu \;=\; \int_0^\infty g(u)du \;=\; 1\,.$$

Hence the solution (6.67) satisfies equation (6.63). Then we must show that it satisfies (6.66). Substitute (6.67) into the following expression

$$\sum_{i=1}^{l+1}\int_{R_+} q_{k+1}(x_1,\ldots,x_{i-1},0,x_i,\ldots,x_l;\,u_1,\ldots u_{i-1},u,u_i,\ldots,u_l)du \tag{6.68}$$
$$+ \quad n\mu\int_{R_+^2} q_{k+1}(x,x_1,\ldots,x_l;\,u,u_1,u_l)dxdu$$
$$= \quad I\{(x_1,\ldots,x_l;\,u_1,\ldots,u_l)\in\Omega_l\}\cdot\prod_{i=1}^{l} g(u_i)\cdot\Bigg(\int_{u_1-x_1}^\infty g(u)e^{-n\mu u}du$$
$$+ \quad e^{-n\mu(u_1-x_1)}\sum_{i=2}^{l+1}\int_{u_i-x_i}^{u_{i-1}-x_{i-1}} g(u)du$$
$$+ \quad n\mu\int_{u_1-x_1}^\infty\int_0^{u-(u_1-x_1)} g(u)e^{-n\mu(u-x)}dxdu\Bigg)\,. \tag{6.69}$$

The last term of (6.69) is equal to

$$n\mu\int_{u_1-x_1}^\infty g(u)e^{-n\mu u}\frac{1}{n\mu}\cdot\left[e^{n\mu[u-(u_1-x_1)]}-1\right]du \;=\; \int_{u_1-x_1}^\infty g(u)\cdot[e^{-n\mu(u_1-x_1)}-e^{-n\mu u}]du\,.$$

Then, going on with the calculation of (6.69), we get

$$= \; I\{(x_1,\ldots,x_l;\,u_1,\ldots,u_l)\in\Omega_l\}\cdot\prod_{i=1}^{l} g(u_i)\cdot e^{-n\mu(u_1-x_1)}\cdot\int_0^\infty g(u)du$$
$$= \; q_k(x_1,\ldots,x_l;\,u_1,\ldots,u_l)\,. \tag{6.70}$$

Applying the equality between (6.68) and (6.70) at the point $(x_1+\xi,\ldots,x_l+\xi;\, u_1,\ldots,u_l)$ and integrating by ξ, we obtain (6.66). Combining (6.67) and (6.62), we get (6.46):

$$\pi_k(x_1,\ldots,x_l;\, u_1,\ldots,u_l) \;=\; I\{(x_1,\ldots,x_l;\, u_1,\ldots,u_l)\}\cdot\omega\lambda^k\cdot\prod_{i=1}^{l} g(u_l)\cdot e^{-n\mu(u_1-x_1)}$$

Finally, we have to integrate (6.46) in order to get the number-in-system steady-state distribution for $k>n$. Making the substitution $u_i=\xi_i+x_i$, we derive

$$\begin{aligned}
\pi_k &= \omega\lambda^k\int_{R_+^{2l}} I\{\xi_1\ge\ldots\ge\xi_l\}\cdot\left[\prod_{i=1}^{l} g(\xi_i+x_i)\right]e^{-n\mu\xi_1}dx_1\ldots dx_l d\xi_1\ldots d\xi_l \\
&= \omega\lambda^k\int_{R_+^{2l}} I\{\xi_1\ge\ldots\ge\xi_l\}\cdot\left[\prod_{i=1}^{l} G(\xi_i)\right]e^{-n\mu\xi_1}d\xi_1\ldots d\xi_l \\
&= \omega\lambda^k\int_0^\infty e^{-n\mu\xi_1}\cdot\frac{1}{(l-1)!}\left[\int_0^{\xi_1}G(\eta)d\eta\right]^{l-1}G(\xi_1)d\xi_1 \\
&= \frac{\omega\lambda^k}{l!}\cdot\int_0^\infty e^{-n\mu\xi_1}\cdot\frac{d}{d\xi_1}\left\{\left[\int_0^{\xi_1}G(\eta)d\eta\right]^{l}\right\}d\xi_1 \\
&= \omega\lambda^k\frac{n\mu}{l!}\cdot\int_0^\infty e^{-n\mu\xi_1}\cdot\left[\int_0^{\xi_1}G(\eta)d\eta\right]^{l}d\xi_1 \\
&\quad \text{(let } \xi\stackrel{\Delta}{=}n\mu\xi_1 \text{ and recall that } l=k-n\text{)} \\
&= \frac{\omega\lambda^k}{(k-n)!}\cdot\int_0^\infty e^{-\xi}\cdot\left[\int_0^{\xi/(n\mu)}G(\eta)d\eta\right]^{k-n}d\xi \\
&= \frac{\omega\lambda^k}{(k-n)!}\cdot\int_0^\infty e^{-\xi}[F(\xi)]^{k-n}d\xi \;=\; \omega\lambda^k F_{k-n},
\end{aligned}$$

according to (6.42).

6.5 M/M/n+G queue: summary of performance measures

Here we summarize *exact* formulae for M/M/n+G performance measures. The following definitions and statements are largely based on Baccelli and Hebuterne [3], but many of them are derived here for the first time.

M/M/n+G primitives.

Recall that the M/M/n+G model requires four input parameters:

λ – arrival rate,

μ – service rate,

n – number of agents,

G – patience distribution ($\bar{G}$ – survival function).

Building blocks.

Define $H(x) \triangleq \int_0^x \bar{G}(u)du$. Note that $H(\infty) = \bar{\tau}$, where $\bar{\tau}$ is the mean patience-time. Introduce the integrals

$$J \triangleq \int_0^\infty \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{6.71}$$

$$J_1 \triangleq \int_0^\infty x \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{6.72}$$

$$J_H \triangleq \int_0^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,. \tag{6.73}$$

In addition, let

$$J(t) \triangleq \int_t^\infty \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{6.74}$$

$$J_1(t) \triangleq \int_t^\infty x \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{6.75}$$

$$J_H(t) \triangleq \int_t^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,. \tag{6.76}$$

Finally, define

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}} = \int_0^\infty e^{-t}\left(1 + \frac{t\mu}{\lambda}\right)^{n-1} dt\ . \tag{6.77}$$

Remark 6.2 A convenient way to calculate $\mathcal{E}$ is via recursion: define

$$\mathcal{E}_k \triangleq \frac{\sum_{j=0}^{k} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}\,, \qquad k \geq 0,$$

and use

$$\mathcal{E}_0 = 1\,; \qquad \mathcal{E}_k = 1 + \frac{k\mu}{\lambda} \cdot \mathcal{E}_{k-1}\,, \quad 1 \leq k \leq n-1\,; \qquad \mathcal{E} = \mathcal{E}_{n-1}\,.$$

List of performance measures:

Many important performance measures of the M/M/n+G queue can be conveniently expressed via the building blocks above and the patience distribution G. Define

P{Ab} – probability to abandon,

P{Sr} – probability to be served,

Q – queue length,

W – waiting time,

V – offered wait (time that a customer with infinite patience would wait).

Then

$$\begin{aligned}
\mathrm{P}\{V>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J}\,, && (6.78)\\
\mathrm{P}\{W>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J}\cdot \bar{G}(0)\,, && (6.79)\\
\mathrm{P}\{\mathrm{Ab}\} &= \frac{1+(\lambda-n\mu)J}{\mathcal{E}+\lambda J}\,, && (6.80)\\
\mathrm{P}\{\mathrm{Ab}\mid \mathrm{V}>0\} &= \frac{1+(\lambda-n\mu)J}{\lambda J}\,, && (6.81)\\
\mathrm{P}\{\mathrm{Sr}\} &= \frac{\mathcal{E}+n\mu J-1}{\mathcal{E}+\lambda J}\,, && (6.82)\\
\mathrm{E}[V] &= \frac{\lambda J_1}{\mathcal{E}+\lambda J}\,, && (6.83)\\
\mathrm{E}[V\mid V>0] &= \frac{J_1}{J}\,, && (6.84)\\
\mathrm{E}[W] &= \frac{\lambda J_H}{\mathcal{E}+\lambda J}\,, && (6.85)\\
\mathrm{E}[Q] &= \frac{\lambda^2 J_H}{\mathcal{E}+\lambda J}\,, && (6.86)\\
\mathrm{E}[V\mid \mathrm{Ab}] &= \frac{(\lambda-n\mu)J_1+J}{(\lambda-n\mu)J+1}\,, && (6.87)\\
\mathrm{E}[W\mid \mathrm{Ab}] &= \frac{J+\lambda J_H-n\mu J_1}{(\lambda-n\mu)J+1}\,, && (6.88)\\
\mathrm{E}[V\mid \mathrm{Sr}] &= \mathrm{E}[W\mid \mathrm{Sr}] \;=\; \frac{n\mu J_1-J}{\mathcal{E}+n\mu J-1}\,, && (6.89)\\
\mathrm{P}\{V>t\} &= \frac{\lambda J(t)}{\mathcal{E}+\lambda J}\,, && (6.90)\\
\mathrm{P}\{W>t\} &= \frac{\lambda \bar{G}(t)J(t)}{\mathcal{E}+\lambda J}\,, && (6.91)\\
\mathrm{E}[V\mid V>t] &= \frac{J_1(t)}{J(t)}\,, && (6.92)
\end{aligned}$$

$$\mathrm{E}[W \mid W > t] \;=\; \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t)J(t)}\,, \tag{6.93}$$

$$\mathrm{P}\{\mathrm{Ab} \mid V > t\} \;=\; \frac{\lambda - n\mu}{\lambda} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda J(t)}\,, \tag{6.94}$$

$$\mathrm{P}\{\mathrm{Ab} \mid W > t\} \;=\; \frac{\lambda - n\mu - G(t)}{\lambda\bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda\bar{G}(t)J(t)}\,. \tag{6.95}$$

Formulae (6.78)-(6.95) are proved in Subsection 6.5.4.

Remark 6.3 In addition to (6.78)-(6.95), one must be able to calculate the four service measures, introduced in Subsection 3.2. The list of formulae above contains expressions for $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{P}\{W > t\}$ and $\mathrm{P}\{\mathrm{Ab}|W > t\}$. The product of the last two provides us with $\mathrm{P}\{W > t; \mathrm{Ab}\}$. The other three service measures are easily derived. For example,

$$\mathrm{P}\{W > t; \mathrm{Sr}\} \;=\; \mathrm{P}\{W > t\} - \mathrm{P}\{W > t; \mathrm{Ab}\}\,.$$

6.5.1 Connection with Erlang-C and Erlang-B

Recall that the M/M/n (Erlang-C) model is equivalent to M/M/n+G with infinite patience. In the notation of the last section, that means $\bar{G}(x) = 1$ and $H(x) = x$, for $0 \le x < \infty$. The offered wait V and the actual wait $W = \min(V, \infty)$ are identical. The building blocks from the last section are now equal to

$$J \;=\; \frac{1}{n\mu - \lambda}\,,$$

$$J_1 \;=\; J_H \;=\; \frac{1}{(n\mu - \lambda)^2}\,,$$

$$J(t) \;=\; \frac{e^{-(n\mu-\lambda)t}}{n\mu - \lambda}\,,$$

and, finally,

$$J_1(t) \;=\; J_H(t) \;=\; \frac{e^{-(n\mu-\lambda)t}}{n\mu - \lambda} \cdot [1 + (n\mu - \lambda)t]\,.$$

(The stability condition $n\mu - \lambda > 0$ must prevail.) The expression for $\mathcal{E}$ remains the same as in (6.77). Now substituting building blocks into the relevant formulae from (6.78)-(6.95), and defining offered load per server by

$$\rho \;\triangleq\; \frac{\lambda}{n\mu}\,,$$

we get several well-known formulae for the Erlang-C model:

$$\begin{aligned} \mathrm{P}\{W>0\} &= \frac{\rho}{\rho+(1-\rho)\mathcal{E}}\,, \qquad (6.96)\\ \mathrm{E}[W|W>0] &= \frac{1}{n\mu-\lambda}\,, \qquad (6.97)\\ \mathrm{P}\{W>t|W>0\} &= e^{-(n\mu-\lambda)t}\,,\\ \mathrm{E}[W|W>t] &= t+\frac{1}{n\mu-\lambda}\,. \end{aligned}$$

(It is straightforward to check that (6.96) is equivalent to classical Erlang-C formula.)

Now note that the M/M/n/n model (Erlang-B) is M/M/n+G with customers that are not willing to wait at all. If wait is encountered, they abandon (blocked in the terms of Erlang-B). In this case, $H(x)\equiv 0$, and the only relevant building block is equal to

$$J = \frac{1}{n\mu}\,,$$

and

$$\mathrm{P}\{W>0\} = \mathrm{P}\{\text{Blocking}\}_{M/M/n/n} = \frac{\rho}{\rho+\mathcal{E}}\,. \qquad (6.98)$$

Formula (6.98) is equivalent to the classical Erlang-B formula:

$$\mathrm{P}\{\text{Blocking}\}_{M/M/n/n} = \frac{\frac{(\lambda/\mu)^n}{n!}}{\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!}}\,.$$

6.5.2 Special case. Exponential patience (M/M/n+M, Erlang-A).

In this case, patience times are exponential with parameter θ. Then

$$H(x) = \frac{1}{\theta}\cdot(1-e^{-\theta x})\,,$$

and the four building blocks are equal to

$$\begin{aligned} J &= \frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)\,,\\ J(t) &= \frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}e^{-\theta t}\right)\,,\\ J_H &= \frac{J}{\theta}-\frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta^2}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}+1}\cdot\gamma\left(\frac{n\mu}{\theta}+1,\frac{\lambda}{\theta}\right)\,,\\ J_H(t) &= \frac{J(t)}{\theta}-\frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta^2}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}+1}\cdot\gamma\left(\frac{n\mu}{\theta}+1,\frac{\lambda}{\theta}e^{-\theta t}\right)\,, \end{aligned}$$

where the incomplete Gamma function $\gamma(x, y)$ was defined in (6.6).

Note, however, that J_1 and $J_1(t)$ cannot be expressed via the incomplete Gamma function. Consequently, those formulae from (6.78)-(6.95) that involve J_1 (e.g. for the average offered wait $\mathrm{E}[V]$) should be calculated either numerically, or by approximations, as discussed in the sequel.

6.5.3 Special case. Deterministic patience (M/M/n+D).

Another important special case is a queue with patience times equal to a constant D. In this case,

$$H(x) \;=\; \begin{cases} x, & 0 \le x \le D \\ D, & x > D \end{cases} .$$

If $\lambda - n\mu \neq 0$,

$$J \;=\; \frac{1}{n\mu - \lambda} - \frac{\lambda}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D} ,$$

$$J(t) \;=\; \begin{cases} \dfrac{1}{n\mu - \lambda} \cdot e^{-(n\mu-\lambda)t} - \dfrac{\lambda}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D} , & t < D \\ \dfrac{1}{n\mu} \cdot e^{\lambda D - n\mu t} , & t \ge D \end{cases}$$

$$J_1 \;=\; \frac{1}{(n\mu - \lambda)^2} - \left[\frac{1}{(n\mu - \lambda)^2} - \frac{1}{(n\mu)^2} + \frac{\lambda D}{n\mu(n\mu - \lambda)} \right] \cdot e^{-(n\mu-\lambda)D} ,$$

$$J_1(t) \;=\; \begin{cases} \dfrac{e^{-(n\mu-\lambda)t} - e^{-(n\mu-\lambda)D}}{(n\mu - \lambda)^2} + \dfrac{te^{-(n\mu-\lambda)t} - De^{-(n\mu-\lambda)D}}{n\mu - \lambda} + \left[\dfrac{D}{n\mu} + \dfrac{1}{(n\mu)^2} \right] \cdot e^{-(n\mu-\lambda)D} , & t < D \\ \left[\dfrac{t}{n\mu} + \dfrac{1}{(n\mu)^2} \right] \cdot e^{\lambda D - n\mu t} , & t \ge D \end{cases}$$

$$J_H \;=\; \frac{1}{(n\mu - \lambda)^2} \cdot [1 - e^{-(n\mu-\lambda)D}] - \frac{\lambda D}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D} ,$$

$$J_H(t) \;=\; \begin{cases} \dfrac{1}{(n\mu - \lambda)^2} \cdot [e^{-(n\mu-\lambda)t} - e^{-(n\mu-\lambda)D}] + \dfrac{t}{n\mu - \lambda} \cdot e^{-(n\mu-\lambda)t} - \dfrac{\lambda D}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D} , & t < D \\ \dfrac{D}{n\mu} \cdot e^{\lambda D - n\mu t} , & t \ge D \end{cases}$$

If $\lambda - n\mu = 0$,

$$J \;=\; D + \frac{1}{n\mu} ,$$

$$J(t) \;=\; \begin{cases} D - t + \dfrac{1}{n\mu}\,, & t < D \\[2ex] \dfrac{1}{n\mu}\cdot e^{\lambda D - n\mu t}\,, & t \geq D \end{cases}$$

$$J_1 \;=\; \frac{D^2}{2} + \frac{D}{n\mu} + \frac{1}{(n\mu)^2}\,,$$

$$J_1(t) \;=\; \begin{cases} \dfrac{D^2 - t^2}{2} + \left[\dfrac{D}{n\mu} + \dfrac{1}{(n\mu)^2}\right]\cdot e^{-(n\mu-\lambda)D}\,, & t < D \\[2ex] \left[\dfrac{t}{n\mu} + \dfrac{1}{(n\mu)^2}\right]\cdot e^{\lambda D - n\mu t}\,, & t \geq D \end{cases}$$

$$J_H \;=\; \frac{D^2}{2} + \frac{D}{n\mu}\,,$$

$$J_H(t) \;=\; \begin{cases} \dfrac{D^2 - t^2}{2} + \dfrac{D}{n\mu}\,, & t < D \\[2ex] \dfrac{D}{n\mu}\cdot e^{\lambda D - n\mu t}\,, & t \geq D \end{cases}$$

6.5.4 Proofs of (6.78)-(6.95)

Here we present the proofs of (6.78)-(6.95), one by one.

(6.78). First, (6.25)-(6.28) and definition (6.77) imply the useful formula

$$\pi_{n-1} \;=\; \frac{\dfrac{1}{(n-1)!}\left(\dfrac{\lambda}{\mu}\right)^{n-1}}{\displaystyle\sum_{j=0}^{n-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^{j} + \frac{\lambda J}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}} \;=\; \frac{1}{\mathcal{E} + \lambda J} \tag{6.99}$$

Now the last equality of (6.37) implies (6.78).

(6.79). Follows from

$$\mathrm{P}\{W > 0 | V > 0\} \;=\; \bar{G}(0)\,.$$

(6.80). Formula (6.29) implies that

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \left(1 - \frac{n\mu}{\lambda}\right)\cdot \mathrm{P}\{V > 0\} + \frac{1}{\mathcal{E} + \lambda J}\,.$$

Now substitute (6.78).

(6.81). Immediate consequence of (6.78) and (6.80).

(6.82).

$$\mathrm{P}\{\mathrm{Sr}\} \;=\; 1 - \mathrm{P}\{\mathrm{Ab}\}\,.$$

(6.83). Results from (6.99) and

$$\mathrm{E}[V] \;=\; \lambda\pi_{n-1}\cdot\int_0^\infty x\exp\{\lambda H(x) - n\mu x\}dx\,.$$

(6.84). Formulae (6.78) and (6.83).

(6.85). According to formula (6.26), the survival function of the virtual wait is given by

$$\mathrm{P}\{V > t\} \;\triangleq\; \bar{V}(t) \;=\; \lambda\pi_{n-1}\int_t^\infty \exp\left\{\lambda H(x) - n\mu x\right\}dx\,.$$

Hence, the average wait is equal to

$$\mathrm{E}[W] \;=\; \int_0^\infty \bar{G}(t)\bar{V}(t)dt \;=\; \lambda\pi_{n-1}\int_0^\infty \bar{G}(t)\cdot\int_t^\infty \exp\left\{\lambda H(x) - n\mu x\right\}dxdt$$

and integrating by parts

$$\mathrm{E}[W] \;=\; \lambda\pi_{n-1}\int_0^\infty H(t)\cdot\exp\left\{\lambda H(t) - n\mu t\right\}dt\,.$$

Then use formula (6.99) and definition (6.73).

(6.86). Follows from (6.85) and Little's formula.

(6.87).

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{\mathrm{E}[V\cdot 1_{\{\tau\le V\}}]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty xv(x)G(x)dx}{\mathrm{P}\{\mathrm{Ab}\}}\,, \tag{6.100}$$

where from (6.26) and (6.99)

$$v(x) \;=\; \frac{\lambda\exp\{\lambda H(x) - n\mu x\}}{\mathcal{E} + \lambda J}\,. \tag{6.101}$$

Integration by parts implies that

$$\int_0^\infty x[\lambda\bar{G}(x) - n\mu]\exp\{\lambda H(x) - n\mu x\}dx$$

$$= \int_0^\infty xd\left[\exp\{\lambda H(x) - n\mu x\}\right] \;=\; -\int_0^\infty \exp\{\lambda H(x) - n\mu x\}dx \;=\; -J\,,$$

and

$$\int_0^\infty xG(x)\exp\{\lambda H(x) - n\mu x\}dx \;=\; \frac{(\lambda - n\mu)J_1 + J}{\lambda}\,, \tag{6.102}$$

which, combined with (6.100), (6.101) and (6.80) implies

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{(\lambda - n\mu)J_1 + J}{(\lambda - n\mu)J + 1}\,.$$

(6.88). Similar to the previous calculation

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \frac{\mathrm{E}[\tau \cdot 1_{\{\tau \le V\}}]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty x\bar{V}(x)dG(x)}{\mathrm{P}\{\mathrm{Ab}\}}\,.$$

Note that

$$d[xG(x) + H(x) - x] = xdG(x)\,. \tag{6.103}$$

Then

$$\frac{\int_0^\infty x\bar{V}(x)dG(x)}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty v(x)\cdot[xG(x) + H(x) - x]dx}{\mathrm{P}\{\mathrm{Ab}\}}$$

(use (6.101) and (6.80))

$$=\; \frac{\lambda\int_0^\infty [xG(x) + H(x) - x]\cdot\exp\{\lambda H(x) - n\mu x\}dx}{1 + (\lambda - n\mu)J} \;=\; \frac{J + \lambda J_H - n\mu J_1}{1 + (\lambda - n\mu)J}\,,$$

where the last equality follows from (6.102) and the definitions of J_1 and J_H.

(6.89). This formula for $\mathrm{E}[V|\mathrm{Sr}]$ can be checked via

$$\mathrm{E}[V] = \mathrm{E}[V|\mathrm{Sr}]\cdot\mathrm{P}\{\mathrm{Sr}\} + \mathrm{E}[V|\mathrm{Ab}]\cdot\mathrm{P}\{\mathrm{Ab}\}\,.$$

Since the event {Sr} is equivalent to {W=V},

$$\mathrm{E}[V|\mathrm{Sr}] \;=\; \mathrm{E}[W|\mathrm{Sr}]\,.$$

(6.90). Follows from (6.101).

(6.91). Consequence of

$$\mathrm{P}\{W > t\} \;=\; \mathrm{P}\{V > t\}\cdot\mathrm{P}\{\tau > t\}\,.$$

(6.92). Follows from definitions of $J(t)$ and $J_1(t)$.

(6.93).

$$\mathrm{E}[W|W > t] \;=\; \frac{\int_t^\infty xw(x)dx}{\mathrm{P}\{W > t\}}\,, \tag{6.104}$$

where $w(x)$ is the waiting-time density. The denominator of (6.104) is equal to

$$\mathrm{P}\{W > t\} \;=\; \bar{G}(t)\bar{V}(t) \;=\; \lambda\pi_{n-1}\bar{G}(t)\int_t^\infty \exp\{\lambda H(x) - n\mu x\}dx\,. \tag{6.105}$$

Calculating the numerator of (6.104):

$$\begin{aligned}\int_t^\infty xw(x)dx \;&=\; \int_t^\infty xv(x)\bar{G}(x)dx + \int_t^\infty x\bar{V}(x)dG(x)\\ &=\; \lambda\pi_{n-1}\left[\int_t^\infty x\bar{G}(x)\exp\{\lambda H(x) - n\mu x\}dx\,+\right.\\ &\left.\quad+\int_t^\infty x\left(\int_x^\infty \exp\{\lambda H(u) - n\mu u\}du\right)dG(x)\right]\end{aligned} \tag{6.106}$$

Use (6.103) to show that the double integral in (6.106) is equal to

$$\begin{aligned}&\int_t^\infty [xG(x) + H(x) - x]\cdot\exp\{\lambda H(x) - n\mu x\}dx\,-\\ &-\;\;[tG(t) + H(t) - t]\cdot\int_t^\infty \exp\{\lambda H(x) - n\mu x\}dx\,.\end{aligned} \tag{6.107}$$

After some terms cancel, we get from (6.104), (6.105), (6.106) and (6.107) that

$$\begin{aligned}\mathrm{E}[W|W > t] \;&=\; \frac{\int_t^\infty [H(x) + t\bar{G}(t) - H(t)]\cdot\exp\{\lambda H(x) - n\mu x\}dx}{\bar{G}(t)\cdot\int_t^\infty \exp\{\lambda H(x) - n\mu x\}dx}\\ &=\; \frac{J_H(t) - (H(t) - t\bar{G}(t))\cdot J(t)}{\bar{G}(t)J(t)}\,.\end{aligned}$$

(6.94).

$$\begin{aligned}\mathrm{P}\{\mathrm{Ab}|V > t\} \;&=\; \frac{\mathrm{P}\{\tau \le V;\, V > t\}}{\mathrm{P}\{V > t\}} \;=\; \frac{\int_t^\infty G(x)v(x)dx}{\int_t^\infty v(x)dx}\\ &=\; \frac{\int_t^\infty G(x)\cdot\exp\{\lambda H(x) - n\mu x\}dx}{J(t)}\,.\end{aligned} \tag{6.108}$$

Using integration by parts

$$\int_t^\infty [\lambda\bar{G}(x) - n\mu]\cdot\exp\{\lambda H(x) - n\mu x\}dx \;=\; -\exp\{\lambda H(t) - n\mu t\}\,.$$

Hence,

$$\int_t^\infty G(x)\cdot\exp\{\lambda H(x) - n\mu x\}dx \;=\; \frac{(\lambda - n\mu)J(t) + \exp\{\lambda H(t) - n\mu t\}}{\lambda}\,. \tag{6.109}$$

Now (6.108) and (6.109) imply

$$\mathrm{P}\{\mathrm{Ab}\mid V > t\} \;=\; \frac{\lambda - n\mu}{\lambda} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda J(t)}\,.$$

(6.95).

$$\mathrm{P}\{\mathrm{Ab}|W>t\} = \frac{\mathrm{P}\{\tau \le V;\, \tau > t\}}{\mathrm{P}\{\min(V,\tau)>t\}} = \frac{\int_t^\infty \bar{V}(x)dG(x)}{\bar{G}(t)\bar{V}(t)}$$

$$= \frac{G(x)\bar{V}(x)|_t^\infty + \int_t^\infty G(x)v(x)dx}{\bar{G}(t)\bar{V}(t)} = \frac{\int_t^\infty G(x)\cdot\exp\{\lambda H(x)-n\mu x\}dx - G(t)J(t)}{\bar{G}(t)J(t)}$$

$$= \frac{\lambda - n\mu - G(t)}{\lambda\bar{G}(t)} + \frac{\exp\{\lambda H(t)-n\mu t\}}{\lambda\bar{G}(t)J(t)}\,.$$

7 Queueing theory: relevant asymptotic results

Although exact formulae for the Erlang-A and M/M/n+G queues are available, they are too complicated for developing guidelines to call center managers. These formulae cannot provide insight to practical questions of the type: "how many additional agents would I need if the arrival rate doubles?", "how sensitive is our model to a possible error in patience estimate?" etc.

Thus, approximations are useful for providing insight and simplifying computations. In addition, more general models (e.g. queues with a general service time distribution) often lack analytical solution. Therefore, approximations can help to extend modelling scope.

There exist the two main types of approximations:

- *Steady-state approximations* provide asymptotic expressions for steady-state performance measures of a queueing system (probability to abandon, average queue etc.)

- *Process-limit approximations* provide asymptotics for the model processes, for example the queue-length process or the offered-wait process.

In our research, we develop steady-state approximations. Below we cover relevant asymptotic results, emphasizing steady-state *QED*, *Efficiency-Driven (ED)* and *Quality-Driven (QD)* approximations, which will be studied in Part IV.

7.1 Classical approximations for queues without abandonment

Extensive research has been done on the so-called *heavy-traffic approximations*, which were first developed by Kingman [45, 46] for G/G/1 and then G/G/n. In this framework, the number of agents n is held fixed and the agents' occupancy ρ converges to the critical value of one (100% utilization) from below. Then the steady-state waiting time and the queue-length in the G/G/n model converge to infinity. However, if these performance measures are properly normalized, their limit steady-state distribution is exponential. See the book of Whitt [68] as a general reference for heavy-traffic approximations.

In our research, which is oriented towards call centers, we are mainly interested in models with a large number of agents n. Formally, we assume that the number of agents n and the arrival rate λ converge to infinity and, then, index a sequence of models by either n or λ. (As a rule, we omit this indexing in formulae.)

For example, consider another version of a heavy-traffic limit for the Erlang-C queue. (See, for example, the Appendix of Gurvich [32].) Fix the service rate μ, let the arrival rate $\lambda \to \infty$ and assume that

$$n \;=\; R + \gamma\,, \tag{7.1}$$

where $R = \lambda/\mu$ is the *offered load* and $\gamma > 0$ is a *service grade* parameter. Then one can verify that

- The delay probability $\mathrm{P}\{W > 0\}$ converges to one.
- Agents occupancy ρ converges to one, since $\lim_{\lambda\to\infty} n(1-\rho) \;=\; \gamma$.
- The steady-state waiting time is approximately exponential with rate $\gamma\mu$.
- Properly normalized and time-changed, the queue-length process converges to a reflected Brownian motion with a negative drift.

The staffing rule (7.1) above is an example of the *efficiency-driven operational regime*: it emphasizes the efficiency side in the quality/efficiency tradeoff, as discussed in Subsection 1.1 of the Introduction.

An additional important approximation was first introduced by Iglehart [37]. Consider the G/G/n queue with fixed μ. Let $\lambda \to \infty$ and assume the staffing relation

$$n \;=\; R\cdot(1+\gamma) \;=\; R + \gamma R\,. \tag{7.2}$$

Then

- The delay probability and the average wait converge to zero.
- The agents occupancy converges to $\dfrac{1}{1+\gamma}$.
- The steady-state number-in-system is asymptotically normal with mean R and standard deviation $\sqrt{R}$.
- Properly normalized, the number-in-system process converges to an Ornstein-Uhlenbeck diffusion process.

The last approximation corresponds to a *quality-driven operational regime*, since it implies especially high performance level.

7.2 QED approximations for queues without abandonment

The classical approximations described above fail to capture an operational regime that can be observed in many well-run call centers. It is first characterized by a high agents' utilization (e.g. 90-98%). However, in contrast to classical heavy-traffic, a non-negligible fraction of the customers (e.g. 30-70%) get service immediately. In addition, average wait is small with respect to the average service time. The paper of Sze [63] described this operational regime, although it focused mainly on classical heavy-traffic.

This high-utilization high-service-level operational regime was introduced in Subsections 3.1 and 4.2 as the QED (Quality & Efficiency Driven) regime. The QED regime for the basic Erlang-C model was formally defined and analyzed by Halfin and Whitt [34]. They fixed the service rate μ and assumed that $\lambda, n \to \infty$. Then the following three asymptotic statements are equivalent:

- The *square root* staffing rule prevails:

$$n \;=\; R + \beta\sqrt{R} + o(\sqrt{R})\,, \qquad \beta > 0, \tag{7.3}$$

  where $\beta > 0$ is the *service grade*.

- Agents' occupancy ρ converges to one, or more precisely

$$\sqrt{n}\cdot(1-\rho) \;\to\; \beta\,. \tag{7.4}$$

- Delay probability converges to a constant:

$$\mathrm{P}\{W>0\} \;\to\; \alpha(\beta) \;\triangleq\; \left[1+\frac{\beta}{h(-\beta)}\right]^{-1}, \tag{7.5}$$

  where $h(\cdot)$ is the hazard rate of the standard normal distribution.

In addition, the square root staffing rule implies the following two approximations for the waiting time in M/M/n:

$$\mathrm{E}[W] \;\sim\; \frac{\alpha(\beta)}{\mu\beta\sqrt{R}} \;\sim\; \frac{\alpha(\beta)}{\mu\beta\sqrt{n}}, \tag{7.6}$$

$$\mathrm{P}\{W>t\} \;\sim\; \alpha(\beta)\cdot e^{-t\beta\mu\sqrt{R}}. \tag{7.7}$$

Summarizing (7.4)-(7.6), note that it is consistent with an informal description of the QED regime in Subsection 4.2: utilization is high; delay probability is neither close to zero, nor to one; and, finally, the average wait converges to zero. Recently, Whitt [71] developed a framework that unified the three Erlang-C approximations: QD, ED and QED. Borst et al. [9] also unifies these approximations via the analysis of staffing and waiting costs.

QED analysis for the Erlang-B (M/M/n/n) model was carried out by Jagerman [39]. He showed that for the staffing rule (7.3) the blocking probability is of order $1/\sqrt{n}$:

$$E_{1,n} \;\triangleq\; \mathrm{P}\{\text{Blocked}\} \;\approx\; \frac{h(-\beta)}{\sqrt{n}}, \qquad -\infty<\beta<\infty\,. \tag{7.8}$$

The G/D/n queueing model (general interarrival times, deterministic service) in the QED framework was analyzed in Jelencovic, Mandelbaum and Momcilovic [41]. The M/M/n/k model with possible busy signals (n agents, $(k-n)$ waiting spaces in queue) was treated by Massey and Wallace [54]. Puhalskii and Reiman [58] prove weak convergence for (the queue and virtual wait) processes of the GI/PH/n system to a complex, multidimensional diffusion process, but not its steady state.

In addition, it turns out that the QED staffing regime can be analyzed with Skill-Based Routing (SBR). Armony and Mandelbaum [1] and Gurvich [32] explore two classical and basic SBR models: $\bigwedge$-design and $\bigvee$-design, respectively.

7.3 Erlang-A approximations

Recall that a basic model that takes into account abandonment is Erlang-A (M/M/n+M), which was treated in Subsection 6.1. Garnett, Mandelbaum and Reiman [29] developed QED approximations for Erlang-A. They established equivalence of the following points of view:

- The *square root* staffing rule prevails:

$$n \;=\; R + \beta\sqrt{R} + o(\sqrt{R})\,, \qquad -\infty < \beta < \infty\,. \tag{7.9}$$

  (Note that here the service grade β can be negative, since abandonment stabilize the system at all staffing levels.)

- Delay probability converges to a constant:

$$\mathrm{P}\{W > 0\} \;\sim\; \left[1 + \frac{h\left(\beta\sqrt{\frac{\mu}{\theta}}\right)}{h(-\beta)\sqrt{\frac{\mu}{\theta}}}\right]^{-1}\,, \tag{7.10}$$

  where θ is the individual abandonment rate in Erlang-A.

- Probability to abandon is of the order $\dfrac{1}{\sqrt{n}}$:

$$\mathrm{P}\{Ab\} \;\sim\; \frac{\delta}{\sqrt{n}}\,,$$

  where $\delta > 0$ depends on β and the μ/θ ratio.

- Asymptotic offered load per agent is

$$\rho \;\triangleq\; \frac{\lambda}{n\mu} \;\approx\; 1 - \frac{\beta + \delta}{\sqrt{n}}\,.$$

7.4 Approximations for M/M/n+G and its extensions

In [70], Whitt develops and validates an approximation for the M/G/n+G model with generally distributed iid service times. In fact, he approximates it by an M/M/n+M(k) Markovian model with abandonment rates that depend on the number-in-system k. The major insight of this paper can be summarized by the following two statements:

- M/G/n+G performance is primarily affected by the service-time distribution through its mean; therefore, numerical approximations for M/M/n+G are of main interest.

- M/G/n+G behavior is primarily affected by the patience-time distribution through its hazard rate near the origin. This statement is consistent with some of our results that will be presented in Part IV.

Naturally, both statements cannot be true for the entire range of the four model parameters. However, they seem to be appropriate for distributions and parameters occurring in call centers: relatively large n and small to moderate, but non-negligible, customers' abandonment.

Whitt [73] provides additional insight into the approximation proposed above. He compares between efficiency-driven approximations of the two models: exact M/G/n+G and approximate M/M/n+M(k) of [70].

Whitt [72] presents a general fluid model (efficiency-driven approximation) for the G/G/n+G queue with general distributions of arrivals services and patience times.

Since Erlang-A and other queueing models with abandonment are very sensitive to changes in the arrival rate (see Whitt [74]), it is important to consider models with uncertainty about the arrival-rate. Recent papers of Whitt [75] and Bassamboo, Harrison and Zeevi [5] study efficiency-driven approximations for such models and develop asymptotic rules of optimal staffing. Note that the ED approximations are cruder than the QED ones, hence they enable analysis of very general models.

Finally, Ward and Glynn [67] use another type of scaling. They explore the M/M/1+M queue (Erlang-A with a single agent), assuming that the arrival rate exceeds or is close to the service rate and the individual abandonment rate is close to zero. It turns out that this system can be approximated by either a reflected Ornstein-Uhlenbeck process or a reflected affine diffusion, depending on how parameters are driven to their limits.

8 Statistical background on the M/M/n+G model

In order to apply the M/M/n+G model in the call center environment, it is essential to be able to estimate its parameters. First, there are three numerical parameters: λ, μ,

and n. Their estimates are usually based on historical ACD data. Below, in Subsections 8.1-8.3, we briefly outline corresponding estimation and prediction procedures.

One must also estimate either the patience distribution or its characteristics that constitute input to the M/M/n+G formulae, either exact or approximate. This issue is treated in Subsection 8.4.

8.1 Estimation of the arrival rate

Arrivals of incoming calls are typically assumed Poisson, with time-varying arrival rates. The goal is to estimate/predict these arrival rates, over short time-intervals (15, 30 minutes or one hour), chosen so that the rates are approximately constant during an interval. Then the time-homogeneous model is applied separately over each such interval.

The goal can be achieved in two stages. First, time-series algorithms are used to predict *daily* volumes, taking into account trends and special days (eg. holidays, "Mondays", special sales). Second, one uses (non)parametric regression techniques for predicting the *fraction* of arrivals per time-interval, out of the daily-total. This fraction, combined with the daily total, yields actual arrival rates per each time-interval. (See [13] for detailed treatment).

8.2 Estimation of the average service time

Service durations are assumed exponential. Average service times tend to be relatively stable from day to day and from hour to hour. (However, they often change depending on the time-of-day! See [13].)

In practice, service consists of several phases, notably talk time, wrap-up time (after-call work), and what is sometimes referred to as auxiliary time. An easier-to-grasp notion is thus "idle-time", namely the time that an agent is immediately accessible for service. It is thus also possible to estimate the average service time during a time interval by:

$$\frac{\text{Total Working Time} - \text{Total Idle Time}}{\text{Number of Served Customers}},$$

where Total Working Time is the product of the Number of Agents by the Interval Duration.

8.3 Estimation of the number of agents

In performance analysis, the number of agents n is an M/M/n+G input. In staffing, n is an output. In both cases, n must be normalized by the *rostered staff factor (RSF)*, or *shrinkage factor*, which accounts for absenteeism, unscheduled breaks etc. (See Cleveland and Mayben [16]). For example, if 100 agents are required for answering calls, in fact more agents (110, 120, ...) should be assigned to shift, depending on RSF. In addition, one should carefully check whether staffing-level data from ACD reports is reliable. Specifically, is planned or actual number of agents reported? Does the ACD data take into account time periods (both scheduled and not) when agents are taking breaks?

In practice, our experience suggests that it is harder to get reliable historical data on the number of agents, than on arrival or service rates. In our application analysis (Section 19 and Subsection 6.1.3) we did not have data on the number-of-agents and thus we had to resort to heuristic procedures in order to estimate it. See Section 19 for details.

8.4 Estimation of the patience-time distribution

Recall the framework that we use in order to incorporate the abandonment phenomenon:

- An arriving customer is equipped with *patience time* τ. Patience times are G-distributed and iid;
- An arriving customer encounters an *offered wait* V;
- Actual waiting time of the customer is equal to $W = \min(\tau, V)$;
- If $\tau \leq V$, the customer abandons; otherwise, the customer gets service.

Note that we observe patience times of only customers that abandon the queue. If a customer gets service, we deduce that his patience τ is larger than the actual wait W. This is a classical example of a *censoring* problem.

Techniques for analyzing censored data have been developed within the well-established statistical branch of Survival Analysis. See Cox and Oakes [17] for a classical treatment, or Fleming and Harrington [25] for an advanced measure-theoretic approach. We now review the application of relevant techniques to our data, which we shall use in Section 19.

Kaplan-Meier estimator. Our data from call centers in a US bank is discrete with a resolution of one second. Specifically, for each call, we observe its waiting time in seconds $(0, 1, 2, \ldots)$ and the call outcome (served or abandoned). Assume that we analyze a sample of calls with the same patience distribution over the sample. Denote by A_k and S_k, respectively, the number of abandonment and the number of service starts exactly at k seconds. Let η_k be the number of customers that is neither served nor abandoned before k seconds (number-at-risk in Survival Analysis).

Then the non-parametric maximum-likelihood estimator of the discrete hazard rate is given by

$$\hat{h}_k \;=\; \frac{A_k}{\eta_k}\,, \qquad k \geq 0\,. \tag{8.1}$$

The Kaplan-Meier (product-limit) estimator of the survival function of patience times is then

$$\hat{\bar{G}}(t) \;=\; \prod_{k \leq t}(1 - \hat{h}_k)\,, \qquad t \geq 0\,. \tag{8.2}$$

The Greenwood formula provides asymptotic variance of the patience survival function:

$$\mathrm{Var}[\hat{\bar{G}}(t)] \;=\; [\hat{\bar{G}}(t)]^2 \cdot \sum_{k \leq t} \frac{A_k}{\eta_k(\eta_k - A_k)}\,. \tag{8.3}$$

Actuarial estimator. In practice, patience and waiting-time distributions are continuous rather than discrete. An integer waiting time $T > 0$ in the database is approximately a rounded wait in $(T-0.5, T+0.5)$. (In fact, given our knowledge about the measurement system in the US bank, it is more like a triangular distribution over $(T-1, T+1)$.)

Therefore, there is a need to modify assumptions of the discreet Kaplan-Meier estimator. Assume that the distribution of patience times is continuous with piecewise-constant hazard rate. Let the hazard be equal to h_j during intervals $[a_{j-1}, a_j)$, $j \geq 1$, with the convention $a_0 = 0$ and $a_0 < a_1 < a_2 \ldots$ Set the interval widths

$$b_j \;\triangleq\; a_j - a_{j-1}\,.$$

Under these assumptions, formulae (8.1)-(8.3) for the MLE should be modified in the following way:

$$\hat{h_k} \;=\; \frac{A_k}{b_k(\eta_{k-1} - 0.5 \cdot (S_k + A_k))}\,, \qquad k \geq 1, \tag{8.4}$$

$$\hat{\bar{G}}(a_j) \;=\; \prod_{k\le j}\left(1-\frac{A_k}{\eta_k'}\right), \qquad j\ge 1\,, \tag{8.5}$$

$$\mathrm{Var}[\hat{\bar{G}}(a_j)] \;=\; [\hat{\bar{G}}(a_j)]^2\cdot\frac{A_k}{\eta_k'(\eta_k'-S_k)}\,, \qquad j\ge 1\,, \tag{8.6}$$

where

$$\eta_k' \;\triangleq\; \eta_{k-1}-\frac{1}{2}S_k$$

is the adjusted number at risk in $[a_{j-1},a_j)$. The estimator (8.5) is called the *actuarial estimator* of the survival function.

Robustness of patience estimates. In a well-run call center, the probability to abandon usually does not exceed several percents. Hence, the censoring affects 95-98% of call-by-call data, taking place at the early stages of wait. In addition, in most of our applications, tele-customers seem to be rather patient, in the sense that most of them are ready to wait several minutes for service. Combining the two facts, it follows that standard methods of survival analysis are often hard for implementation in the call center environment. For example, assume that we try to estimate the average patience via the tail formula:

$$\widehat{\mathrm{E}[\tau]} \;=\; \int_0^\infty \hat{\bar{G}}(t)dt\,, \tag{8.7}$$

where $\hat{\bar{G}}(t)$ is the Kaplan-Meier estimator (8.2) or (8.5). The value of (8.7) strongly depends on the survival estimate for large t which, according to our experience, turns out to be very unreliable: the number of large waiting times is relatively small, and many of them can be, in fact, due to bugs and measurement errors in the system.

In contrast, waiting data for the smaller time values (say, up to 1 minute) is far more extensive and reliable. Hence, it is desirable to develop and use estimates that are based on this data only, which is yet to be done.

Independence between observations. The theory behind formulae (8.1)-(8.6) assumes that patience times and censoring (offered wait) times are independent random variables. Such independence is plausible for patience times. In tele-queues, it is also reasonable to assume that patience of a customer is independent of his offered wait. (Although, announcements informing customers on their expected wait can lead to violation

of this assumption.) However, successive offered-wait (censoring) times are clearly dependent. We assume that given a large number of observations from different days, the effect of such dependence is negligible.

Estimation of the individual abandonment rate in the Erlang-A model. If patience is assumed exponential, one can use the relation (6.14) in order to estimate the individual abandonment rate θ. The average wait in queue, $\mathrm{E}[W]$, and the fraction of customers abandoning, $\mathrm{P}\{\mathrm{Ab}\}$, are in fact standard ACD data outputs, thus, providing the means for estimating θ as follows:

$$\hat{\theta} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\%\text{Abandonment}}{\text{Average Wait}}\,.$$

The formula above is also equivalent to the MLE for a censored exponential parameter.

9 Asymptotic behavior of integrals

We have seen that the building blocks of the M/M/n+G model have an integral form (recall formulae (6.71)-(6.77)). In Sections 15-17 we shall calculate various approximations for these building blocks and, consequently, for the M/M/n+G performance measures. To this end, we now develop a general method and prove several lemmas that will help us in the task.

9.1 The Laplace method

In the proofs of Part IV, we repeatedly derive asymptotic approximations for integrals that are expressed in the form

$$\int_0^\infty x^m \cdot e^{-f_\lambda(x)} dx\,, \qquad \lambda \to \infty\,. \tag{9.1}$$

As a rule, $f_\lambda(0) = 0$, for all $\lambda > 0$, and $f_\lambda(x) \to \infty$, as $\lambda \to \infty$, for all $x > 0$. Note that the exponential term rapidly converges to zero, for $x > 0$. Hence, one could expect that, as $\lambda \to \infty$, the value of (9.1) depends mainly on behavior of the integrand near the origin.

An important special case is given by

$$\int_0^\infty x^m \cdot \exp\left\{-b\lambda^k x^l\right\} \;=\; \frac{\Gamma\left(\frac{m+1}{l}\right)}{l b^{\frac{m+1}{l}}} \cdot \lambda^{-\frac{k(m+1)}{l}}\,, \tag{9.2}$$

where $k \geq 0$, $l > 0$, $b > 0$ and $m \geq 0$. If $m = 0$ one gets

$$\int_0^\infty \exp\left\{-b\lambda^k x^l\right\} \;=\; \frac{\Gamma\left(\frac{1}{l}\right)}{l b^{1/l}} \cdot \lambda^{-k/l} \tag{9.3}$$

But, generally, (9.1) cannot be calculated analytically, in which case we derive its approximation in the spirit of de Bruijn [20]. The general approach is to show that $\int_\delta^\infty x^m \cdot e^{-f_\lambda(x)} dx$ is negligible for some $\delta > 0$ (δ can depend on λ). Then $\int_0^\delta x^m \cdot e^{-f_\lambda(x)} dx$ is approximated using the Taylor expansion of $f_\lambda(x)$ near the origin and formulae (9.2)-(9.3) above.

This technique is referred to in [20] as the *Laplace method* for the calculation of integrals. We now apply it to derive several asymptotic statements.

9.2 Asymptotic results

Lemma 9.1 Let b_1, k_1, l_1, l_2 be positive numbers and let b_2, k_2, m be non-negative. In addition, assume that l_1 and l_2 are integers. Consider a function $r_1 = \{r_1(\lambda), \quad \lambda > 0\}$ such that $r_1(\lambda) \sim \lambda^{k_1}, \quad \lambda \to \infty$. Finally, assume that

$$\frac{k_1}{l_1} \;>\; \frac{k_2}{l_2}\,. \tag{9.4}$$

Then

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx$$

$$= \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} \;+\; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right), \qquad \lambda \to \infty\,. \tag{9.5}$$

and

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx$$

$$= \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 [b_1 r_1(\lambda)]^{\frac{m+1}{l_1}}} \;-\; \frac{b_2 \Gamma\left(\frac{m+l_2+1}{l_1}\right)}{l_1 b_1^{\frac{m+l_2+1}{l_1}}} \cdot \lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}} \;+\; o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right). \tag{9.6}$$

Remark 9.1 Note that the main term in the right hand side of (9.5) is equal to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1\lambda^{k_1}x^{l_1}\right\} dx\,.$$

Thus, the relation (9.4) determines the "dominant" term in the exponent. Moreover, the second term in (9.6) is equal to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1\lambda^{k_1}x^{l_1}\right\} \cdot b_2\lambda^{k_2}x^{l_2} dx\,.$$

Therefore, Lemma 9.1 states, informally, that

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1} - b_2\lambda^{k_2}x^{l_2}\right\} dx \;\approx\; \int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1}\right\} \cdot [1 - b_2\lambda^{k_2}x^{l_2}] dx\,.$$

Remark 9.2 We can generalize (9.5) to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1} - \sum_{i=2}^n b_i\lambda^{k_i}x^{l_i}\right\} dx \;=\; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} + o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right), \qquad \lambda \to \infty\,,$$

as long as $\dfrac{k_1}{l_1} > \dfrac{k_i}{l_i}$ prevails for $2 \le i \le n$.

We shall also need the following slightly different version of Lemma 9.1.

Lemma 9.2 In addition to assumptions of Lemma 9.1, let $k_1 > k_2$ and assume that the function $r_2 = \{r_2(\lambda), \quad \lambda > 0\}$ satisfies $r_2(\lambda) = o(\lambda^{k_2}), \quad \lambda \to \infty$. Then

$$\begin{aligned}&\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1} - b_2 r_2(\lambda)x^{l_2}\right\} dx \\ &= \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} + o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right). \end{aligned} \tag{9.7}$$

Remark 9.3 Note that $r_2(\lambda)$ does not need to be positive, which is in contrast to the corresponding term λ^{k_2} in Lemma 9.1.

Lemma 9.3 Let $b, k, l, \delta > 0$, integer $m \ge 0$, and $-\infty < n < \infty$. Assume that the function $r(\lambda) \sim \lambda^k, \quad \lambda \to \infty$. Define a function

$$S(\lambda) \;\triangleq\; \int_{\delta\lambda^n}^\infty x^m \cdot \exp\left\{-br(\lambda)x^l\right\} dx\,, \qquad \lambda > 0\,,$$

and assume

$$nl + k > 0\,. \tag{9.8}$$

Then there exists $\nu > 0$ such that

$$S(\lambda) \;=\; o(e^{-\lambda^\nu})\,. \tag{9.9}$$

9.3 Proofs of Lemmata 9.1-9.3

Proof of Lemma 9.1.

Define

$$I \stackrel{\Delta}{=} \int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx$$

and

$$I_A \stackrel{\Delta}{=} \int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 \lambda^{k_2} x^{l_2}] dx \,.$$

Formula (9.2) and straightforward calculations imply

$$I_A = \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 [b_1 r_1(\lambda)]^{\frac{m+1}{l_1}}} - \frac{b_2 \Gamma\left(\frac{m+l_2+1}{l_1}\right)}{l_1 b_1^{\frac{m+l_2+1}{l_1}}} \cdot \lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}} + o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right).$$

Now

$$|I - I_A| = o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right) \tag{9.10}$$

will imply Lemma 9.1. If $x > 0$ and $\lambda^{k_2} x^{l_2} \leq 1$, then exists $C > 0$ such that

$$|\exp\{-b_2 \lambda^{k_2} x^{l_2}\} - (1 - b_2 \lambda^{k_2} x^{l_2})| \leq C \lambda^{2k_2} x^{2l_2} \,.$$

Define $\delta \stackrel{\Delta}{=} \lambda^{-k_2/l_2}$ and note that the condition $\lambda^{k_2} x^{l_2} \leq 1$ is equivalent to $x \leq \delta$. Now

$$\int_0^\delta x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot |\exp\{-b_2 \lambda^{k_2} x^{l_2}\} - (1 - b_2 \lambda^{k_2} x^{l_2})| dx$$

$$\leq C \cdot \int_0^\infty \lambda^{2k_2} x^{m+2l_2} \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} dx$$

$$= \frac{C \lambda^{2k_2}}{l_1 [b_1 r_1(\lambda)]^{\frac{m+2l_2+1}{l_1}}} \cdot \Gamma\left(\frac{m+2l_2+1}{l_1}\right) = O\left(\lambda^{2k_2 - \frac{k_1(m+2l_2+1)}{l_1}}\right) = o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right),$$

where the last equality follows from (9.4). In order to complete the proof, we show that the remainder $\int_\delta^\infty$ of the integrals can be ignored. Specifically, there exists $\nu > 0$ such that

$$\int_\delta^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx = o\left(e^{-\lambda^\nu}\right)$$

and

$$\int_\delta^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 \lambda^{k_2} x^{l_2}] dx = o\left(e^{-\lambda^\nu}\right).$$

The last two statements follow from Lemma 9.3. (Condition (9.8) applies due to (9.4).)
■

Proof of Lemma 9.2.

The proof is similar to the proof of Lemma 9.1. The integration domain is again divided by $\delta = \lambda^{-k_2/l_2}$. For large λ the inequality $x \le \delta$ implies $|x^{l_2} r_2(\lambda)| \le 1$, which, in turn, implies

$$|\exp\{-b_2 r_2(\lambda) x^{l_2}\} - (1 - b_2 r_2(\lambda) x^{l_2})| \;\le\; C[r_2(\lambda)]^2 x^{2l_2}$$

for some $C > 0$. Then one shows that

$$\left| \int_0^\delta x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} dx - \int_0^\delta x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \right|$$
$$= \; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right),$$

and

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \;=\; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} \;+\; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right).$$

The last step is to prove "exponential bounds":

$$\int_\delta^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} dx \;=\; o\left(e^{-\lambda^\nu}\right), \quad \nu > 0, \tag{9.11}$$

and

$$\int_\delta^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \;=\; o\left(e^{-\lambda^\nu}\right), \quad \nu > 0.$$

In order to get (9.11), the condition $k_1 > k_2$ is needed. It enables us to find $0 < C_1 < 1$, such that for $x > \delta$ and λ large enough,

$$\exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} \;<\; \exp\left\{-b_1 C_1 r_1(\lambda) x^{l_1}\right\},$$

and

$$\exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] \;<\; \exp\left\{-b_1 C_1 r_1(\lambda) x^{l_1}\right\},$$

Now we can apply Lemma 9.3. (Its proof appears below.) ■

Proof of Lemma 9.3.

We perform a change of variables

$$z = br(\lambda) x^l, \quad x = \left(\frac{z}{br(\lambda)}\right)^{1/l}, \quad dx = \frac{dz}{br(\lambda)} \left(\frac{z}{br(\lambda)}\right)^{1/l-1},$$

getting

$$S(\lambda) \;=\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \int_{C_1 r(\lambda)\lambda^{nl}}^{\infty} e^{-z} z^{\frac{m+1}{l}-1} dz\,, \tag{9.12}$$

where C_1 and C_2 are positive constants. Under condition (9.8), the lower bound $C_1 r(\lambda)\lambda^{nl}$ of the integral in (9.12) converges to infinity. Therefore, there exists $\alpha > 0$ such that for λ large enough,

$$S(\lambda) \;\leq\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \int_{C_1 r(\lambda)\lambda^{nl}}^{\infty} e^{-\alpha z} dz \;=\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \exp\{-C_3 r(\lambda)\lambda^{nl}\}\,,$$

where C_3 is a positive constant. Since $r(\lambda)\lambda^{nl} \sim \lambda^{nl+k}$, we can easily find $\nu > 0$ such that (9.9) is satisfied. ■

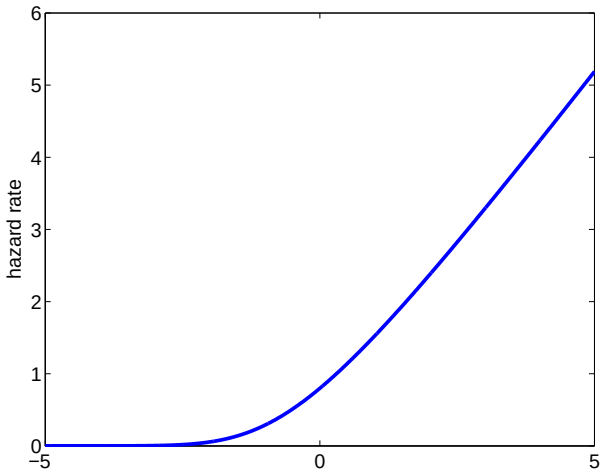
10 Some properties of the normal hazard-rate

In Part IV we shall extensively use the *hazard rate* function of the standard normal distribution

$$h(x) \;=\; \frac{\phi(x)}{1-\Phi(x)} \;=\; \frac{\phi(x)}{\bar{\Phi}(x)}\,, \tag{10.1}$$

where $\Phi(x)$ is its cumulative distribution function, $\bar{\Phi}(x) = 1-\Phi(x)$ is the survival function and $\phi(x) = \Phi'(x)$ is the density.

Figure 17: **Normal hazard rate**



The derivative of the normal hazard rate is equal to

$$h^{'}(x) = h(x) \cdot (h(x) - x) \tag{10.2}$$

and, consequently,

$$h(x) - x = \left[\ln h(x)\right]^{'}.$$

The second derivative is

$$h^{''}(x) = h(x) \cdot (2h^2(x) - 3xh(x) + x^2 - 1)\,. \tag{10.3}$$

Theorem 1.3 from Durrett [21] states that

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\phi(x) \leq \bar{\Phi}(x) \leq \frac{1}{x}\phi(x)\,, \quad \text{for } x > 0.$$

Then it follows that:

$$h(x) \geq x, \qquad -\infty \leq x \leq \infty\,,$$
$$h(x) \leq \frac{x^3}{x^2 - 1}, \qquad x > 1,$$

and

$$|h(x) - x| \to 0\,, \quad \text{as } x \to \infty\,. \tag{10.4}$$

It is well-known that h is an increasing function (see Gupta and Gupta [31] for a general treatment of multivariate normal case). Surprisingly, we have not found anywhere a proof that h is convex and we shall need this fact. So we constructed an indirect proof, based on the convexity of the Erlang-B formula [40], in the following way. Define the function

$$B(s,a) \triangleq \left[a\int_0^\infty e^{-at}(1+t)^s dt\right]^{-1}.$$

For $a > 0$ and integer $s > 0$ it can be shown that

$$B(s,a) = \left[\sum_{i=0}^{s}\frac{a^i}{i!}\right]^{-1} \cdot \frac{a^s}{s!}\,.$$

The last expression is equal to the Erlang-B blocking probability in the M/M/s/s system with $a = \dfrac{\lambda}{\mu}$. It has been proved in [40] that $B(s,a)$ is convex in s in $[0,\infty)$, for all $a > 0$. Now define

$$\tilde{B}(\beta,a) \triangleq \sqrt{a} \cdot B(a + \beta\sqrt{a}, a)\,.$$

Obviously $\tilde{B}(\beta, a)$ is also convex in β over $(-\sqrt{a}, \infty)$. The QED result for the Erlang-B system, derived by Jagerman [39], implies that

$$\tilde{B}(\beta, a) \to h(-\beta) \qquad (a \to \infty)\,.$$

The pointwise limit of a sequence of convex functions is convex as well, implying that h is convex.

Finally, formula (10.4) and the convexity of h imply

$$h'(x) < 1\,, \qquad -\infty < x < \infty\,. \tag{10.5}$$

Part III
Impact of customers' patience on delay and abandonment

11 Some new theoretical results

11.1 Patience-induced order relations for performance measures

We consider the following problem. Fix the parameters λ, μ, n of the M/M/n+G queue, so that only the patience distribution G can be varied. Is it possible to derive any relation for different performance measures of the M/M/n+G, based on some order relation between patience distributions? Is there any distribution that brings those performance measures to their maximum (minimum)?

These problems are practically important, for example, if the patience (maximal waiting time) depends not only on the customer but on the system management as well. Consider, for example, *overflows* when a customer that has already been waiting in queue is rerouted to some alternative resource: a free agent, another queue or a VRU. (The latter option usually leads to customers' dissatisfaction, but, nevertheless, it is practiced.) Protocols in real-time communication systems (e.g. the Internet) also rely on time bounds for the maximal wait in queues.

The following two statements provide some answers to the questions formulated above.

Lemma 11.1 Consider an M/M/n+G queue with fixed parameters λ, μ, n. Assume that for two patience-time distributions G_1 and G_2, the inequality

$$\int_0^x \bar{G}_1(\eta)d\eta \;\geq\; \int_0^x \bar{G}_2(\eta)d\eta \tag{11.6}$$

prevails for all $x > 0$, where $\bar{G}_1$ and $\bar{G}_2$ are the corresponding survival functions. Let $\mathrm{P}^i\{\mathrm{Ab}\}$, $\mathrm{P}^i\{\mathrm{Ab}|V>0\}$, $\mathrm{P}^i\{V>0\}$ and $\mathrm{P}^i\{W>0\}$, $i=1,2$, denote the steady-state characteristics of the corresponding M/M/n+G$_i$ queue. Then,

a. $\mathrm{P}^1\{V>0\} \;\geq\; \mathrm{P}^2\{V>0\}$; $\mathrm{P}^1\{W>0\} \;\geq\; \mathrm{P}^2\{W>0\}$.
b. $\mathrm{P}^1\{\mathrm{Ab}\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}\}$; $\mathrm{P}^1\{\mathrm{Ab}|V>0\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}|V>0\}$.

Theorem 11.1 Consider the M/M/n+G queue with fixed parameters λ, μ, n and a *fixed* average patience time $\bar{\tau}$. Then the deterministic distribution of patience G_d (every customer is willing to wait exactly $\bar{\tau}$) has the following "extremal" properties among all patience-time distributions with average $\bar{\tau}$:

a. The deterministic distribution maximizes the steady-state probabilities of wait $\mathrm{P}\{W > 0\}$ and $\mathrm{P}\{V > 0\}$.
b. The deterministic distribution minimizes the steady-state probabilities to abandon $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{P}\{\mathrm{Ab}|V > 0\}$.
c. The deterministic distribution maximizes the steady-state average wait $\mathrm{E}[W]$.
d. The deterministic distribution maximizes the steady-state average queue length $\mathrm{E}[Q]$.

Remarks.

1. It will be shown that Statements **a** and **b** of Theorem 11.1 are corollaries of Statements **a** and **b** from Lemma 11.1, respectively. However, inequality (11.6) does not imply order relations for average wait or average queue. An example is provided in Section 12 (see comments adjacent to Figure 25).

2. Assume that the patience-time distribution G_1 is *stochastically larger* than G_2: $\bar{G}_1(x) \geq \bar{G}_2(x), \quad x \geq 0$. Then condition (11.6) prevails automatically. Bhattacharya and Ephremides [6] proved that the former conventional stochastic order implies the corresponding inequality between the probabilities to abandon even in the general G/G/n+G case (non-Poisson arrivals, non-exponential service).

3. One can check that for two distribution G_1 and G_2 with the same mean, condition (11.6) is equivalent to $G_1 \leq_{cx} G_2$ in the sense of *stochastic convex order*. See Shaked and Shanthikumar [64] for details.

The proofs of Lemma 11.1 and Theorem 11.1 are given in Subsection 14.1.

11.2 Light-traffic results

Now we fix the parameters μ, n and the patience distribution G and derive several asymptotic formulae for small λ. One of the goals is to identify the slope of "probability to abandon versus average wait" near zero. This slope sometimes remains stable for light to

moderate (or even large) loads.

Lemma 11.2 Consider M/M/n+G queues with all parameters, except the arrival rate, being fixed. Assume that the arrival rate $\lambda \to 0$. (Below we index steady-state performance measures by a subscript λ.) Then

$$\lim_{\lambda\to 0} \frac{\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[W]} \;=\; \alpha_1 \;\triangleq\; \frac{1}{\int_0^\infty \bar{G}(x)e^{-n\mu x}dx} - n\mu\,. \tag{11.7}$$

The meaning of α_1 is the abandonment rate given one customer in the queue. In addition,

$$\lim_{\lambda\to 0} \mathrm{P}_\lambda\{\mathrm{Ab}|W>0\} \;=\; 1 - n\mu\int_0^\infty \bar{G}(x)e^{-n\mu x}dx \;=\; \mathrm{P}\{\tau < \exp(n\mu)\}\,, \tag{11.8}$$

$$\lim_{\lambda\to 0} \mathrm{E}_\lambda[W|W>0] \;=\; \int_0^\infty \bar{G}(x)e^{-n\mu x}dx \;=\; \mathrm{E}[\tau \wedge \exp(n\mu)]\,, \tag{11.9}$$

where the patience τ is independent of the $\exp(n\mu)$ random variable. See Subsection 14.2 for proofs.

Remark. Formulae (11.8) and (11.9) can be explained intuitively. Consider a lightly loaded M/M/n+G queue and assume that a customer encounters wait. Since the arrival load is small, it is highly probable that this customer is the only one in queue. Then the offered wait is $\exp(n\mu)$ distributed, which implies the relations above.

12 Empirically-driven experiments

We consider M/M/n+G queues with service rate $\mu = 1$ (minutes will be used as time units, for concreteness) and $n = 10$ agents. Several patience distributions were studied, most of which had an average patience $\bar{\tau} = 2$. We varied the arrival rate λ from 1 to 50, in step 0.25, then calculated performance measures and summarized the results graphically. A Matlab program, based on Brandt and Brandt [11], was used for calculations. Here we present a sample of examples that are related to the following topics:

- The relation between the probability to abandon and average wait, in particular how close it is to being linear.
- Explanations of linearity or non-linearity of the above relation.

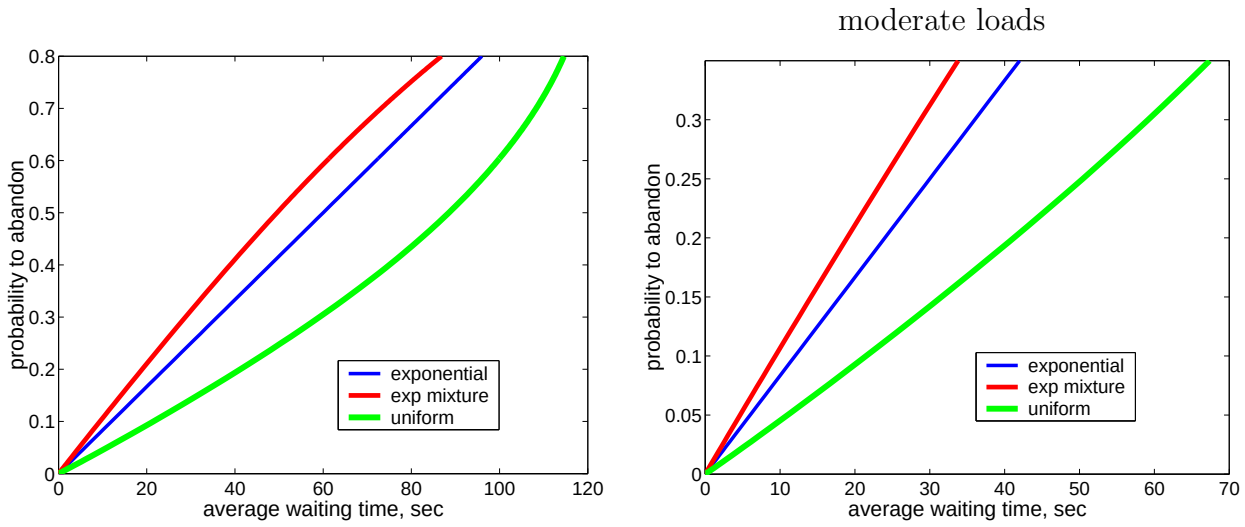
- Checking some theoretical results for M/M/n+G, exact and asymptotic.
- Exploring the relation between performance measures and the arrival rate, for various patience distributions.

Remark 12.1 Below we perform numerical experiments with $n = 10$. However, our conclusions seem to be correct for larger values of n. (See, for example, Figure 6 from Introduction with $n = 100$.)

12.1 Examples of a linear relation between P{Ab} and E[W].

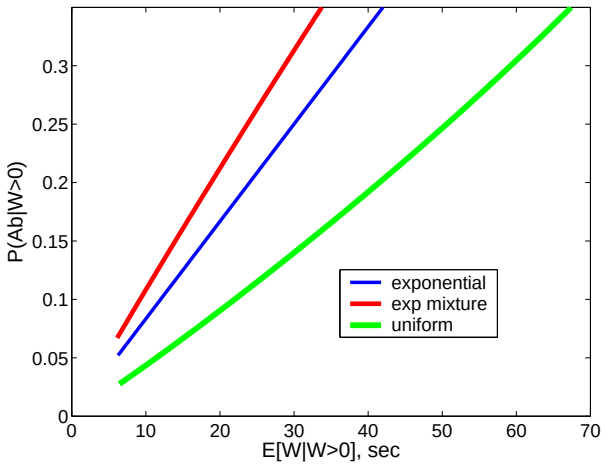
Example 1. We start with comparing the following three patience distributions: exponential with mean 2 minutes, uniform on [0,4] and hyperexponential (50-50% mixture of two independent exponentials with means 1 and 3 minutes). The first plot of Figure 18 depicts the corresponding relations between the probability to abandon and average wait, as λ varies from 1 to 50. The second plot shows the same relation, restricted to loads that are not extremely high (probability to abandon less than 35%). Finally, Figure 19 presents the same performance measures but conditioned on positive wait.

Figure 18: **Probability to abandon vs. average wait**



The first plot of Figure 18 illustrates the general form of P{Ab} / E[W] curves, given λ that varies from zero to infinity. Those curves always connect the origin and the point $(\bar{\tau}, 1)$ ($\bar{\tau} = 2$ minutes, or 120 seconds, in our case). The reason is that the average wait converges to the average patience, as $\lambda \to \infty$.

Figure 19: **Probability to abandon vs. average wait: delayed customers**



From relation (6.14) we know that exponential patience implies a linear curve, which is supported by Figures 18-19. The curves for other distributions need not be linear. For example, we observe (the first plot of Figure 18) the convex curve for the uniform distribution and the concave curve for the exponential mixture.

However, if we consider "reasonable loads" (the second plot of Figure 18) the two non-exponential curves are strikingly close to linear patterns. The same phenomenon is observed for conditional values (Figure 19), as well as for uniform distributions with different average in Appendix.

Finally, we check the light-traffic formulae from Lemma 2. The abandonment rates α_1, given one customer in the queue, are equal to 0.5, 0.2565 and 0.6563 for exponential, uniform and mixed exponential distributions, respectively. (These values are very close to 0.5, 0.25 and 2/3 – the patience densities at zero: the Remark below Theorem 2 explains this phenomenon.) The ratio between P{Ab} and $E[W]$ for the smallest arrival rate ($\lambda = 3$) is equal to (0.5, 0.2589, 0.6533), which conforms to Lemma 2. Checking formula

(11.8), for the light-traffic limit of the conditional probability to abandon, we get the vector (0.0476, 0.0250, 0.0616), which is highly plausible in view of Figure 19.

Example 2. We again consider three patience distributions:

- Delayed exponential distribution: all customers are willing to wait at least 0.25 (15 seconds); then their patience is governed by an exp(mean = 1.75) distribution.
- Exponential distribution with balking: 10% of the customers balk (leave immediately) if they encounter queue; the rest 90% of the customers are equipped with an exponential patience (mean = $\frac{20}{9}$), so that the overall mean equals 2.
- The survival curve for regular customers, based on the call center data from Figure 2, has been used in order to produce the third patience time distribution. We refer to it below as "cc data". The following operations have been performed with the data:

    1. The data has been normalized to 2, which is the average patience for other theoretical distributions in this section.
    2. Exponential smoothing was performed for the tail of the distribution (for time values larger than 1.2, which is equivalent to 6 minutes in the initial scale). The reason is that estimates of survival functions are very unreliable for large time values (the data is heavily censored, see [13]). Note that the linear pattern of the "cc data" curve for very large loads (in the first plot of Figure 20) is a consequence of the exponential smoothing.

Figures 20-21 were plotted using the same algorithm as in Figures 18-19.

We observe that the cc data curve for moderate loads (the second plot of Figure 20) is close to a linear pattern. However, it is noticeably concave. In fact, the concave pattern for small loads is plausible due to the first peak of the hazard rate in the second plot of Figure 5 in Section 2: a fraction of the customers abandons almost immediately if a positive wait is encountered. We also observe that the nearly perfect linear pattern of Figure 7 is in fact somewhat concave near zero.

The graphs of the two theoretical distributions in Figure 21 can be used, to some extent, for validation of Figure 8 from Subsection 4.1. First, Figure 21 demonstrates the

theoretically predicted linear curve with y-positive intercept for the exponential distribution with balking. (Recall the second plot in Figure 8.) The conditional curve of the delayed exponential distribution in Figure 21 is close to a straight line as well, which provides analogy with the first plot of Figure 8. (Note that in both cases, all customers are delayed, either in VRU, or in queue.)

Figure 20: **Probability to abandon vs. average wait**

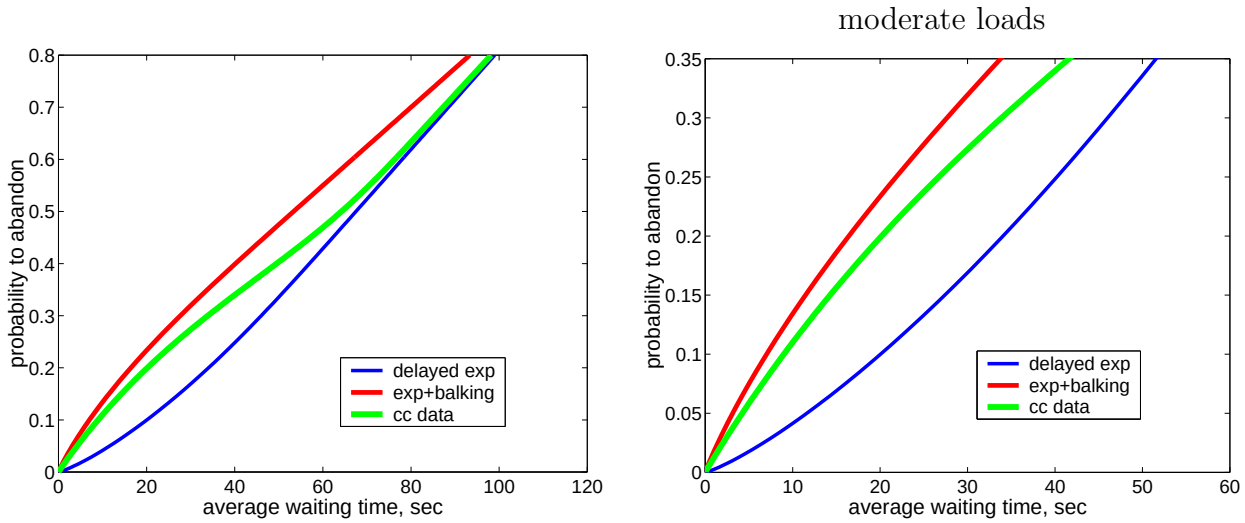
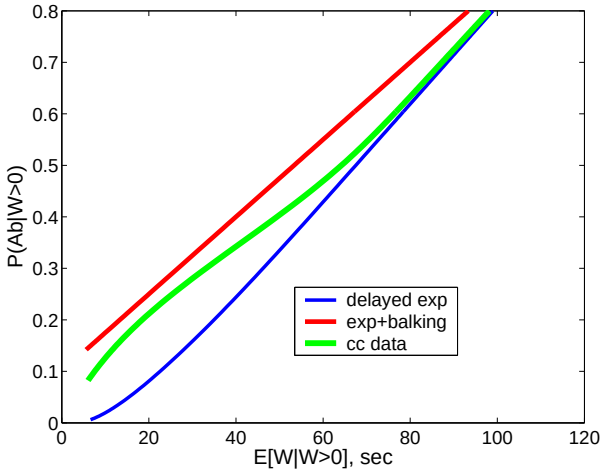


Figure 21: **Probability to abandon vs. average wait: delayed customers**



It was observed in [13] that the service-time distribution in our call center is very close to lognormal. Therefore, it is natural to compare between M/M/n+G results, illustrated by Figures 20 and 21, and M/G/n+G results with lognormal service distribution. Since theoretical results on the M/G/n+G queueing system are not available, we resort to simulation.

We simulated the M/G/n+G queue with $n = 10$, the "cc data" patience used above and lognormal service times with mean 1 and coefficient of variation equal to 1.2 (approximately the same as in our call center data). The arrival rate was varied from 3 to 15. Figure 22 demonstrates that the performance measures of the two systems are indeed very similar. The first plot shows that the probabilities to abandon are almost indistinguishable. (This fact conforms to the conclusion of Boxma and de Waal [10] that observed only mild sensitivity of the probability to abandon with respect to the service-time distribution.) In the second plot, we observe a small difference in waiting time. As a result, the lognormal P{Ab} / E[W] curve in the third plot is close to the real-data straight-line pattern of Figure 7 from Subsection 4.1.

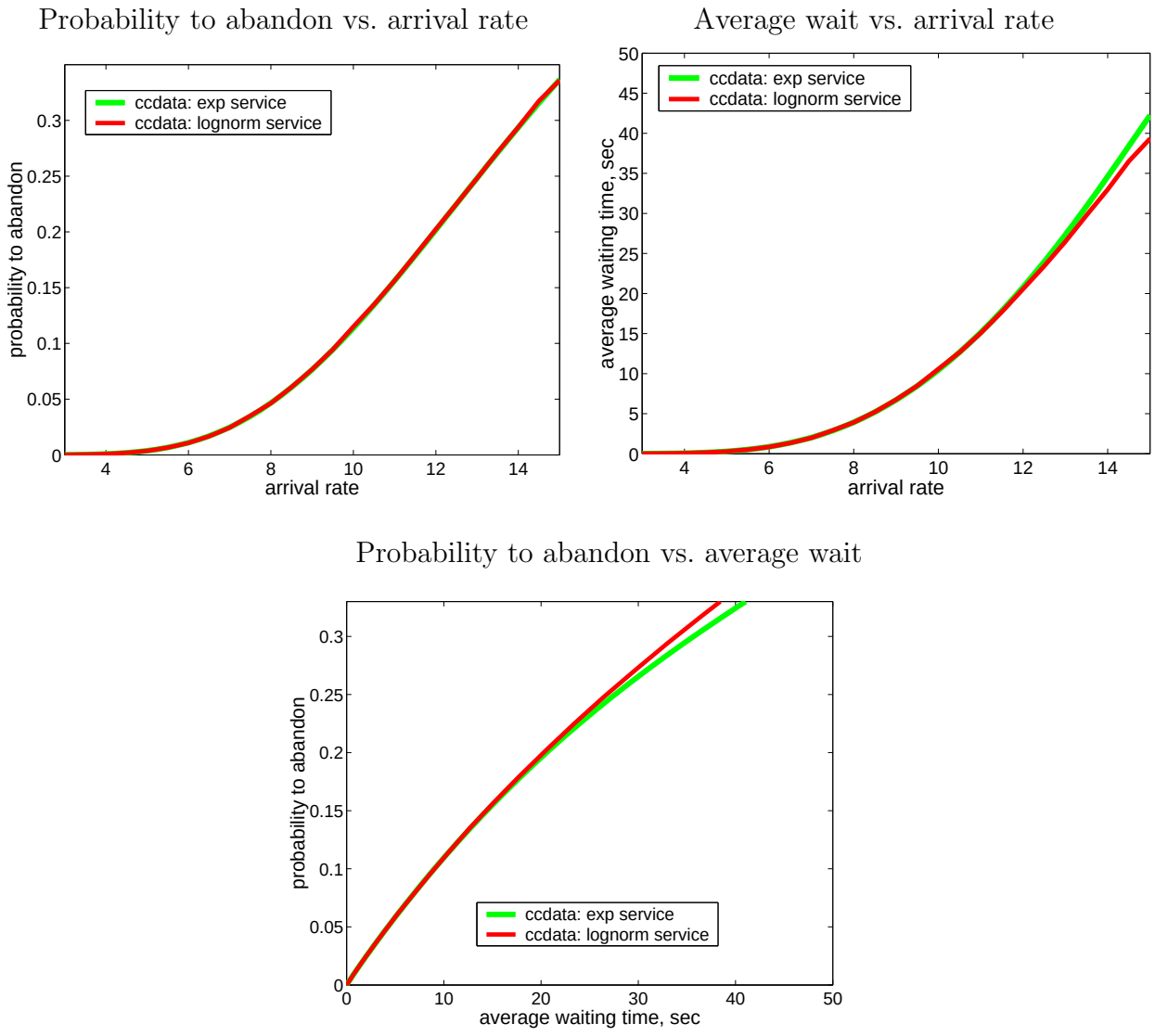
Summarizing, the patterns of Figure 7 and our "cc data" curves are similar. The observed difference can be attributed to various discrepancies between the call center environment and the M/M/n+G model: the number of servers is not constant over the day, the system does not always enter steady-state, priorities and skill-based routing, etc. Yet we believe that simple queueing models, such as Erlang-A (see Chapter 8 of Brown et al. [13]) or M/M/n+G, could turn out very useful in the analysis of complex call centers.

12.2 Examples of a strictly non-linear relation between P{Ab} and E[W].

Example 3. Here we present four patience distributions that give rise to non-linear patterns of dependence between the probability to abandon and average wait:

- Deterministic distribution: all customers are willing to wait *exactly* 2 minutes.
- Erlang (Gamma) distribution with two exponential phases, each with the mean equal to one minute.
- Lognormal distribution with both average and standard deviation equal to 2.

Figure 22: **Comparing M/M/n+G and M/G/n+G (lognormal service)**

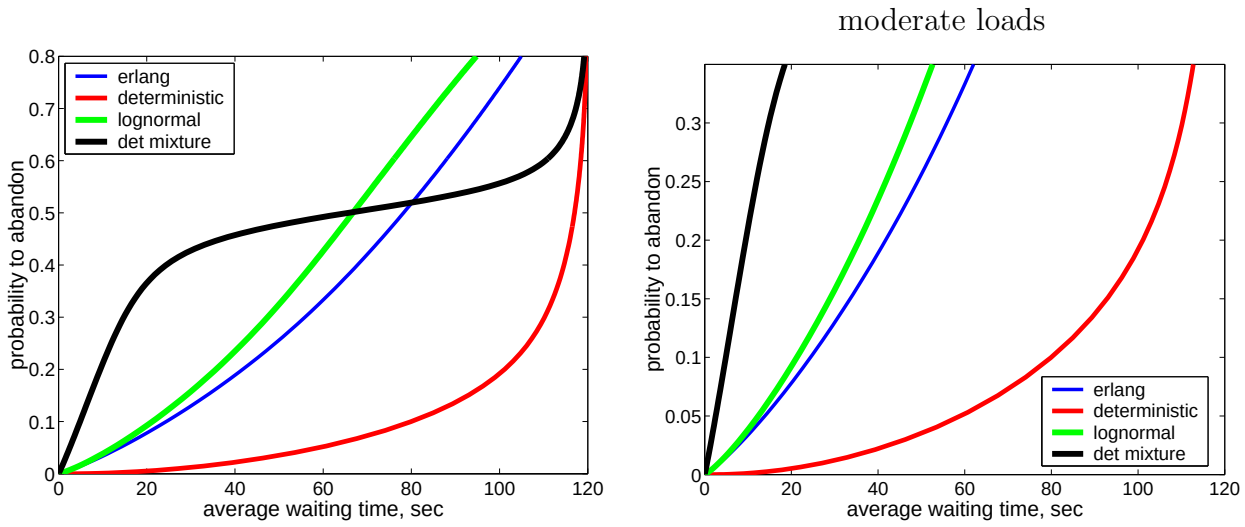


- The fourth distribution is a 50-50% mixture of two constants: 0.2 and 3.8.

Note that the densities of all the above distributions vanish at the origin. In general, the theoretical cases with strictly non-linear relations between P{Ab} and E[W] are usually characterized by patience density that, at the origin, either vanishes or exhibits some "unstable" behavior.

In our example, illustrated by Figure 23, the deterministic curve is strictly convex and lies below all other plots, as could be expected from Theorem 11.1. The patterns

Figure 23: **Probability to abandon vs. average wait**



of Erlang and lognormal are similar. Finally, a deterministic mixture provides a peculiar curve, which starts concave, and turns into convex. This can be explained as follows. When the loads are light to moderate, the customers with short patience abandon. (The curve is almost linear in this range.) For larger loads, the probability to abandon remains almost constant while the average wait increases: indeed, all customers with short patience abandoned and those with long patience still prevail. Eventually, the long-patience customers start abandoning as well.

12.3 Abandonment rate as a function of queue length

In the framework of our experiments, all the M/M/n+G parameters, except for λ, have been fixed. Therefore, from Little's formula and (6.48),

$$\frac{\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[W]} \;=\; \frac{\lambda\cdot\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[Q]} \;=\; \frac{\sum_{l=1}^{\infty}\alpha_l\cdot\pi_{n+l}(\lambda)}{\sum_{l=1}^{\infty} l\cdot\pi_{n+l}(\lambda)}\,, \tag{12.10}$$

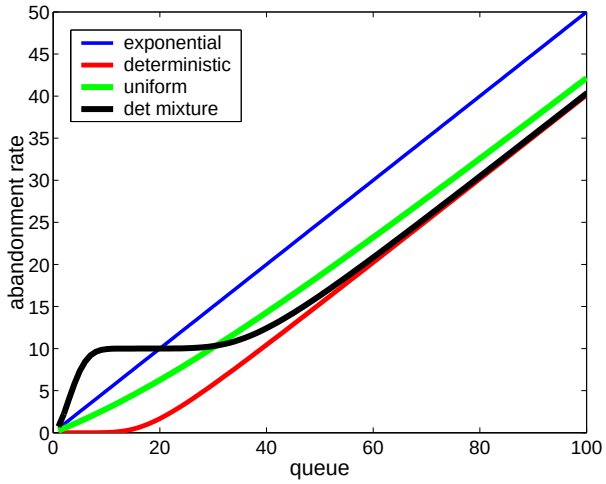
where α_l are the abandonment rates given l customers in the queue. For example, in the case of exp(θ) patience:

$$\alpha_l = \theta\cdot l \quad \text{and} \quad \frac{\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[W]} \equiv \theta\,.$$

Figure 24 supports the claim that the expression in (12.10) is approximately linear with respect to λ, if the abandonment rates α_l are approximately linear with respect to the queue length l.

We plotted four curves of abandonment rates in Figure 24, using the patience distributions from Examples 1 and 3. One observes a clear connection between the curves from these examples (Figures 18 and 23) and Figure 24: the exponential curve is exactly linear, the uniform curve is close to linear and the deterministic curve is strictly convex. (In the deterministic case, there is almost no abandonment if the queue is small.)

Figure 24: **Abandonment rate given queue**



The abandonment plot for the deterministic mixture also conforms to the corresponding curve at Figure 23: linear increase for small queues, then a "plateau" (all short-patience customers abandon) and, finally, linear increase again.

Figure 24 provides the opportunity to verify our equations (6.49) and (6.50), both taken from Brandt and Brandt [12]. We observe that the four curves share the same slope $1/\bar{\tau} = 0.5$, for large values of l.

12.4 Dependence of E[W] and P{Ab} on varying arrival rates

Figure 25 shows the graph of E[W], for M/M/n+G with three patience distributions. We are already familiar with the deterministic and uniform distributions from our previous examples.

The third patience distribution is the following: 25% of the customers balk immediately if they encounter a queue, 25% are willing to wait exactly 2 min and 50% have a uniform patience on [2,4]. (Hence the average patience is 2 minutes.)

Figure 25 illustrates two facts. First, as predicted by Theorem 11.1, the deterministic curve is maximal. Second, note that the other two curves cannot be ordered uniformly in λ. However, the uniform distribution G_1 is larger than its mixture counterpart G_2 in the sense of the order relation (11.6) from Lemma 11.1. Therefore, this relation does *not* imply any order for average waits.

Figure 25: **Average wait vs. arrival rate**

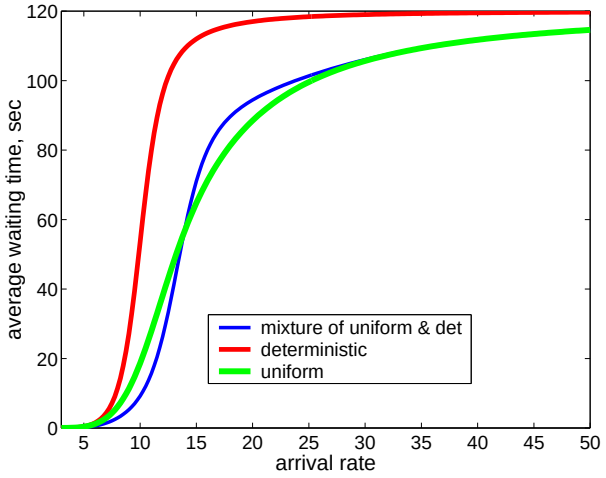
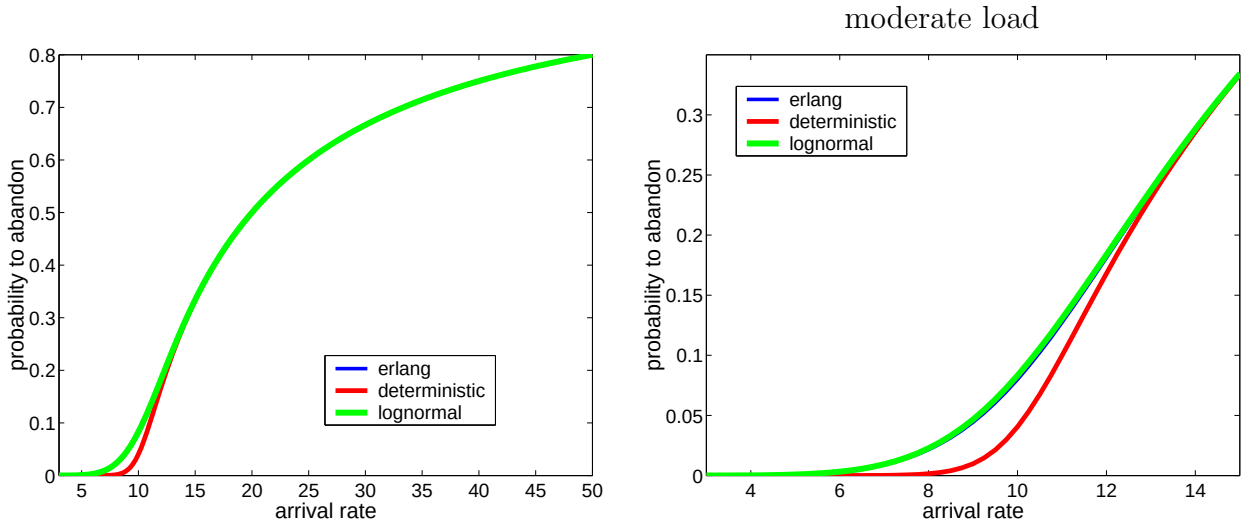


Figure 26: **Probability to abandon vs. arrival rate**



Finally, we present in Figure 26 abandonment versus arrival rates for the three distributions: deterministic, Erlang and lognormal (see Example 3). Note that the probabilities to abandon are rapidly reaching the "fluid limit" $1-(1\bigvee\rho)^{-1}$ (see [29]), where ρ is the offered load per server. (For example, if $\lambda=50$, then $\rho=5$ and $\mathrm{P}\{\mathrm{Ab}\}\approx 0.8$.) Figure 26 shows that even for moderate loads, the abandonment curves of two different distributions (Erlang and lognormal) can be almost indistinguishable. Overall, it seems that variations in the patience distribution usually implies larger changes in the average wait and less significant changes in the probability to abandon. (The most significant changes for the probability to abandon occur around $\rho=1$; see also Boxma and de Waal [10].)

12.5 Quantitative verification of linearity: ratio and curvature

So far we have relied on a visual inspection to identify linearity (or non-linearity) of the relation between the probability to abandon and average wait. Two simple methods, illustrated by Figure 27, can be used to verify this quantitatively.

First, the ratio between the two performance measures can be plotted. The first plot in Figure 27 shows clearly the constant relation for the exponential distribution; this relation is only slightly increasing (especially if the load is not very high) for the uniform distribution and it is strictly non-constant for the deterministic distribution.

The second plot presents the curvature of the "P{Ab} vs. E[W]" graphs. (See [57], page 119, for the definition of curvature. Recall that zero curvature corresponds to a straight line.) Consult Figure 25 in order to verify that the curvature of the deterministic distribution increases steeply even for moderate loads and the curvature of the uniform distribution is high for only the very large loads.

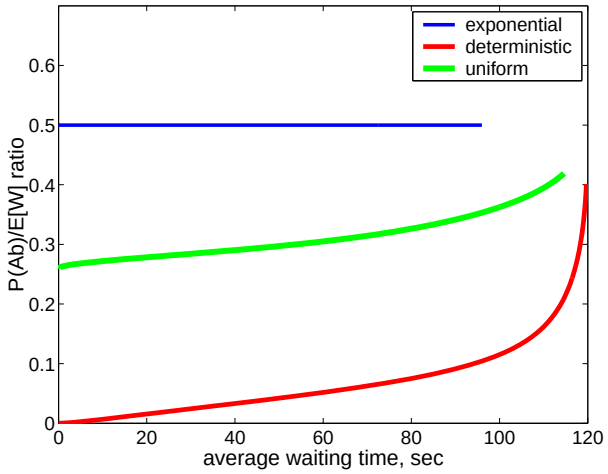
13 Conclusions of Part III

We have studied the impact of the patience distribution on M/M/n+G performance. In particular, some interesting extremal properties of the M/M/n+D queue were established.

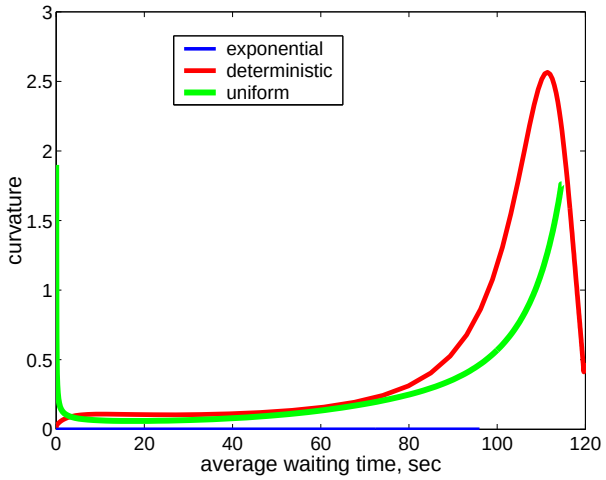
Concerning the specific phenomenon discussed in Subsection 4.1 (e.g. Figure 7), it has turned out that the linear relation between the probability to abandon and average wait prevails, both practically and theoretically, in a much broader context than the

Figure 27: **Two methods to evaluate linearity**

Ratio between probability to abandon and average wait



Curvature of P{Ab} vs. E[W]



Erlang-A model with exponential patience. To be more specific, an exact linear relation holds theoretically for exponential patience (Figure 18). It also holds practically for many patience distributions in the sense that, over realistic parameter values, the relation is close to being linear (second plot in Figure 18). In addition, such a relation was established in the light-traffic framework (Lemma 11.2). There are exceptions, however, as apparent for the deterministic case (Figure 23).

14 Proofs of theoretical results

In Subsection 6.3 we surveyed a relevant material from Baccelli and Hebuterne [3], which constitutes a necessary theoretical background for Lemma 11.1 and Theorem 11.1. The proofs of Lemma 11.1 and Theorem 11.1 are presented in Subsection 14.1. We proceed with the proof of Lemma 2 (Subsection 14.2) which uses the review of Brandt and Brandt [11, 12] in Subsection 6.4.

14.1 Proofs of Lemma 11.1 and Theorem 11.1

We start with the proof of Lemma 11.1, then derive its corollaries: **a** and **b** of Theorem 11.1 and, finally, proceed to **c** and **d** that require more complicated proofs.

Proof of Lemma 11.1.

a. Formula (6.28) and inequality (11.6) imply that $J_1 \geq J_2$ (where the values J_1 and J_2 correspond to distributions G_1 and G_2). From (6.78) one gets that $\mathrm{P}^i\{V > 0\}$ is an increasing function of J_i. (We assume that λ, μ and n are fixed; note that $\mathcal{E}$ from (6.78) does not depend on patience distribution.). Therefore, the relation

$$\mathrm{P}^1\{V > 0\} \;\geq\; \mathrm{P}^2\{V > 0\}$$

prevails.

Survival functions $\bar{G}_i$ are right-continuous. Therefore, inequality (11.6) implies $\bar{G}_1(0) \geq \bar{G}_2(0)$. Now

$$\mathrm{P}^1\{W > 0\} \;\geq\; \mathrm{P}^2\{W > 0\}$$

follows from (6.79).

b. A relation between the probability to abandon and utilization, namely

$$\rho \;=\; \frac{\lambda}{n\mu} \cdot (1 - \mathrm{P}\{\mathrm{Ab}\})$$

implies that the inequality for probabilities to abandon follows from the inequality for utilizations: $\rho_1 \geq \rho_2$. Utilizations can be calculated by

$$\rho_i \;=\; 1 - \sum_{j=0}^{n-1} \pi_j^i \left(1 - \frac{j}{n}\right), \quad i = 1, 2.$$

Formulae (6.25) and (6.27) imply that $\pi_j^1 \leq \pi_j^2$, $0 \leq j \leq n-1$. Therefore, the inequality for utilizations holds.

Finally, the inequality $\mathrm{P}^1\{\mathrm{Ab}|V > 0\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}|V > 0\}$ follows from the definition of conditional probability and the two previous statements.

Proof of Theorem 11.1.

a+b. Define a function

$$H(x) \;=\; \int_0^x \bar{G}(\eta) d\eta\,.$$

Lemma 11.1 implies that if the deterministic distribution G_d uniformly maximizes $H(x)$ over all possible survival functions with mean $\bar{\tau}$, then **a+b** holds. Note that $H(x)$ has the following properties:

$$\begin{cases} H(0) \;=\; 0, \\ H(\infty) \;=\; \bar{\tau}, \\ H(x) \;\text{ is non-decreasing}, \quad x \geq 0, \\ H^{'}(0) \;\leq\; 1\,, \\ H^{'}(x) \;\text{ is non-increasing}, \quad x \geq 0. \end{cases} \tag{14.1}$$

The uniformly maximal survival function that satisfies the constraints (14.1) is:

$$H(x) \;=\; \min(x, \bar{\tau}), \quad x \geq 0\,, \tag{14.2}$$

which corresponds to the survival function of the deterministic distribution: $\bar{G}(\eta) = I\{\eta < \bar{\tau}\}$.

c+d. The bulk of the proof is showing that deterministic patience maximizes $\mathrm{E}[W|V>0]$. Then the statement for the average wait will follow from Lemma 11.1, part **a**. In addition, Little's formula will imply **d**.

First, we present expression for the conditional average wait, which is convenient to analyze:

$$\mathrm{E}[W|V>0] \;\triangleq\; W_0(H) \;=\; \frac{\int_0^\infty H(t)\cdot\exp\{\lambda H(t) - n\mu t\}dt}{\int_0^\infty \exp\{\lambda H(t) - n\mu t\}dt}\,, \tag{14.3}$$

where

$$H(t) = \int_0^t \bar{G}(u)du\,. \tag{14.4}$$

Formula (14.3) follows from (6.78), (6.85) and definitions (6.71), (6.73).

Now we can formulate the optimization problem. Define by S the set of functions that satisfy constraints (14.1). We need to prove that

$$H^D(x) \;\triangleq\; \min(x, \bar{\tau}) \;=\; \arg\max_{H\in S} W_0(H)$$

where $W_0(H)$ is defined by (14.3). It can be verified that W_0 is continuous on S in the uniform metrics

$$\rho(H_1, H_2) \;=\; \max_{0\leq t\leq\infty} |H_1(t) - H_2(t)|\,.$$

(Check the numerator and the denominator of (14.3) separately.)

Introduce

$$H_b^D(x) \triangleq \min(x/b, \bar{\tau}), \quad b > 1, \tag{14.5}$$

which is the family of functions that, as can be easily checked using (14.4), corresponds to patience distributions that take values $b\bar{\tau}$ and 0 with probabilities $1/b$ and $(1 - 1/b)$ respectively. In other words, customers either balk immediately (if they encounter queue) or are ready to wait a deterministic time $b\bar{\tau}$.

The proof will proceed via 3 steps.

Step 1. For any $H \in S$ that does not belong to the class H_b^D, there exists $\tilde{H} \in S$ s.t. $W_0(\tilde{H}) > W_0(H)$.

Step 2. For any $b > 1$, $W_0(H^D) > W_0(H_b^D)$.

Steps 1 and 2 imply that H^D is the only "candidate" for being $\arg\max_{H \in S} W_0(H)$. So, naturally

Step 3. $H^D = \arg\max_{H \in S} W_0(H)$.

Proof of Step 1. First, we shall calculate the *variation* of W_0:

$$\delta W_0(H, \delta H) \triangleq \frac{\partial}{\partial \alpha} \left[W_0 \left(H(t) + \alpha \delta H(t) \right) \right]_{\alpha = 0} \tag{14.6}$$

The set S is convex:

$$H_1, H_2 \in S, \ 0 \leq \alpha \leq 1 \ \Rightarrow \ \alpha H_1 + (1 - \alpha) H_2 \in S \, .$$

Remark. Below we shall use the following fact that is a consequence of definition (14.6): Let $H \in S$. Assume, for some δH, that $\delta W_0(H, \delta H) > 0$ and $(H + \delta H) \in S$. Then there exists $\alpha \in (0, 1)$ such that $W_0(H + \alpha \delta H) > W_0(H)$. Since S is a convex set, also $(H + \alpha \delta H) \in S$, for all $\alpha \in (0, 1)$. ■

Straightforward calculations imply that

$$\begin{aligned} \delta W_0(H, \delta H) &= \frac{\partial}{\partial \alpha} \left[\frac{\int_0^\infty (H(t) + \alpha \delta H(t)) \cdot \exp\{\lambda (H(t) + \alpha \delta H(t)) - n\mu t\} dt}{\int_0^\infty \exp\{\lambda (H(t) + \alpha \delta H(t)) - n\mu t\} dt} \right]_{\alpha = 0} \\ &= \int_0^\infty \delta H(t) \cdot \exp\{\lambda H(t) - n\mu t\} \cdot r(t) dt \, , \end{aligned} \tag{14.7}$$

where

$$r(t) = \int_0^\infty (1 + \lambda H(t) - \lambda H(x)) \cdot \exp\{\lambda H(x) - n\mu x\} dx \tag{14.8}$$

Note that $r(t)$ is an increasing function (in fact, strictly increasing at points t such that $H(t) \neq \bar{\tau}$). In addition, $r(\infty) > 0$. Hence, two cases can be considered:

- The value $r(0) \geq 0$ and $r(t)$ is positive for $t \in (0, \infty)$;
- the value $r(0) < 0$ and there exists a unique t^* that solves the equation $r(t) = 0$.

In the first case, for any $H \not\equiv H^D$ take $H + \delta H \equiv H^D$. Since δH is non-negative, $\delta W_0(H, \delta H) > 0$. Then there exists $\alpha \in (0, 1)$ such that $W_0(H + \tilde{\alpha}\delta H) > W_0(H)$.
In the second case, choose

$$(H + \delta H_1)(t) = \begin{cases} (tH(t^*))/t^*, & 0 \leq t \leq t^* \\ H(t), & t > t^* \end{cases} \tag{14.9}$$

Clearly $(H + \delta H_1) \in S$. If H is linear on $[0, t^*]$, $(H + \delta H_1) \equiv H$. Otherwise, note that on $[0, t^*]$ both δH_1 and r are negative. Hence, according to (14.7), $\delta W_0(H, \delta H_1) > 0$ and W_0 cannot attain its maximum at H.

If $(H + \delta H_1) \equiv H$ on $[0, t^*]$, we define $h = H'(t^*-)$ and try

$$(H + \delta H_2)(t) = \begin{cases} H(t), & 0 \leq t \leq t^* \\ \min(H(t^*) + h \cdot (t - t^*), \bar{\tau}), & t > t^* \end{cases} \tag{14.10}$$

Again $(H + \delta H_2) \in S$ and if $\delta H_2 \not\equiv 0$, $\delta W_0(H, \delta H_2) > 0$.
Hence, the only functions that can possibly bring W_0 to a maximum are those with $\delta H_1 \equiv 0$ and $\delta H_2 \equiv 0$. However, such functions (linear until they reach $\bar{\tau}$) constitute exactly the class H_b^D. ■

Proof of Step 2. We must maximize the functional defined in (14.3) over the one-parameter family of the functions H_b^D, introduced in (14.5). The problem is solved by "brute force", proving that $\dfrac{\partial}{\partial b} W_0(H_b^D)$ is negative and, hence, the maximum is attained at $b = 1$. Integration using definitions (14.3) and (14.5) shows that, if $\lambda - n\mu b \neq 0$,

$$\frac{\partial}{\partial b} W_0(H_b^D) = \frac{\partial}{\partial b} \frac{\dfrac{1}{b}\displaystyle\int_0^{b\bar{\tau}} t \cdot \exp\left\{\frac{(\lambda - n\mu b)t}{b}\right\} dt + \int_{b\bar{\tau}}^{\infty} \bar{\tau} \cdot \exp\{\lambda\bar{\tau} - n\mu t\} dt}{\displaystyle\int_0^{b\bar{\tau}} \exp\left\{\frac{(\lambda - n\mu b)t}{b}\right\} dt + \int_{b\bar{\tau}}^{\infty} \exp\{\lambda\bar{\tau} - n\mu t\} dt}$$

$$= \frac{\partial}{\partial b} \frac{\left[\frac{b\bar{\tau}}{\lambda - n\mu b} \cdot e^{\bar{\tau}(\lambda - n\mu b)} - \frac{b}{(\lambda - n\mu b)^2} \cdot e^{\bar{\tau}(\lambda - n\mu b)} + \frac{b}{(\lambda - n\mu b)^2}\right] + \frac{\bar{\tau}}{n\mu} e^{\bar{\tau}(\lambda - n\mu b)}}{\frac{b}{\lambda - n\mu b} \cdot [e^{\bar{\tau}(\lambda - n\mu b)} - 1] + \frac{e^{\bar{\tau}(\lambda - n\mu b)}}{n\mu}}$$

$$\triangleq \frac{\partial}{\partial b}\frac{f(b)}{g(b)},$$

where

$$f(b) = \left[\frac{\lambda\bar{\tau}}{(\lambda-n\mu b)n\mu} - \frac{b}{(\lambda-n\mu b)^2}\right]e^{\bar{\tau}(\lambda-n\mu b)} + \frac{b}{(\lambda-n\mu b)^2} \tag{14.11}$$

and

$$g(b) = \frac{1}{\lambda-n\mu b}\left[\frac{\lambda}{n\mu}e^{\bar{\tau}(\lambda-n\mu b)} - b\right]. \tag{14.12}$$

We shall need also the derivatives:

$$f^{'}(b) = e^{\bar{\tau}(\lambda-n\mu b)}\left[\frac{(\lambda+n\mu b)\bar{\tau}}{(\lambda-n\mu b)^2} - \frac{\lambda\bar{\tau}^2}{\lambda-n\mu b} - \frac{\lambda+n\mu b}{(\lambda-n\mu b)^3}\right] + \frac{\lambda+n\mu b}{(\lambda-n\mu b)^3}$$

and

$$g^{'}(b) = e^{\bar{\tau}(\lambda-n\mu b)}\left[\frac{\lambda}{(\lambda-n\mu b)^2} - \frac{\lambda\bar{\tau}}{\lambda-n\mu b}\right] - \frac{\lambda}{(\lambda-n\mu b)^2}.$$

One must show that

$$f(b)g^{'}(b) - f^{'}(b)g(b) \geq 0, \quad b \geq 1. \tag{14.13}$$

Then some algebra provides us with:

$$(\lambda-n\mu b)^2\cdot[f(b)g^{'}(b) - f^{'}(b)g(b)] =$$

$$\frac{\lambda^2}{n\mu(\lambda-n\mu b)^2}\cdot e^{2\bar{\tau}(\lambda-n\mu b)} - \left[b\lambda\bar{\tau}^2 + b\bar{\tau} + \frac{\lambda^2+n^2\mu^2b^2}{n\mu(\lambda-n\mu b)^2} + \frac{\lambda\bar{\tau}}{n\mu}\right]\cdot e^{\bar{\tau}(\lambda-n\mu b)} + \frac{n\mu b^2}{(\lambda-n\mu b)^2}$$

Multiply the last expression by $n\mu(\lambda-n\mu b)^2$ and denote $\tilde{\mu} = n\mu b$. We then get

$$\lambda^2 e^{2\bar{\tau}(\lambda-\tilde{\mu})} - [\lambda\tilde{\mu}\bar{\tau}^2(\lambda-\tilde{\mu})^2 + \tilde{\mu}\bar{\tau}(\lambda-\tilde{\mu})^2 + \lambda^2 + \tilde{\mu}^2 + \lambda\bar{\tau}(\lambda-\tilde{\mu})^2]\cdot e^{\bar{\tau}(\lambda-\tilde{\mu})} + \tilde{\mu}^2.$$

Now we change variables: $x = \lambda - \tilde{\mu}$ (note that $x > -\tilde{\mu}$) and transform the last expression to

$$(x+\tilde{\mu})^2 e^{2\bar{\tau}x} - [\tilde{\mu}(x+\tilde{\mu})^2\bar{\tau}^2x^2 + \tilde{\mu}\bar{\tau}x^2 + (x+\tilde{\mu})^2 + \tilde{\mu}^2 + (x+\tilde{\mu})\bar{\tau}x^2]\cdot e^{\bar{\tau}x} + \tilde{\mu}^2. \tag{14.14}$$

It is easy to check that (14.14) is zero for $x = 0$. Differentiating it with respect to x, we get

$$e^{\bar{\tau}x}(e^{\bar{\tau}x}\cdot r(x) - h_2(x)),$$

where

$$r(x) = 2\tilde{\mu} + 2\tilde{\mu}^2\bar{\tau} + (2+4\tilde{\mu}\bar{\tau})x + 2\bar{\tau}x^2$$

and

$$h_2(x) \;=\; 2\tilde{\mu} + 2\tilde{\mu}^2\bar{\tau} + (2 + 6\tilde{\mu}\bar{\tau} + 2\tilde{\mu}^2\bar{\tau}^2)x + (\tilde{\mu}^2\bar{\tau}^3 + 4\bar{\tau} + 5\tilde{\mu}\bar{\tau}^2)x^2 + (\tilde{\mu}\bar{\tau}^3 + \bar{\tau}^2)x^3\,.$$

Let $h_1(x) \stackrel{\Delta}{=} e^{\bar{\tau}x} \cdot r(x)$. Assume $\boldsymbol{x > 0}$. Then

$$e^{\bar{\tau}x} \;>\; 1 + \bar{\tau}x + \frac{\bar{\tau}^2x^2}{2}$$

and

$$h_1(x) \;>\; r(x) \cdot \left(1 + \bar{\tau}x + \frac{\bar{\tau}^2x^2}{2}\right) \;=\; h_2(x) + (\tilde{\mu}\bar{\tau}^3 + 2\bar{\tau}^2)x^3 + \bar{\tau}^3x^4 \;>\; h_2(x)\,. \quad (14.15)$$

If $\boldsymbol{x < 0}$, then

$$e^{\bar{\tau}x} \;<\; 1 + \bar{\tau}x + \frac{\bar{\tau}^2x^2}{2}$$

and

$$h_1(x) \;<\; h_2(x) + 2\bar{\tau}^2x^3 + \bar{\tau}^3x^3(\tilde{\mu} + x) < h_2(x)\,, \quad (14.16)$$

where the last inequality follows from $-\tilde{\mu} < x < 0$. According to (14.15) and (14.16), the derivative of expression (14.14) is negative for $x < 0$ and positive for $x > 0$. Hence (14.14) is non-negative, implying (14.13) and, in turn, Step 2.

The calculations above are not valid for $b = \dfrac{\lambda}{n\mu}$ (due to division by zero in (14.11) and (14.12)). However, that does not create any problem: W_0 is continuous in the uniform metric and, therefore, continuous in b over H_b^D. ■

Proof of Step 3. Define the subset $\tilde{S} \subseteq S$, where

$$H \in \tilde{S} \;\text{ iff }\; \exists\, T \;\text{ s.t. }\; H(T) = \bar{\tau}\,,$$

and introduce $\tilde{S}_T$ by

$$H \in \tilde{S}_T \;\text{ iff }\; H(T) = \bar{\tau}\,.$$

For all $T > 0$, the set $\tilde{S}_T$ is closed (in the uniform metric), uniformly bounded and uniformly equicontinuous (since a derivative of any function from S is bounded by 1). The Arzela-Ascoli theorem (see [47], for example) implies that $\tilde{S}_T$ is a compact set. Since W_0 is a continuous functional in the uniform metrics, it must attain a maximum on $\tilde{S}_T$.

If $H \in \tilde{S}_T$ and $H \notin H_b^D$ then, using the technique from Step 1, we can find $\tilde{H} \subseteq \tilde{S}_T$ such that $W_0(\tilde{H}) > W_0(H)$. (Note that the transformations (14.9) and (14.10), described

in Step 1, are $\tilde{S} \to \tilde{S}$.) Now Step 2 implies that H^D remains the only candidate to maximize W_0 on $\tilde{S}_T$, for all $T \geq \bar{\tau}$. Since T is arbitrary, H^D maximizes W_0 over $\tilde{S}$ as well.

Finally, note that any $H \in S$ can be approximated in the uniform metric by a sequence $\{H_n\} \in \tilde{S}$. Since W_0 is continuous, $W_0(H^D) \geq W_0(H)$. Moreover, $W_0(H^D) = W_0(H)$ is impossible for $H^D \not\equiv H$ since the value of the functional at H can always be improved by the methods from Step 1 and Step 2. ■

14.2 Proof of Lemma 11.2

We use the results from Brandt and Brandt [11, 12]. Note that if the patience time is finite almost surely, then the steady-state distribution exists for all values of λ, μ, n and the inverse of the normalization constant in (6.43) is finite. Hence, the sequence $\{F_j\}$ is bounded (in fact, converges to zero) and

$$g^{-1} \;=\; \sum_{j=0}^{n-1} \frac{n! \cdot \lambda^j \mu^{n-j}}{j!} + \sum_{j=0}^{\infty} \lambda^{n+j} F_j \;=\; n! \cdot \mu^n + o(\lambda)\,, \qquad (\lambda \to 0)\,.$$

Recall from (6.45) that $\pi_{n+l} = g \cdot \lambda^{n+l} F_l$, $l > 0$. Little's formula implies that the average waiting time is

$$\mathrm{E}[W] \;=\; \frac{\sum_{l=1}^{\infty} l \cdot \pi_{n+l}}{\lambda} \;=\; g\lambda^{n-1} \sum_{l=1}^{\infty} \lambda^l F_l \;=\; \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n F_1 + o(\lambda^n)\,. \tag{14.17}$$

Applying integration by parts:

$$F_1 \;\triangleq\; \int_0^{\infty} F(\xi) e^{-\xi} d\xi \;=\; \int_0^{\infty} \bar{G}(x) e^{-n\mu x} dx\,.$$

From (6.48) and (6.45), the probability to abandon is

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{\sum_{l=1}^{\infty} \alpha_l \cdot \pi_{n+l}}{\lambda} \;=\; \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \alpha_1 F_1 + o(\lambda^n)\,. \tag{14.18}$$

In order to validate the second equality of (14.18), note that from formula (6.49),

$$\alpha_l \;=\; \frac{l}{\mathrm{E}[\tau]} + o(l)\,, \qquad (l \to \infty)\,.$$

Hence, taking into account that

$$\sum_{l=2}^{\infty} l\lambda^l \;=\; \frac{\lambda}{(1-\lambda)^2} - \lambda \;\sim\; 2\lambda^2\,, \qquad (\lambda \to 0)\,,$$

note that

$$\sum_{l=2}^{\infty} \alpha_l \pi_{n+l} \;=\; g \cdot \lambda^n \sum_{l=2}^{\infty} \alpha_l \lambda^l F_l \;=\; o(\lambda^{n+1})\,, \qquad (\lambda \to 0)\,,$$

which implies (14.18). Now using formulae (14.17) and (14.18), we prove (11.7).

From the PASTA principle,

$$\mathrm{P}\{W > 0\} \;=\; \sum_{l=0}^{\infty} \pi_{n+l} \;=\; g\lambda^n + o(\lambda^n) \;=\; \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + o(\lambda^n)\,, \qquad (\lambda \to 0)\,.$$

Now we derive formula (11.8), using (14.18):

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{P}\{W > 0\}} \;\sim\; \alpha_1 F_1 \;=\; 1 - n\mu F_1 \;=\; \mathrm{P}\{\tau < \exp(n\mu)\}\,, \qquad (\lambda \to 0)\,,$$

and one gets (11.9) using (14.17)

$$\mathrm{E}[W|W > 0] \;=\; \frac{\mathrm{E}[W]}{\mathrm{P}\{W > 0\}} \;\sim\; F_1 \;=\; \mathrm{E}[\tau \wedge \exp(n\mu)] \qquad (\lambda \to 0)\,.$$

The last two equations imply

$$\lim_{\lambda \to 0} \frac{\mathrm{P}_\lambda\{\mathrm{Ab}|W > 0\}}{\mathrm{E}_\lambda[W|W > 0]} \;=\; \alpha_1\,.$$

Part IV
Asymptotic operational regimes in the M/M/n+G queue

15 QED operational regime

15.1 Formulation of results

15.1.1 Main case: patience distribution with a positive density at the origin

Consider an M/M/n+G queue. Fix the service rate μ and the patience distribution G. Assume that the arrival rate $\lambda \to \infty$ and the staffing level n is given by

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \qquad \lambda \to \infty\,, \quad -\infty < \beta < \infty. \tag{15.1}$$

We develop asymptotics of different performance measures, as $\lambda \to \infty$ ($n \to \infty$). The idea (adopted also in the following sections of Part IV) is, first, to derive asymptotic expressions for the building blocks J, $\mathcal{E}$ and J_1 that were defined in (6.71), (6.77) and (6.72), respectively. Then we continue with performance measures, using relevant formulae from the list (6.78)-(6.95).

Remark 15.1 All performance measures and queue characteristics in the statements of Part IV should be indexed by λ. As a rule, we omit this indexing. All asymptotic results are, by default, valid given $\lambda \to \infty$ (or, the same, $n \to \infty$).

Lemma 15.1 (Building blocks (6.71), (6.77), (6.72)) Define the patience-time density by $g = \{g(x), x \geq 0\}$. Assume that the density exists at the origin and its value $g(0) \triangleq g_0$ is strictly positive. Then, in the QED operating regime, namely $\lambda \to \infty$ and n as in (15.1), we have

a.

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \tag{15.2}$$

where

$$\hat{\beta} \;\triangleq\; \beta\sqrt{\frac{\mu}{g_0}}\,. \tag{15.3}$$

b.

$$\mathcal{E} \;=\; \sqrt{n}\cdot\frac{1}{h(-\beta)}+o(\sqrt{n})\,. \tag{15.4}$$

c.

$$J_1 \;=\; \frac{1}{n}\cdot\frac{1}{\mu g_0}\left[1-\frac{\hat{\beta}}{h(\hat{\beta})}\right]+o\left(\frac{1}{n}\right)\,. \tag{15.5}$$

d. Define

$$J_2 \;\triangleq\; \int_0^\infty x^2\cdot\exp\left\{\lambda\int_0^x G(u)du-n\mu x\right\}dx\,. \tag{15.6}$$

Then

$$J_2 \;=\; \frac{1}{n^{3/2}}\cdot\frac{1}{(\mu g_0)^{3/2}}\left[\frac{\hat{\beta}^2+1}{h(\hat{\beta})}-\hat{\beta}\right]+o\left(\frac{1}{n^{3/2}}\right)\,. \tag{15.7}$$

Theorem 15.1 (Performance measures) Under the assumptions of Lemma 15.1, the performance measures of the M/M/n+G queueing system in the QED regime can be approximated by:

a. The probability of wait converges to a constant that depends on β and $\frac{g_0}{\mu}$:

$$\mathrm{P}\{\mathrm{W}>0\} \;\sim\; \left[1+\sqrt{\frac{g_0}{\mu}}\cdot\frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}\,. \tag{15.8}$$

In addition, if $\lambda\to\infty$ and $\mathrm{P}\{\mathrm{W}>0\}\to\alpha$, with $0<\alpha<1$, then

$$n \;=\; \frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}+o(\sqrt{\lambda})\,, \tag{15.9}$$

where $\alpha=\left[1+\sqrt{\frac{g_0}{\mu}}\cdot\frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}$.

b. The probability-to-abandon of delayed customers decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1}{\sqrt{n}}\cdot\sqrt{\frac{g_0}{\mu}}\cdot\left[h(\hat{\beta})-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right)\,. \tag{15.10}$$

The probability to abandon P{Ab} also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (15.8) with (15.10).

c. The average offered wait of delayed customers decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{g_0\mu}}\cdot\left[h(\hat{\beta})-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right)\,. \tag{15.11}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\dfrac{1}{\sqrt{n}}$ and can be approximated by the product of (15.8) and (15.11).

d. The average waiting time is of the same order as the average offered wait:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]\,. \tag{15.12}$$

In addition,

$$\mathrm{P}[\mathrm{Ab} \mid W > 0] \;\sim\; \mathrm{P}[\mathrm{Ab} \mid V > 0]\,.$$

e. The ratio between probability to abandon and average wait converges to the (positive) value of patience density at the origin:

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}|W > 0\}}{\mathrm{E}[W|W > 0]} \;\sim\; g_0\,. \tag{15.13}$$

f. The average virtual offered wait of reneging customers decreases at rate $\dfrac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[\frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.14}$$

Assume, in addition, that the average patience time $\bar\tau = \mathrm{E}[\tau] < \infty$ and that the patience time has a continuous density at the origin. Then

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{2\sqrt{g_0\mu}} \left[\frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta\right] + o\left(\frac{1}{\sqrt{n}}\right), \tag{15.15}$$

or, in other words,

$$\mathrm{E}[W|\mathrm{Ab}] \;\sim\; \frac{1}{2} \cdot \mathrm{E}[V|\mathrm{Ab}] \qquad (n \to \infty)\,.$$

Moreover, the following inequality prevails:

$$\frac{1}{2} \cdot \left[\frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta\right] \;<\; h(\hat\beta) - \hat\beta \;<\; \frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta\,, \qquad -\infty < \hat\beta < \infty. \tag{15.16}$$

In conjunction with **c** and **d**, (15.16) implies corresponding asymptotic order relations between $\mathrm{E}[W|\mathrm{Ab}]$, $\mathrm{E}[W|W > 0]$, $\mathrm{E}[V|V > 0]$ and $\mathrm{E}[V|\mathrm{Ab}]$. In words, the average offered wait of *reneging* customers exceeds (asymptotically) the average waiting time of *delayed*

customers which, in turn, exceeds the average actual wait of reneging customers. (See also Remark 15.3 below).

g. The asymptotic distribution of wait, or *Total Service Factor(TSF)*, is given by the product of the right-hand side of (15.8) with

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \,\middle|\, W > 0 \right\} \;\sim\; \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{\frac{g_0}{\mu}} \cdot t\right)}{\bar{\Phi}(\hat{\beta})}\,, \qquad t \geq 0\,. \tag{15.17}$$

h. The probability to abandon, given delay in queue, is asymptotically equal to

$$\mathrm{P}\left\{ \mathrm{Ab} \,\middle|\, \frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \right\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{g_0}{\mu}}\right) - \hat{\beta} \right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.18}$$

i. The average wait, given delay in queue, is asymptotically equal to

$$\mathrm{E}\left[W \,\middle|\, \frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \right] \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{g_0\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{g_0}{\mu}}\right) - \hat{\beta} \right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.19}$$

Parts **h** and **i** together imply a generalization of part **e**:

$$\frac{\mathrm{P}\left\{ \mathrm{Ab} \,\middle|\, W > \frac{t}{\sqrt{n}} \right\}}{\mathrm{E}\left[W \,\middle|\, W > \frac{t}{\sqrt{n}} \right]} \;\sim\; g_0\,, \qquad t \geq 0. \tag{15.20}$$

Remark 15.2 The asymptotic statement (15.13) provides additional support for the practically observed linear relation between probability-to-abandon and average wait. (Recall Figure 7.)

Remark 15.3 Figure 28 illustrates the asymptotic relations

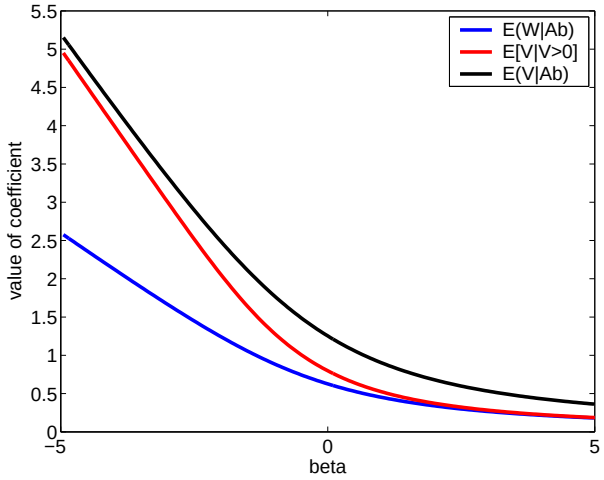
$$\mathrm{E}[W|\mathrm{Ab}] \;<\; \mathrm{E}[W|W>0] \;\approx\; \mathrm{E}[V|V>0] \;<\; \mathrm{E}[V|\mathrm{Ab}]\,,$$

or, formulating rigorously,

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[W|\mathrm{Ab}]}{\mathrm{E}_\lambda[W|W>0]} < 1; \qquad \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[W|W>0]}{\mathrm{E}_\lambda[V|V>0]} = 1; \qquad \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V|V>0]}{\mathrm{E}_\lambda[V|\mathrm{Ab}]} < 1.$$

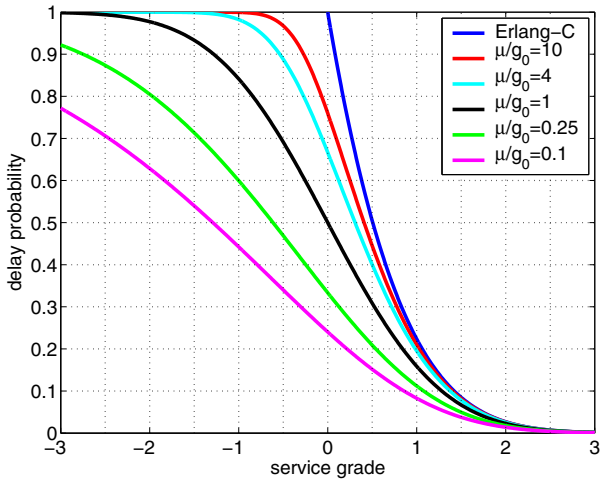
The three curves display the coefficients $\frac{1}{2}\cdot\left[\frac{1}{h(\hat{\beta})-\hat{\beta}} - \hat{\beta}\right]$, $h(\hat{\beta})-\hat{\beta}$ and $\frac{1}{h(\hat{\beta})-\hat{\beta}}-\hat{\beta}$, which correspond to $\mathrm{E}[W|\mathrm{Ab}]$, $\mathrm{E}[V|V>0]$ and $\mathrm{E}[V|\mathrm{Ab}]$, respectively.

Figure 28: **Comparison between wait formulae**



Remark 15.4 Figure 29 illustrates the dependence (15.8) between the service grade and probability the probability of wait, over varying values of the ratio μ/g_0. In addition, we plotted the curve (7.5) for the Erlang-C queue, which is meaningful for positive β only. Note that for large values of μ/g_0 (very patient customers) the Erlang-A curves are close to the Erlang-C curve.

Figure 29: **Asymptotic relations between service grade and delay probability**



Remark 15.5 The special case $\beta = 0$.

Note that when $\beta = 0$ in (15.1), the staffing level asymptotically corresponds to the simple rule that does not take into account stochastic considerations: assign the number of agents equal to the *offered load* $\frac{\lambda}{\mu}$. In Erlang-C, this "naive" approach would lead to system instability. However, in M/M/n+g (which is a much better fit to the real world of call centers than Erlang-C) one would get a reasonable-to-good performance level. Specifically,

$$\mathrm{P}\{W > 0\} \;\sim\; \frac{\sqrt{\frac{\mu}{g_0}}}{\sqrt{\frac{\mu}{g_0}}+1} \;=\; \frac{1}{1+\sqrt{g_0/\mu}}\,, \tag{15.21}$$

$$\mathrm{P}\{\mathrm{Ab}\} \;\sim\; \sqrt{\frac{2}{\pi n}}\cdot\frac{1}{1+\sqrt{\frac{\mu}{g_0}}}\,, \tag{15.22}$$

$$\mathrm{P}\left\{\mathrm{Ab} \mid W>0\right\} \;\sim\; \sqrt{\frac{2}{\pi n}}\cdot\sqrt{\frac{g_0}{\mu}}\,, \tag{15.23}$$

$$\mathrm{E}[W] \;\sim\; \sqrt{\frac{2}{\pi n}}\cdot\frac{1}{g_0+\sqrt{\mu g_0}}\,, \tag{15.24}$$

$$\mathrm{E}[W \mid W>0] \;\sim\; \sqrt{\frac{2}{\pi n}}\cdot\sqrt{\frac{1}{g_0\mu}}\,, \tag{15.25}$$

$$\mathrm{P}\left\{\frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\,\middle|\, W>0\right\} \;\sim\; 2\bar{\Phi}\left(\frac{1}{2}+t\sqrt{\frac{g_0}{\mu}}\right)\,, \qquad t\geq 0\,. \tag{15.26}$$

For example, if the service rate μ is equal to the individual abandonment rate θ, and $\beta = 0$, 50% of customers would get service immediately upon arrival. (Check it in Figure 29. Note that for Erlang-C, 50% delay probability corresponds to $\beta = 0.5$.) This suggests why some call centers that are managed using simplified deterministic models actually perform at reasonable service levels. (One obtains the "right answer" from the "wrong reasons".)

Note that performance measures (15.21)-(15.26) are functions of the ratio between the service rate μ and the patience density at zero g_0 (after dividing the average wait in (15.24)-(15.25) by the average service time).

Remark 15.6 Formula (15.8) generalizes the statement for the Erlang-A queue (exponential patience), derived in Garnett et al. [29]. Namely,

$$\mathrm{P}\{W>0\} \;\sim\; w\left(-\beta, \sqrt{\mu/\theta}\right)\,, \tag{15.27}$$

where

$$w(x,y) \;=\; \left[1+\frac{h(-xy)}{yh(x)}\right]^{-1}, \quad h(x) \;=\; \frac{\phi(x)}{1-\Phi(x)}$$

and θ is the abandonment rate (parameter of the exponential patience). Straightforward calculations reveal the equivalence between formulae (15.27) and (15.8), if we substitute g_0 instead of θ to (15.27). (Note that θ is indeed the density of $\exp(\theta)$ at the origin.)

Approximations for other performance measures (for example, the probability to abandon and average wait) were also derived in [29]. However, they do not coincide exactly with our approximations. The reason is that in Theorem 15.1 the lead asymptotic term is always presented explicitly with respect to n or λ. On the other hand, the approximation formulae in [29] do not display the lead term. For example, the [29] analogue to our formula (15.10) is as follows:

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;\approx\; \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta}+\sqrt{\theta/(n\mu)})}\,.$$

15.1.2 Patience distribution with density vanishing near the origin

Lemma 15.2 (Building blocks) Assume that the density of patience time at the origin $g_0=0$; that the first $(k-1)$ derivatives vanish as well: $g^{(i)}(0)=0,\ 1\le i\le k-1$, and that the k-th derivative is positive: $g^{(k)}(0)\stackrel{\Delta}{=}g_{0k}>0$.

For $\beta\neq 0$ (positive or negative) let the QED staffing level be

$$n=\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}+o(\sqrt{\lambda})\,. \tag{15.28}$$

If $\beta=0$ let

$$n=\frac{\lambda}{\mu}+o\left(\lambda^s\right)\,, \tag{15.29}$$

for some $s<\dfrac{1}{k+2}$.

The asymptotic expression for $\mathcal{E}$ coincides with (15.4) for all the theorems of Section 15. The approximations for J and J_1 are given by the following formulae:

a. If $\beta>0$

$$J \;=\; \frac{1}{n\mu-\lambda}-\frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+3}}+o\left(\frac{1}{\lambda^{(k+1)/2}}\right)\,, \tag{15.30}$$

$$J_1 \;=\; \frac{1}{(n\mu-\lambda)^2}-\frac{(k+3)\cdot\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+4}}+o\left(\frac{1}{\lambda^{(k+2)/2}}\right)\,. \tag{15.31}$$

b. If $\beta = 0$

$$J = \frac{1}{k+2} \cdot \left[\frac{(k+2)!}{\lambda g_{0k}}\right]^{1/(k+2)} \cdot \Gamma\left(\frac{1}{k+2}\right) + o\left(\frac{1}{\lambda^{1/(k+2)}}\right), \tag{15.32}$$

$$J_1 = \frac{1}{k+2} \cdot \left[\frac{(k+2)!}{\lambda g_{0k}}\right]^{2/(k+2)} \cdot \Gamma\left(\frac{2}{k+2}\right) + o\left(\frac{1}{\lambda^{2/(k+2)}}\right). \tag{15.33}$$

c. If $\beta < 0$

$$J \sim \exp\left\{\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\}$$
$$\cdot \sqrt{2\pi k!} \cdot (\lambda g_{0k})^{-1/(2k+2)} \cdot ((k+1)!(\lambda - n\mu))^{-k/(2k+2)}, \tag{15.34}$$

$$J_1 \sim \left(\frac{-\beta\sqrt{\mu}(k+1)!}{g_{0k}\sqrt{\lambda}}\right)^{1/(k+1)} \cdot J. \tag{15.35}$$

Remark 15.7 Expression (15.34) increases exponentially due to the $(\lambda - n\mu)^{(k+2)/(k+1)}$ term in the exponent.

Theorem 15.2 (Performance measures) Under the assumptions of Lemma 15.2, the performance measures of M/M/n+G are approximated by:

a. Probability of wait.

If $\beta > 0$, the probability of wait coincides (asymptotically) with the Erlang-C approximation (7.5):

$$\mathrm{P}\{\mathrm{W} > 0\} \sim \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}. \tag{15.36}$$

If $\beta = 0$, the probability to get service immediately converges to zero at rate $\frac{1}{n^{k/(2k+4)}}$:

$$\mathrm{P}\{\mathrm{W} = 0\} = \frac{1}{n^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{k/(2k+4)}}\right). \tag{15.37}$$

If $\beta < 0$, the probability to get service immediately decreases to zero at an exponential rate:

$$\mathrm{P}\{\mathrm{W} = 0\} \approx \exp\left\{-\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\}$$

$$\cdot \frac{g_{0k}^{1/(2k+2)} \cdot (-\beta(k+1)!)^{k/(2k+2)}}{\lambda^{k/(4k+4)} \cdot \mu^{(k+2)/(4k+4)} \cdot \sqrt{2\pi k!} \cdot h(-\beta)} \,. \tag{15.38}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{n^{(k+1)/2}} \cdot \frac{g_{0k}}{(\beta\mu)^{k+1}} + o\left(\frac{1}{n^{(k+1)/2}}\right). \tag{15.39}$$

If $\beta = 0$

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{n^{(k+1)/(k+2)}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{(k+1)/(k+2)}}\right). \tag{15.40}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.41}$$

c. Average offered waiting time.

If $\beta > 0$, the average offered wait is given by the Erlang-C approximation (7.6):

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{1}{\beta\mu\sqrt{n}}\,. \tag{15.42}$$

If $\beta = 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{n^{1/(k+2)}} \cdot \frac{\Gamma\left(\frac{2}{k+2}\right)}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{(k+2)!}{\mu g_{0k}}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{1/(k+2)}}\right). \tag{15.43}$$

If $\beta < 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)} + o\left(\frac{1}{n^{1/(2k+2)}}\right). \tag{15.44}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]\,. \tag{15.45}$$

Remark 15.8 The value $-\beta/\sqrt{n}$ in formula (15.41) is the minimal reneging rate that is required to avoid queue explosion. Indeed, one can check that $-\beta/\sqrt{n}$ is asymptotically equivalent to the "fluid limit" of the probability-to-abandon [29] $1 - 1/\rho$, given $n \to \infty$.

Example. Phase-type patience times. An important special case of distributions, described in Theorem 15.2, is phase-type (see Asmussen [2] or Issaev [38]). Here we study the behavior of the phase-type density near zero, which is essential if one is to apply Theorem 15.2.

Definition. Consider a continuous-time Markov process $\{X = X_t,\ t \geq 0\}$ with a finite state-space $\{1, 2, \ldots, k, \Delta\}$, where $1, 2, \ldots, k$ are transient states and Δ is the absorbing state. The distribution of X is characterized by:

- *Initial distribution* $\bar{q} = (q_1, \ldots, q_k)$, where $q_i = \mathrm{P}\{X_0 = i\},\ 1 \leq i \leq k$ (the process cannot start from the absorbing state).
- *Phase-type generator* R, a $k \times k$ matrix of transition rates between the transient states. We know that $R_{kk} < 0$, $R_{kj} \geq 0$ for $k \neq j$, and $\sum_{i=1}^{k} R_{ki} \leq 0$.
- *Absorbtion intensities* $\bar{r} = (r_1, \ldots, r_k)'$. Overall, the generator of X can be written as
$$Q \ = \ \begin{pmatrix} R & \bar{r} \\ 0, \ldots, 0 & 0 \end{pmatrix},$$
where every row in Q sums up to zero: $\bar{r} = -R \cdot \bar{1}$.

Let
$$T \ \stackrel{\Delta}{=} \ \inf\{t > 0 :\ X(t) = \Delta\}$$
denote the absorbtion time. Then $F_T(t) = \mathrm{P}_q\{T \leq t\}$ is a *phase-type distribution* with parameters $(\bar{q},\ R)$.

The cumulative distribution function of the phase-type distribution with parameters $(\bar{q}, R)$ is,
$$F_T(t) = 1 - \bar{q} \exp\{Rt\}\bar{1}\,,$$
and it has a density
$$f_T(t) = \bar{q} \exp\{Rt\}\bar{r}\,. \tag{15.46}$$
In order to apply Theorem 15.2, we must calculate the density at the origin and its derivatives. From (15.46), the density at the origin is
$$f_T(0) \ = \ \bar{q}\bar{r}$$

and its n-th derivative (for convenience, we denote also $f_T^{(0)}(t) \triangleq f_T(t)$)

$$f_T^{(n)}(0) \;=\; \bar{q}R^n\bar{r}\,. \tag{15.47}$$

Theorem 15.3 (Phase-Type patience) Represent the transient states of the underlying Markov process of a phase-type distribution by a directed graph. Two states j and k are connected if and only if $R_{jk} > 0$. For any initial state j ($q_j > 0$), let L_j denote the number of states in a minimal path that connects j with the absorbing state Δ. Define

$$L \;\triangleq\; \min_j L_j\,. \tag{15.48}$$

(For example, $L = n$ for the Erlang distribution with n phases and $L = 1$ for the hyper-exponential distribution.)

Then $f_T^{(L-1)}(0) > 0$. Moreover, if $L \geq 2$, then $f_T^{(i)}(0) = 0$ for $0 \leq i \leq L-2$. ■

Now Theorem 15.3 and formula (15.47) enable us to apply Theorem 15.2 to phase-type distributions.

15.1.3 Delayed distribution of patience

Lemma 15.3 (Building blocks) Assume that the density of patience time vanishes over the interval $[0, c]$, for some $c > 0$. (That means that all customers are willing to wait at least c.) Take c to be maximal in the sense that the density of patience time is positive at c: $g_c > 0$. For $\beta \neq 0$ (both negative and positive) consider the staffing level

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})\,.$$

For $\beta = 0$ let

$$n = \frac{\lambda}{\mu} + a\,, \qquad -\infty < a < \infty\,. \tag{15.49}$$

a. If $\beta > 0$

$$J \;=\; \frac{1}{n\mu-\lambda} - \frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{\beta\sqrt{\mu}} - \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} \;+\; o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\right), \tag{15.50}$$

$$\hat{\beta}_c \;\triangleq\; \beta\sqrt{\frac{\mu}{g_c}}\,. \tag{15.51}$$

If $\beta = 0$ and $a \neq 0$

$$J \sim \frac{1}{\mu a} \cdot (1 - e^{-\mu a c}) . \tag{15.52}$$

If $\beta = 0$ and $a = 0$

$$J \sim c . \tag{15.53}$$

If $\beta < 0$

$$J \sim \frac{e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} . \tag{15.54}$$

b. If $\beta > 0$

$$J_1 = \frac{1}{\lambda} \cdot \frac{1}{\beta^2 \mu} + o\left(\frac{1}{\lambda}\right) . \tag{15.55}$$

If $\beta = 0$ and $a \neq 0$

$$J_1 \sim \frac{1}{\mu^2 a^2} \cdot (1 - e^{-\mu a c}) - \frac{c e^{-\mu a c}}{\mu a} . \tag{15.56}$$

If $\beta = 0$ and $a = 0$

$$J_1 \sim \frac{c^2}{2} . \tag{15.57}$$

If $\beta < 0$

$$J_1 \sim \frac{c e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} . \tag{15.58}$$

Remark 15.9 In the case $\beta = 0$, performance measures are very sensitive to the remaining term $n - \lambda/\mu$. Therefore, in (15.49) this term is asymptotically small in comparison to $o(\sqrt{\lambda})$ in the other cases.

Theorem 15.4 (Performance measures) Under the assumptions of Lemma 15.3, the performance measures of the M/M/n+G system with delayed patience distribution are approximated by:

a. Probability of wait.

If $\beta > 0$, the asymptotic probability of wait coincides with the Erlang-C approximation (7.5) (or (15.36)):

$$\mathrm{P}\{\mathrm{W} > 0\} \sim \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1} . \tag{15.59}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{a}{1 - e^{-\mu a c}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.60}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \frac{\pi}{2} \cdot \frac{1}{\mu c} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.61}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; e^{-c(\lambda - n\mu)} \cdot \frac{\frac{1}{h(-\beta)\sqrt{\mu}}}{-\frac{1}{\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}}. \tag{15.62}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;\sim\; \frac{e^{-c(n\mu - \lambda)}}{\sqrt{\lambda}} \cdot \left\{\beta\sqrt{\mu} - \frac{\beta^2 \mu}{h(\hat{\beta}_c)\sqrt{g_c}}\right\}. \tag{15.63}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{a e^{-\mu a c}}{1 - e^{-\mu a c}} + o\left(\frac{1}{n}\right). \tag{15.64}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{1}{\mu c} + o\left(\frac{1}{n}\right). \tag{15.65}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.66}$$

(See Remark 15.8 on page 118.)

c. Average offered waiting time.

If $\beta > 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\beta\mu} + o\left(\frac{1}{\sqrt{n}}\right) \tag{15.67}$$

(Erlang-C approximation).

If $\beta = 0$ and $a \neq 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{1}{\mu a} - \frac{c e^{-\mu a c}}{1 - e^{-\mu a c}}. \tag{15.68}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{c}{2}. \tag{15.69}$$

If $\beta < 0$

$$\mathrm{E}[V] \;\sim\; \mathrm{E}[V \mid V > 0] \;\sim\; c\,. \tag{15.70}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]\,. \tag{15.71}$$

Remark 15.10 Formulae (15.68)-(15.71) imply that, for $\beta \leq 0$, average wait (both offered and actual) converges to positive constants. That distinguishes the case of delayed distributions from Theorems 15.1 and 15.2, where $\mathrm{E}[W]$ converged to zero.

The important case of deterministic patience times gives rise to similar statements:

Theorem 15.5 (Deterministic patience) Assume that patience time is deterministic and equal to $c > 0$.

a. Probability of wait.

If $\beta > 0$:

$$\mathrm{P\{W > 0\}} \;\sim\; \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}. \tag{15.72}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P\{W = 0\}} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{a}{1 - e^{-\mu a c}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.73}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P\{W = 0\}} \;=\; \frac{1}{\sqrt{n}} \cdot \frac{\pi}{2} \cdot \frac{1}{\mu c} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.74}$$

If $\beta < 0$

$$\mathrm{P\{W = 0\}} \;\sim\; e^{-c(\lambda - n\mu)} \cdot \frac{-\beta}{h(-\beta)}\,. \tag{15.75}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P\{Ab|}W > 0\} \;\sim\; \frac{e^{-c(n\mu - \lambda)}}{\sqrt{\lambda}} \cdot \beta\sqrt{\mu}\,. \tag{15.76}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P\{Ab|}W > 0\} \;=\; \frac{1}{n} \cdot \frac{a e^{-\mu a c}}{1 - e^{-\mu a c}} + o\left(\frac{1}{n}\right). \tag{15.77}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{1}{\mu c} + o\left(\frac{1}{n}\right). \tag{15.78}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.79}$$

c. Average offered waiting time.

If $\beta > 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\beta\mu} + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.80}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{1}{\mu a} - \frac{ce^{-\mu ac}}{1 - e^{-\mu ac}}. \tag{15.81}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{c}{2}. \tag{15.82}$$

If $\beta < 0$

$$\mathrm{E}[V] \;\sim\; \mathrm{E}[V \mid V > 0] \;\sim\; c. \tag{15.83}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]. \tag{15.84}$$

15.1.4 Patience with balking

Lemma 15.4 (Building blocks) Consider the QED operational regime

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \qquad \lambda \to \infty.$$

Assume that the patience-time distribution has an atom at zero. In other words, if wait is encountered, customers abandon immediately with probability $\mathrm{P}\{\mathrm{Blk}\} > 0$, or $\bar{G}(0) = 1 - \mathrm{P}\{\mathrm{Blk}\}$. Assume, in addition, that the survival function $\bar{G}$ is differential at the origin: $\bar{G}'(0) = -g_0$. (Here g_0 is the right-side derivative of the patience-time distribution function at the origin.) Then

a.

$$J \;=\; \frac{1}{\lambda \cdot \mathrm{P}\{\mathrm{Blk}\} + (n\mu - \lambda)} - \frac{g_0}{\lambda^2 \cdot \mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right). \tag{15.85}$$

b.

$$J_1 \;=\; \frac{1}{n^2\mu^2\mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^2}\right) . \tag{15.86}$$

Theorem 15.6 (Performance measures) Under the assumptions of Lemma 15.4, the performance measures of the M/M/n+G queueing system in the QED regime can be approximated by:

a. Probability to encounter queue decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{V > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right) . \tag{15.87}$$

Probability of wait decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{W > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{(1-\mathrm{P}\{\mathrm{Blk}\}) \cdot h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right) . \tag{15.88}$$

b. Conditional probability to abandon $\mathrm{P}\{\mathrm{Ab}|V > 0\}$ converges to the balking probability:

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \mathrm{P}\{\mathrm{Blk}\} + \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right) . \tag{15.89}$$

Conditional probability to abandon $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ decreases at rate $\frac{1}{n}$:

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\} \cdot (1-\mathrm{P}\{\mathrm{Blk}\})} + o\left(\frac{1}{n}\right) . \tag{15.90}$$

The unconditional probability to abandon decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1}{\sqrt{n}} \cdot h(-\beta) + o\left(\frac{1}{\sqrt{n}}\right) . \tag{15.91}$$

c. Conditional average offered wait $\mathrm{E}[V|V > 0]$ decreases at rate $\frac{1}{n}$:

$$\mathrm{E}[V|V > 0] \;=\; \frac{1}{n} \cdot \frac{1}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right) . \tag{15.92}$$

The average offered wait decreases at rate $\frac{1}{n^{3/2}}$:

$$\mathrm{E}[V] \;=\; \frac{1}{n^{3/2}} \cdot \frac{h(-\beta)}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^{3/2}}\right) . \tag{15.93}$$

d. Conditional average waiting time $\mathrm{E}[W|W>0]$ decreases at rate $\frac{1}{n}$:

$$\mathrm{E}[W|W>0] \;=\; \frac{1}{n}\cdot\frac{1}{\mu\cdot\mathrm{P}\{\mathrm{Blk}\}}+o\left(\frac{1}{n}\right). \tag{15.94}$$

The average wait $\mathrm{E}[W]$ decreases at rate $\frac{1}{n^{3/2}}$:

$$\mathrm{E}[W] \;=\; \frac{1}{n^{3/2}}\cdot\frac{(1-\mathrm{P}\{\mathrm{Blk}\})\cdot h(-\beta)}{\mu\cdot\mathrm{P}\{\mathrm{Blk}\}^2}+o\left(\frac{1}{n^{3/2}}\right). \tag{15.95}$$

Remark 15.11 Consider two events:

- $\{V>0\}$, which means that "a customer did not get service immediately";
- $\{W>0\}$, which means "positive actual wait".

In fact, $\{W>0\}$ is equivalent to $\{V>0,\tau>0\}$.

In Theorems 15.1-15.5, we did not distinguish between $\{V>0\}$ and $\{W>0\}$ since $\mathrm{P}\{\tau=0\}=0$. However, in the cases with balking we must be careful, calculating conditional probabilities like (15.89) and (15.90).

Remark 15.12 In contrast to Theorem 15.1, probabilities of wait (both $\mathrm{P}\{W>0\}$ and $\mathrm{P}\{V>0\}$) decrease at rate $O\left(\frac{1}{\sqrt{n}}\right)$. If the balking probability $\mathrm{P}\{\mathrm{Blk}\}=1$, (15.87) is equivalent to Jagerman's [44] QED result for M/M/n/n (Erlang-B), cited in (7.8). In addition, (15.91) demonstrates that the M/M/n+G queue with any positive fraction of balking implies, in the QED regime, the same fraction of lost customers as in M/M/n/n. In this sense, Balking turns out to be equivalent to Blocking.

Formula (15.89) provides insight into this striking similarity. We observe that in the M/M/n+G queue with balking, the fraction of customers that abandon after positive wait is negligible (the second term of (15.89)), which makes it similar to Erlang-B, where all lost customers abandon immediately.

Note that, given positive offered wait, a fixed proportion of customers abandon, and the system is similar to M/M/n+G in the quality-driven regime (fixed ρ). This is the reason why the second term of (15.89) and formula (15.92) will have counterparts in the quality-driven results (16.7) and (16.8) later on.

15.1.5 Patience with scaled balking

Below we treat a special case of M/M/n+G which, in practical terms, corresponds to small yet non-negligible balking.

Lemma 15.5 (Building blocks) Assume that the patience distribution depends on the system size n. Specifically, let the balking probability $\mathrm{P}_n\{\mathrm{Blk}\} = \dfrac{p_b}{\sqrt{n}}$, for some $p_b > 0$. Assume that the derivative of the survival function $\bar{G}_n$ at the origin is independent of the system size: $\bar{G}_n'(0) = -g_0$. Then

a.

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \tag{15.96}$$

where

$$\hat{\beta} \;\triangleq\; (\beta + p_b) \cdot \sqrt{\frac{\mu}{g_0}}\,. \tag{15.97}$$

b.

$$J_1 \;=\; \frac{1}{n\mu g_0}\left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right). \tag{15.98}$$

Theorem 15.7 (Performance measures) Under the assumptions of Lemma 15.5, the performance measures of the M/M/n+G queueing system in the QED regime can be approximated by:

a. The probability of delay and positive offered wait converge to a constant that depends on β, p_b and $\dfrac{g_0}{\mu}$:

$$\mathrm{P}\{V > 0\} \;\sim\; \mathrm{P}\{W > 0\} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{15.99}$$

where $\hat{\beta}$ is defined by formula (15.97).

b. Conditional probabilities to abandon decrease at rate $\dfrac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.100}$$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.101}$$

The unconditional probability to abandon P{Ab} also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (15.100) and (15.99).

c. Conditional average offered wait $\mathrm{E}[V|V>0]$ decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{15.102}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (15.102) and (15.99).

d. The average waiting time is equivalent to the average offered wait:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W>0] \;\sim\; \mathrm{E}[V \mid V>0] \tag{15.103}$$

e. The ratio between the probability to abandon of delayed customers and average wait of delayed customers converges to the value of the patience density at the origin:

$$\frac{\mathrm{P}\{\mathrm{Ab}|W>0\}}{\mathrm{E}[W|W>0]} \;\sim\; g_0\,. \tag{15.104}$$

Remark 15.13 Under scaled balking, we observe a clear similarity with the main case described in Theorem 15.1 (positive patience density at the origin). Some results (formulae (15.99), (15.100) and (15.102)) have exact counterparts in Theorem 15.1: the service grade β should be replaced by $(\beta + p_b)$. We also derived the linear relation (15.104), although this result does not prevail for the corresponding unconditional performance measures.

15.2 Numerical experiments

We proceed with analyzing the quality of the approximations, derived in Theorems 15.1, 15.2 and 15.4. Four patience distributions are chosen for the following analysis, all with their means equal to 2:

- *Uniform* distribution on [0,4]: illustrates Theorem 15.1 with $g_0 = 0.25$;
- *Hyperexponential* distribution (mixture of two exponentials, with means 1 and 3 respectively): conforms to Theorem 15.1 with $g_0 = 2/3$;

- *Erlang* (Gamma) distributions, two exponential phases, each with the mean equal to 1: Theorem 15.2 with $k = 1$ and $g_{01} = 1$;

- *Delayed exponential* distribution equal to 1 + exp(mean=1): Theorem 15.4, with $c = 1$ and $g_c = 1$.

We consider seven values of the service grade β that vary from -1.5 to 1.5 in step 0.5. For each value of the service grade we perform the following experiment. M/M/n+G queues with the service rate $\mu = 1$ are considered. Arrival rate λ increases from 20 to 1000 with a varying step (44 values of λ overall). The number of agents n increases according to the QED staffing rule:

$$n = \left[\frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}\right], \tag{15.105}$$

where, as usual, the square brackets in (15.105) denote the nearest integer value.

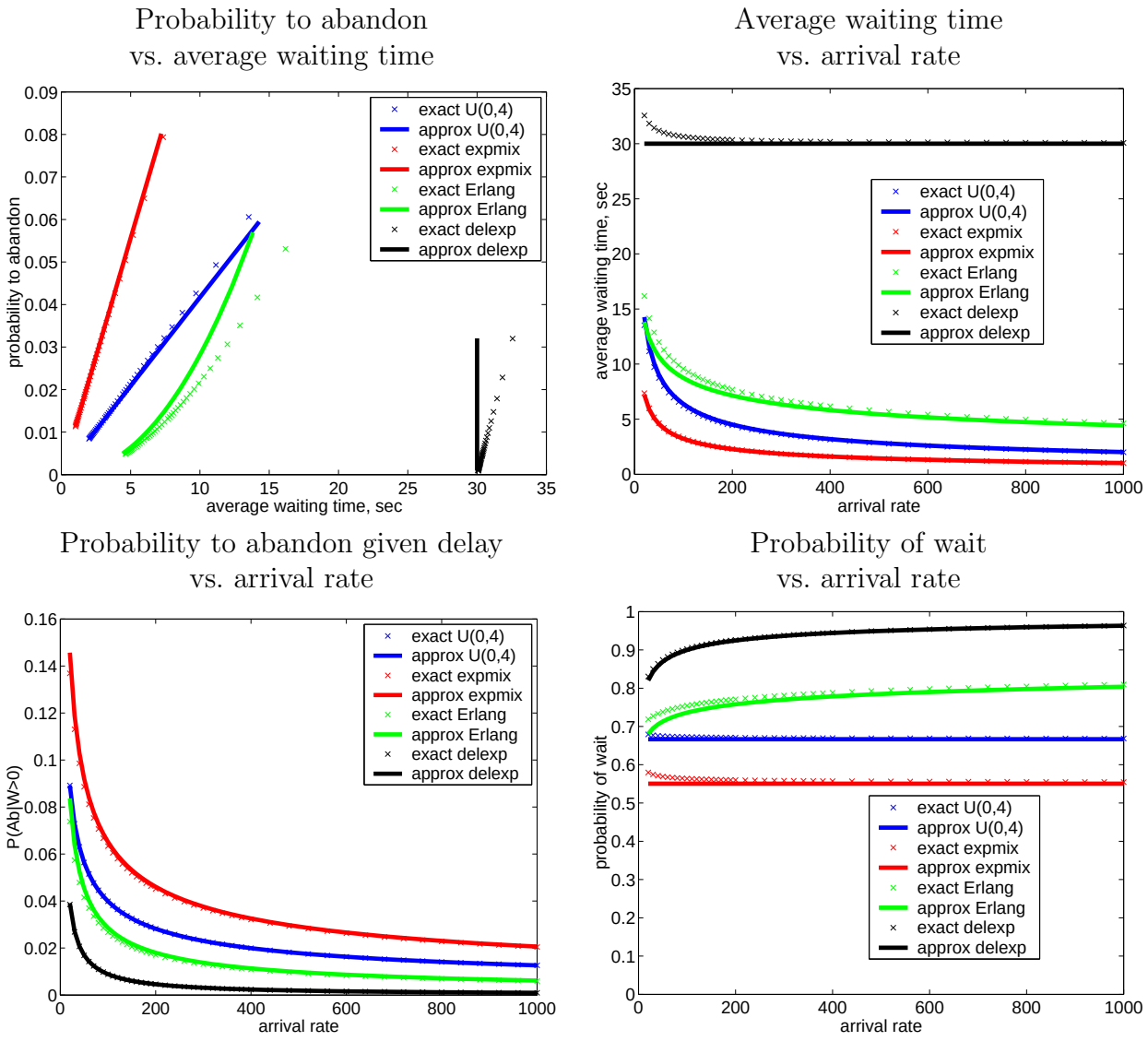
Then, for each M/M/n+G queue in consideration, exact (Baccelli and Hebuterne [3]) and approximate calculations are performed. The results are presented by four graphs for each service grade. The first graph presents a scatterplot of the probability to abandon against average wait. The other three plots show three different performance characteristics, as they change with the arrival rate: the average wait, the probability to abandon of delayed customers and the probability of wait. Solid lines are for approximations, and x's are for exact values.

Below we continue with presentations of seven special cases and finish with some general conclusions.

Example 1 (Figure 30): $\beta = 0$. A very good fit between approximations and exact values is observed for $\lambda > 100$ (and the fit is reasonable even for small arrival rates starting with $\lambda = 20$). Note the straight-line curves for the first two distributions in the first plot. (The two distributions with $g_0 = 0$ give rise to non-linear curves.) The average wait for the Erlang distribution drops slower than in the main case ($n^{-1/3}$ vs. $1/\sqrt{n}$) and the conditional probability to abandon decreases faster ($n^{-2/3}$ vs. $1/\sqrt{n}$). In the last case (delayed exponential), the probability to abandon decreases at rate $1/n$ and the average wait converges to a constant $c/2$ (30 seconds). The last plot demonstrates that the probability of wait rapidly converges to a constant in the main case. In the Erlang-

distribution case, it converges to one very slowly ($\mathrm{P}\{W{=}0\} \sim n^{-1/6}$), and, finally, for the delayed distribution $\mathrm{P}\{W{=}0\} \sim n^{-1/2}$.

Figure 30: **Service grade $\beta = 0$, performance measures and approximations**



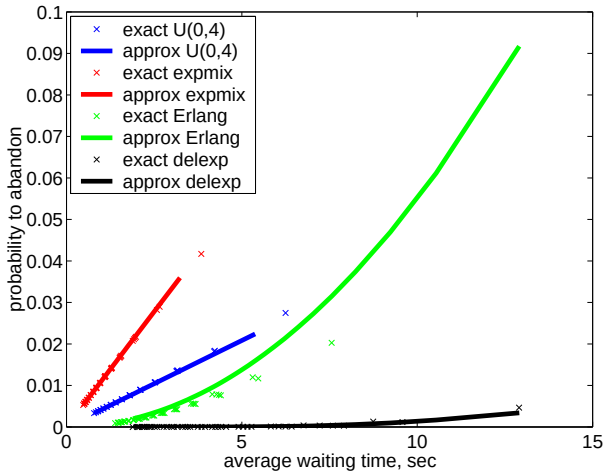
Example 2 (Figure 31): $\beta = 0.5$. The approximations for the first two distributions are excellent again. The slopes of the two corresponding curves in the first plot remain the same as in Figure 30: 0.25 and 2/3, respectively. Note that, in contrast to Figure 30, the difference between exact values and approximations does not decrease monotonically

on λ. That is due to approximation of the QED staffing level in (15.105) by the nearest integer value.

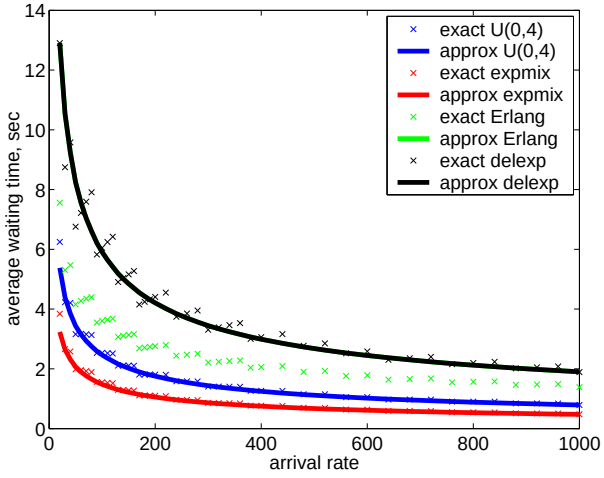
The delayed exponential distribution also demonstrates very good fit: the average wait and the probability of wait are very close to the Erlang-C approximation, and the probability to abandon decreases exponentially.

Figure 31: **Service grade $\beta = 0.5$, performance measures and approximations**

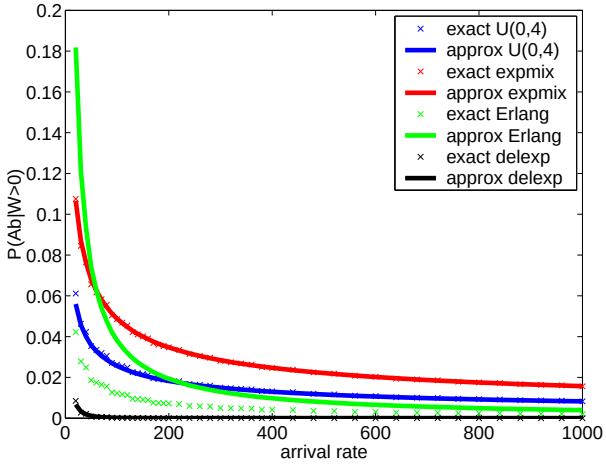
Probability to abandon
vs. average waiting time



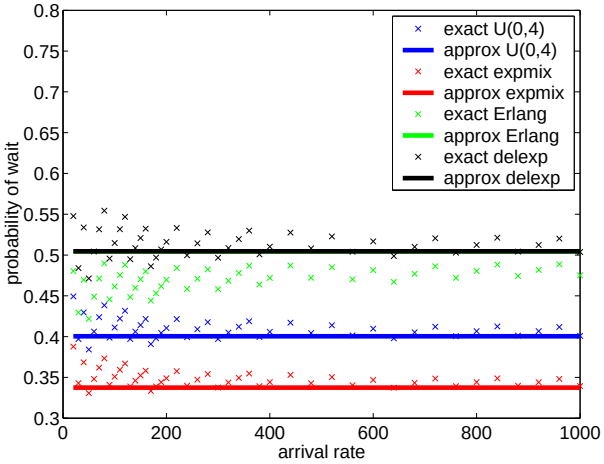
Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate

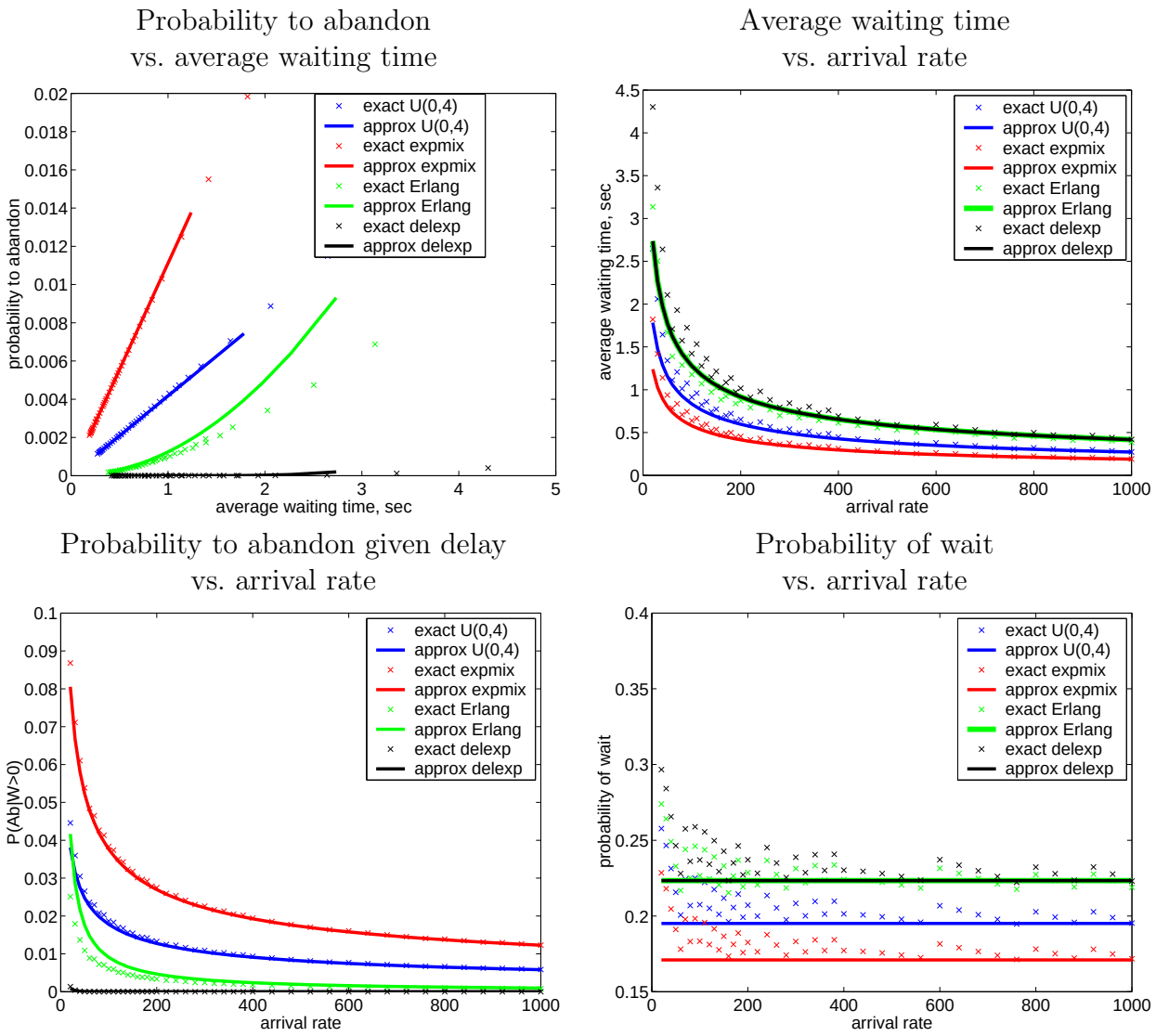


Probability of wait
vs. arrival rate



However, quality of approximations for the Erlang distribution is not so good (in fact, the worst one among all special cases considered in this subsection). Approximations for the average wait and the probability of wait coincide with Erlang-C formulae (and, therefore, with the approximation for delayed exponential). The fit of $\mathrm{P}\{W > 0\}$ is not bad at all. However, the fit of $\mathrm{E}[W]$ is less good and the fit of $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ is the worst of all. The reason seems to be unstableness of approximation (15.39) for small positive service grades β.

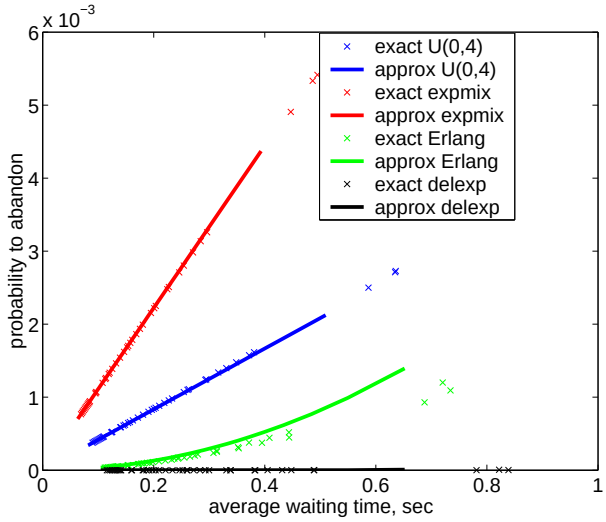
Figure 32: **Service grade $\beta = 1$, performance measures and approximations**



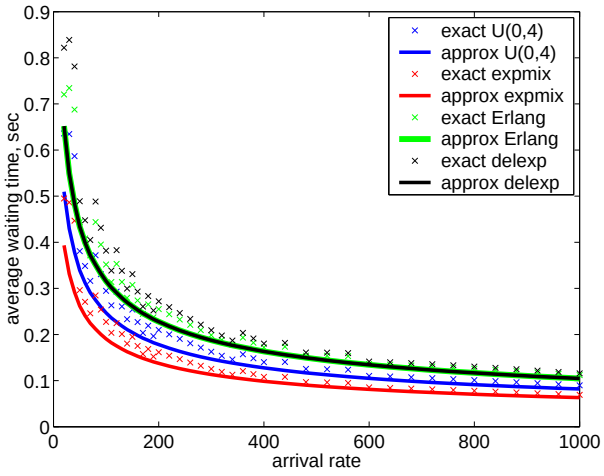
Example 3 (Figure 32): $\beta = 1$. Now the approximations for the Erlang distribution are much better than in Figure 31. In particular, the fit of $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ graph is reasonable for small values of λ and good for large values. (Recall from formula (15.39) that conditional probability to abandon decreases at rate $1/n$.) In the delayed exponential case, the probability to abandon is negligible for all values of λ.

Figure 33: **Service grade $\beta = 1.5$, performance measures and approximations**

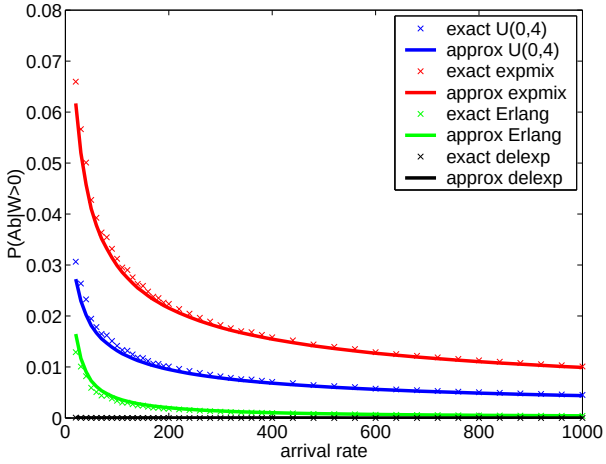
Probability to abandon
vs. average waiting time



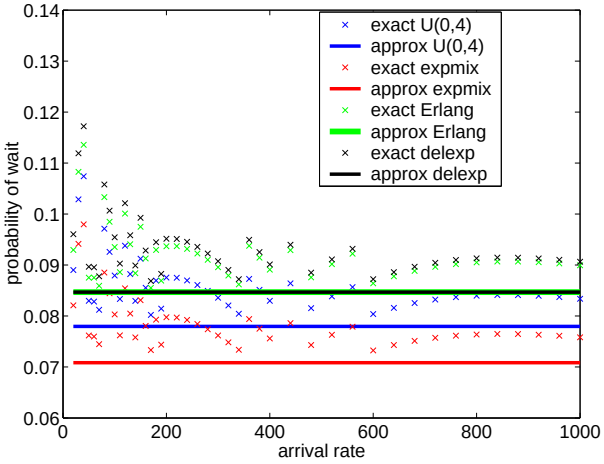
Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate

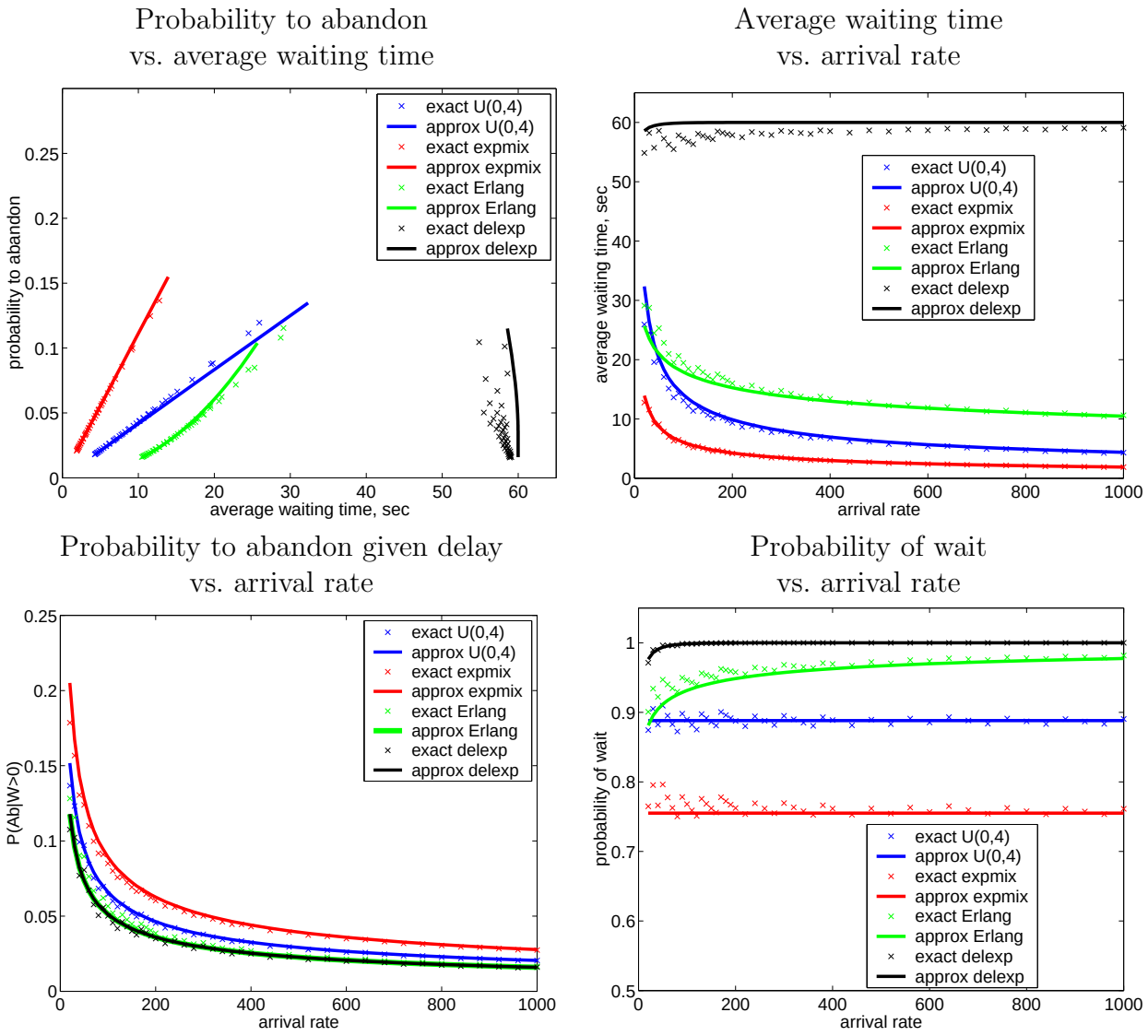


Probability of wait
vs. arrival rate



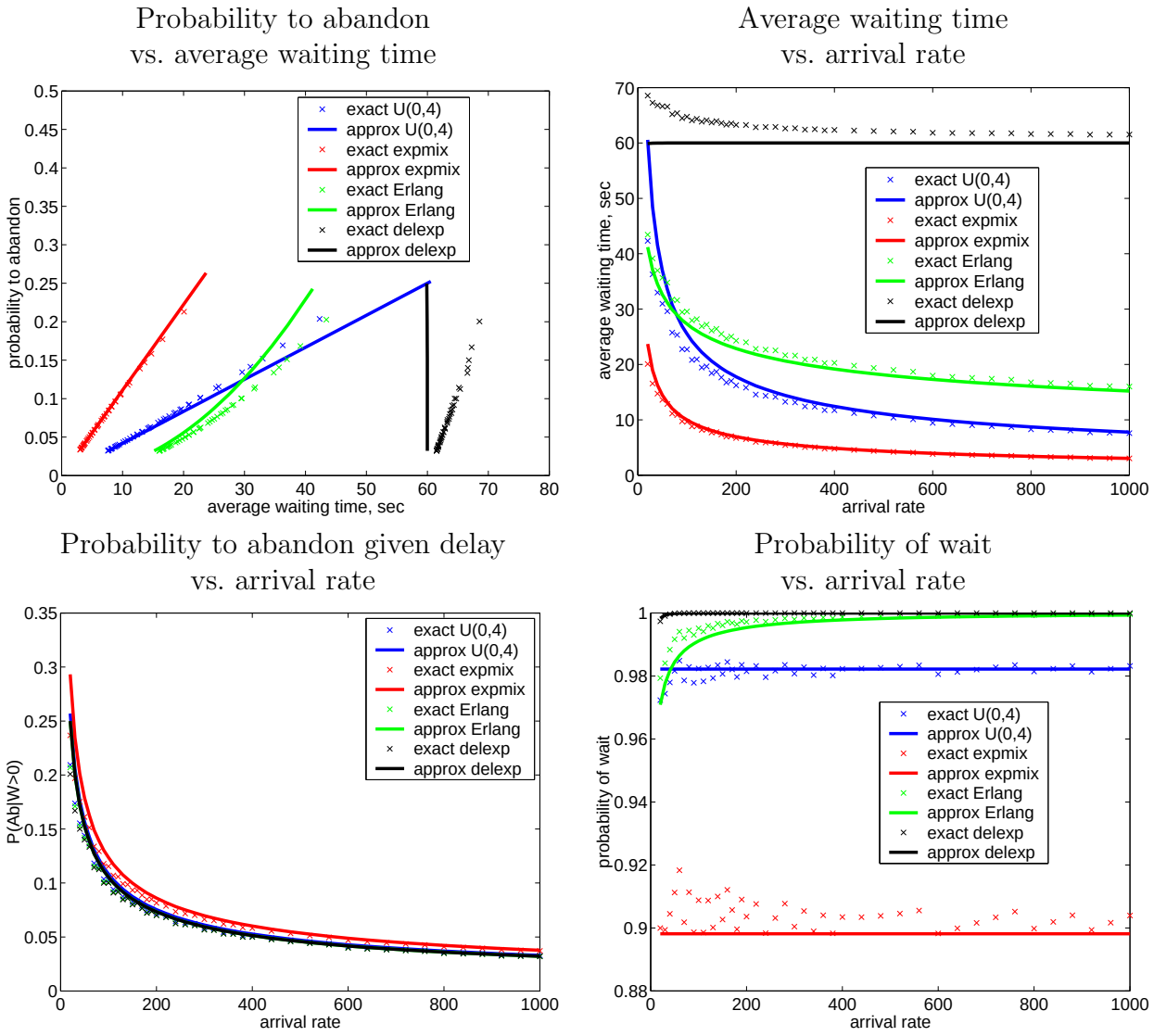
Example 4 (Figure 33): $\beta = 1.5$. The curves are similar to the curves in Figure 32. Note that the relative difference between the probability-of-wait approximations decreases significantly while we proceed from Figure 31 to Figure 33. In fact, it is easy to show from properties of the normal hazard (Section 10) that approximation (15.8) is equivalent to the Erlang-C approximation (7.5) given $\beta \to \infty$.

Figure 34: **Service grade $\beta = -0.5$, performance measures and approximations**



Example 5 (Figure 34): $\beta = -0.5$. The fit for the first two distributions ($g_0 > 0$) is fine. The approximation for $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ coincides for the last two distributions with $g_0 = 0$. (See Remark 15.8.) The average wait decreases at rate $n^{-1/4}$ for the Erlang patience and converges to delay time in the delayed exponential case. Finally, in the last two cases probability of wait converges to one exponentially (but with very different rates).

Figure 35: **Service grade $\beta = -1$, performance measures and approximations**



Example 6 (Figure 35): $\beta = -1$. Here we encounter two interesting phenomena. First, the conditional probabilities to abandon start to be very similar for the four distributions and close to $-\beta/\sqrt{n}$. (Recall formula (15.10) and take into account that $h(\hat{\beta})$ is small for large negative β.)

Another interesting phenomenon is observed in the last plot: $\mathrm{P}\{W > 0\}$ curve for exponential mixture is relatively far from the uniform one. To explain it, note that for large negative β

$$\mathrm{P}\{W = 0\} \;\approx\; \frac{\sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}}{1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}} \;\approx\; \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\,,$$

recall that the normal hazard $h(\cdot)$ decreases rapidly for large negative $\hat{\beta}$, and that the absolute value of $\hat{\beta}$ is larger for the uniform distribution. (Recall definition (15.3).)

Example 7 (Figure 36): $\beta = -1.5$. The tendencies observed in the previous figure (e.g. $-\beta/\sqrt{n}$ limit for the probabilities to abandon) are even more striking at this plot. Note also that the average wait approximation for the delayed exponential seems to underestimate the exact values for small and moderate loads.

General Conclusions.

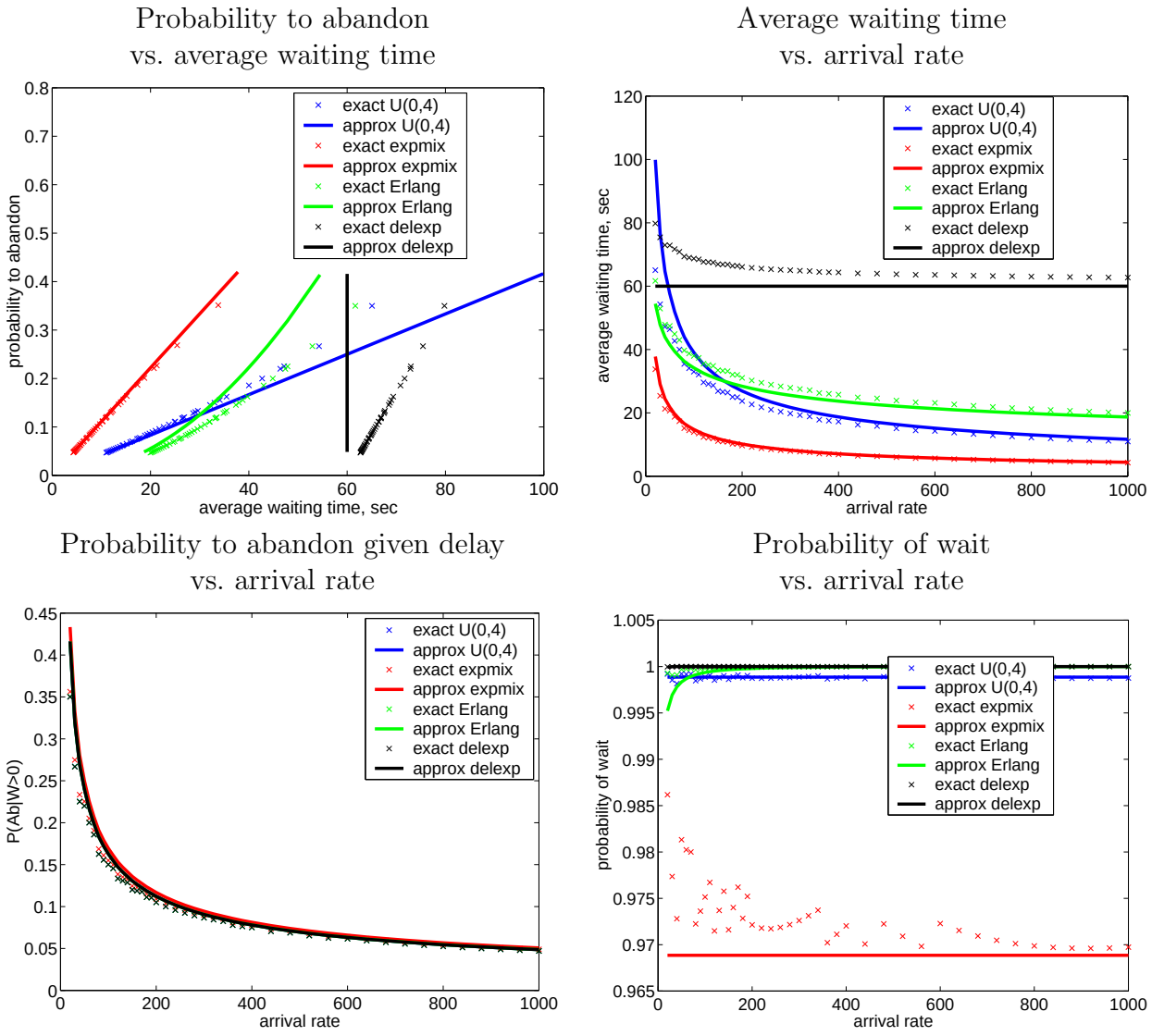
- Overall, the QED approximations are very good even for moderate staffing levels. Below (Subsections 16.2 and 17.2) we compare them with the quality-driven and the efficiency-driven approximations observing that, in most cases, the QED approximations are preferable.

- In the main case ($g_0 > 0$), the linear $\mathrm{P}\{\mathrm{Ab}\}$ / $\mathrm{E}[W]$ relation is confirmed for all values of the service grade.

- For relatively large positive β we observe convergence to the Erlang-C asymptotic formulae for the average wait and the probability of wait.

- For relatively large negative β the probability to abandon converges to $-\beta/\sqrt{n}$ for all distributions in consideration. (Recall Remark 15.8 after Theorem 15.2.)

Figure 36: **Service grade $\beta = -1.5$, performance measures and approximations**



15.3 Proofs of the QED results

Proof of Lemma 15.1.

a. First, we present the proof for the case when the QED staffing rule prevails *exactly*:

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}\,. \tag{15.106}$$

At the end of the proof of **a**, we show how to generalize it, in order to accommodate (15.1).

Define

$$h_\lambda(x) \triangleq \int_0^x \left[\lambda(\bar{G}(u)-1) - \beta\sqrt{\lambda\mu}\right] du\,. \tag{15.107}$$

Then, under the staffing (15.106)

$$J = \int_0^\infty \exp\{h_\lambda(x)\}dx\,. \tag{15.108}$$

The proof uses the Taylor expansion of $\bar{G}$ near the origin: $\bar{G}(u) \approx 1 - g_0 u$, approximating J by

$$J_A = \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx\,.$$

It is well-known (see, for example, de Bruijn [20], page 65) that $\forall \epsilon > 0\ \exists\, \delta > 0$ such that

$$|\bar{G}(u) - 1 + g_0 \cdot u| \le \epsilon u \quad \text{for } u \in [0,\delta]\,. \tag{15.109}$$

(Recall that $\bar{G}(0) = 1$, $\bar{G}'(0) = -g_0$). Then

$$-g_0 u - \epsilon u \le \bar{G}(u) - 1 \le -g_0 u + \epsilon u\,, \quad u \in [0,\delta]\,.$$

The integrand (15.107) is bounded by

$$-(g_0+\epsilon)\lambda u - \beta\sqrt{\lambda\mu} \;\le\; \lambda(\bar{G}(u)-1) - \beta\sqrt{\lambda\mu} \;\le\; -(g_0-\epsilon)\lambda u - \beta\sqrt{\lambda\mu}$$

for $u \in [0,\delta]$. Integrating twice the above inequalities we get

$$\begin{aligned} \int_0^\delta \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda(g_0+\epsilon)x^2}{2}\right\} dx &\le \int_0^\delta \exp\left\{h_\lambda(x)\right\} dx \\ &\le \int_0^\delta \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda(g_0-\epsilon)x^2}{2}\right\} dx\,. \end{aligned} \tag{15.110}$$

Now we need to construct a bound for $\int_\delta^\infty \exp\{h_\lambda(x)\}dx$, showing that, given $\lambda \to \infty$, the asymptotic behavior of $\int_0^\infty \exp\{h_\lambda(x)\}dx$ depends only on the values of $h_\lambda(x)$ near the origin.

Since $g_0 > 0$, the patience survival function $\bar{G}$ is strictly decreasing at the origin. Take

$$\alpha \triangleq 1 - \frac{1+\bar{G}(\delta/2)}{2} > 0\,. \tag{15.111}$$

Then, for λ large enough,

$$h_\lambda(x) \;=\; \int_0^x \left[\lambda(\bar{G}(u)-1)-\beta\sqrt{\lambda\mu}\right] du \;=\; \int_0^{\delta/2}\cdots + \int_{\delta/2}^x \cdots$$
$$\leq\; -\frac{\delta}{2}\beta\sqrt{\lambda\mu} - \int_{\delta/2}^x \alpha\lambda du \;=\; -\frac{\delta}{2}\beta\sqrt{\lambda\mu} - \alpha\lambda\left(x-\frac{\delta}{2}\right).$$

Integrating,

$$\int_\delta^\infty \exp\{h_\lambda(x)\}\,dx \;\leq\; \exp\left\{\frac{\alpha\lambda\delta}{2}-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}\cdot\frac{e^{-\alpha\lambda\delta}}{\alpha\lambda}$$
$$=\; \frac{\exp\left\{-\frac{\alpha\lambda\delta}{2}-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}}{\alpha\lambda} = o\left(e^{-\nu\lambda}\right),\quad \nu>0\,.$$

In other words,

$$\left|\int_0^\delta \exp\{h_\lambda(x)\}\,dx - \int_0^\infty \exp\{h_\lambda(x)\}\,dx\right| \;=\; o\left(e^{-\nu\lambda}\right). \tag{15.112}$$

Using identical arguments, the same relation between $\int_0^\delta$ and $\int_0^\infty$ can be derived for the two other integrals from (15.110).

Now, transforming expressions in (15.110), we get

$$\exp\left\{\frac{\beta^2\mu}{2(g_0+\epsilon)}\right\}\cdot\int_0^\infty \exp\left\{\frac{-\lambda(g_0+\epsilon)\left[x+\frac{\beta\sqrt{\lambda\mu}}{\lambda(g_0+\epsilon)}\right]^2}{2}\right\}dx \;+\; o\left(e^{-\nu\lambda}\right)$$
$$\leq\; \int_0^\infty \exp\{h_\lambda(x)\}\,dx$$
$$\leq\; \exp\left\{\frac{\beta^2\mu}{2(g_0-\epsilon)}\right\}\cdot\int_0^\infty \exp\left\{\frac{-\lambda(g_0-\epsilon)\left[x+\frac{\beta\sqrt{\lambda\mu}}{\lambda(g_0-\epsilon)}\right]^2}{2}\right\}dx \;+\; o\left(e^{-\nu\lambda}\right).$$

Calculating integrals:

$$\exp\left\{\frac{\beta^2\mu}{2(g_0+\epsilon)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda(g_0+\epsilon)}}\cdot\bar{\Phi}\left(\beta\sqrt{\frac{\mu}{\lambda(g_0+\epsilon)}}\right)+o\left(e^{-\nu\lambda}\right)$$
$$\leq\; J\;\leq\; \exp\left\{\frac{\beta^2\mu}{2(g_0-\epsilon)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda(g_0-\epsilon)}}\cdot\bar{\Phi}\left(\beta\sqrt{\frac{\mu}{\lambda(g_0-\epsilon)}}\right)+o\left(e^{-\nu\lambda}\right).$$

If λ is large enough,

$$(1-\epsilon)\sqrt{\frac{2\pi}{\lambda(g_0+\epsilon)}}\;\exp\left\{\frac{\beta^2\mu}{2(g_0+\epsilon)}\right\}\cdot\bar{\Phi}\left(\beta\sqrt{\frac{\mu}{g_0+\epsilon}}\right)\leq J$$
$$\leq\; (1+\epsilon)\sqrt{\frac{2\pi}{\lambda(g_0-\epsilon)}}\;\exp\left\{\frac{\beta^2\mu}{2(g_0-\epsilon)}\right\}\cdot\bar{\Phi}\left(\beta\sqrt{\frac{\mu}{g_0-\epsilon}}\right).$$

Now using the definition (10.1) of the normal hazard rate $h(x)$

$$(1-\epsilon)\frac{1}{\sqrt{\lambda(g_0+\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0+\epsilon}}\right)} \;\le\; J \;\le\; (1+\epsilon)\frac{1}{\sqrt{\lambda(g_0-\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0-\epsilon}}\right)}\,.$$

Since ϵ is arbitrary, and

$$\lambda \sim n\mu \qquad (\lambda, n \to \infty)$$

in the QED regime, we get the statement (15.2).

Finally, assume that the QED staffing rule prevails asymptotically in the sense of (15.1). Then $\forall \tilde{\epsilon} > 0$, for large λ we have

$$\frac{\lambda}{\mu} + (1-\tilde{\epsilon})\beta\sqrt{\frac{\lambda}{\mu}} \;\le\; n \;\le\; \frac{\lambda}{\mu} + (1+\tilde{\epsilon})\beta\sqrt{\frac{\lambda}{\mu}}\,,$$

and

$$\int_0^\infty \exp\left\{\int_0^x \left[\lambda\bar{G}(u) - (\lambda + (1+\tilde{\epsilon})\beta\sqrt{\lambda\mu})\right]du\right\}dx \;\le\; J$$
$$\le\; \int_0^\infty \exp\left\{\int_0^x \left[\lambda\bar{G}(u) - (\lambda + (1-\tilde{\epsilon})\beta\sqrt{\lambda\mu})\right]du\right\}dx\,.$$

Now we can proceed with the proof above for the exact QED staffing and get the same statement using that $\tilde{\epsilon}$ is arbitrary.

Remark 15.14 The following three main ideas, that are part of the so-called *Laplace method* (see de Bruijn [20]), were applied in the proof above:

- Using the *Taylor expansion* near the origin, we approximated

  $$J^\delta \;\triangleq\; \int_0^\delta \exp\{h_\lambda(x)\}dx$$

  by

  $$J_A^\delta \;\triangleq \int_0^\delta \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\}dx\,.$$

  Specifically, we have shown that $\forall\epsilon > 0\ \exists\delta > 0$ such that

  $$1-\epsilon \;\le\; \liminf_{\lambda\to\infty}\frac{J^\delta}{J_A^\delta} \;\le\; \limsup_{\lambda\to\infty}\frac{J^\delta}{J_A^\delta} \;\le\; 1+\epsilon\,.$$

- *Exponential bounds* for $\int_\delta^\infty$ integrals were developed. This enabled us to prove that $J \sim J_A$, given $\lambda \to \infty$.

- In the end, we explained how to replace the exact staffing (15.106) by the asymptotic staffing (15.1).

These three steps will be used repeatedly in the other proofs of Part IV. In case the proofs are very similar to **a**, we shall omit details and simply refer to the "Laplace method".

b. In the QED regime,

$$\mathcal{E} \;=\; \int_0^\infty e^{-t}\left(1+\frac{\mu t}{\lambda}\right)^{\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1} dt\,.$$

Changing variables: $(t=\lambda x,\, x=\frac{t}{\lambda})$, we get

$$\begin{aligned}\mathcal{E} \;&=\; \lambda\int_0^\infty e^{-\lambda x}(1+\mu x)^{\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1}dx\\ &=\; \lambda\int_0^\infty \exp\left\{-\lambda x+\left(\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1\right)\ln(1+\mu x)\right\}dx\,. \qquad (15.113)\end{aligned}$$

It is well known that

$$\ln(1+\mu x)=\mu x-\frac{\mu^2x^2}{2}+O(x^3)\,,\qquad x\to 0\,, \qquad (15.114)$$

Substitute into formula (15.113) the first two terms of the Taylor expansion (15.114):

$$\begin{aligned}\mathcal{E}_A \;&\triangleq\; \lambda\cdot\int_0^\infty \exp\left\{-\lambda x+\left(\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1\right)\left(\mu x-\frac{\mu^2x^2}{2}\right)\right\}dx\\
&=\; \lambda\cdot\int_0^\infty \exp\left\{\beta\sqrt{\lambda\mu}\cdot x-\frac{\lambda\mu x^2}{2}-\mu x-\frac{\beta\sqrt{\lambda\mu^3}x^2}{2}+\frac{\mu^2x^2}{2}\right\}dx\\
&=\; \lambda\cdot\exp\left\{\frac{(\beta\sqrt{\lambda\mu}-\mu)^2}{\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2}\right\}\cdot\int_0^\infty \exp\left\{-\left(\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2\right)\left[\frac{x-\frac{\beta\sqrt{\lambda\mu}-\mu}{\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2}}{2}\right]^2\right\}dx\\
&\sim\; \exp\left\{\frac{\beta^2}{2}\right\}\cdot\lambda\sqrt{\frac{2\pi}{\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2}}\cdot\Phi\left(\frac{\beta\sqrt{\lambda\mu}-\mu}{\sqrt{\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2}}\right)\\
&\sim\; \sqrt{\frac{2\pi\lambda}{\mu}}\cdot\exp\left\{\frac{\beta^2}{2}\right\}\cdot\Phi(\beta)\;\sim\;\frac{\sqrt{n}}{h(-\beta)}\,.\end{aligned}$$

The Laplace argument, based on the Taylor expansion, ensures that $\mathcal{E}\sim\mathcal{E}_A$, given $\lambda\to\infty$. Here we provide a sketch of this argument. First, note that $\forall\,\epsilon>0\;\;\exists\,\delta>0$ such that,

$$\mu x-\frac{(\mu^2+\epsilon)x^2}{2}\le\ln(1+\mu x)\le\mu x-\frac{(\mu^2-\epsilon)x^2}{2},\quad 0\le x\le\delta\,.$$

These inequalities enable approximations for $\int_0^\delta$ (see Remark 15.14). Then we set *exponential bounds* $o(e^{-\nu\lambda})$ for $\int_\delta^\infty$ integrals. Specifically, for the integral (15.113) use that $\forall\delta>0\ \exists a<1$ such that if $x>\delta$ then $\ln(1+\mu x)\le a\mu x$.

Then, lower and upper bounds for expression (15.113) are equal to

$$\lambda\int_0^\infty \exp\left\{-\lambda x+\left(\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1\right)\left(\mu x-\frac{(\mu^2+\epsilon)x^2}{2}\right)\right\}dx+o(e^{-\nu\lambda})\,,\quad \nu>0,$$

and

$$\lambda\int_0^\infty \exp\left\{-\lambda x+\left(\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1\right)\left(\mu x-\frac{(\mu^2-\epsilon)x^2}{2}\right)\right\}dx+o(e^{-\nu\lambda})\,,\quad \nu>0.$$

Finally, we use that ϵ is arbitrary, completing the proof of (15.4).

c. In the QED regime,

$$J_1 \;=\; \int_0^\infty x\cdot\exp\{h_\lambda(x)\}dx \;=\; \int_0^\infty x\cdot\exp\left\{\int_0^x\left[\lambda(\bar G(u)-1)-\beta\sqrt{\lambda\mu}\right]du\right\}dx\,.$$

Straightforward calculations imply that

$$\begin{aligned} J_{1A}\;\triangleq\;\int_0^\infty x\cdot\exp\left\{-x\beta\sqrt{\lambda\mu}-\frac{\lambda g_0x^2}{2}\right\}dx \;&=\; \frac{1}{\lambda g_0}-\frac{\beta}{\lambda g_0}\sqrt{\frac{2\pi\mu}{g_0}}\exp\left\{\frac{\beta^2\mu}{2g_0}\right\}\left[1-\Phi\left(\beta\sqrt{\frac{\mu}{g_0}}\right)\right]\\ &=\; \frac{1}{n\mu g_0}\left[1-\frac{\hat\beta}{h(\hat\beta)}\right]+o\left(\frac{1}{n}\right)\,. \end{aligned}\tag{15.115}$$

Asymptotic equivalence between J_1 and J_{1A} is demonstrated via the Laplace argument, using inequality (15.109). Approximation for $\int_0^\delta$ integrals is proved very similarly to **a**. Consider the second part of the argument, exponential bounds for the $\int_\delta^\infty$ integrals above:

$$\int_\delta^\infty x\cdot\exp\{h_\lambda(x)\}dx\;\le\;\int_\delta^\infty x\cdot\exp\left\{-\alpha\lambda\left(x-\frac{\delta}{2}\right)-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}dx\,,$$

(α was defined in (15.111)).

$$\begin{aligned} &=\;\exp\left\{\alpha\lambda\frac{\delta}{2}-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}\cdot\int_\delta^\infty xe^{-\alpha\lambda x}dx\\ =\;\exp\left\{\alpha\lambda\frac{\delta}{2}-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}&\cdot\left[\frac{\delta}{\alpha\lambda}e^{-\alpha\lambda\delta}+\frac{1}{(\alpha\lambda)^2}e^{-\alpha\lambda\delta}\right]\;=\;o(e^{-\nu\lambda}),\quad \nu>0\,. \end{aligned}\tag{15.116}$$

The tail part of the J_{1A} integral,

$$\int_\delta^\infty x\cdot\exp\left\{-x\beta\sqrt{\lambda\mu}-\frac{\lambda g_0x^2}{2}\right\}dx\,,$$

can be treated as a special case of J (linear survival function near zero).

d. First, calculate the integral

$$J_{2A} \triangleq \int_0^\infty x^2 \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx = \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} \cdot \int_0^\infty x^2 \exp\left\{\frac{-\lambda g_0 \left(x + \frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}\right)^2}{2}\right\} dx\,,$$

(changing variables)

$$= \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} \cdot \int_{\frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}}^\infty \left(y - \frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}\right)^2 \exp\left\{\frac{-\lambda g_0 y^2}{2}\right\} dy\,.$$

and, after exact calculations,

$$= \frac{\sqrt{2\pi}}{(\lambda g_0)^{3/2}}\left(1 + \frac{\beta^2\mu}{g_0}\right)\left[1 - \Phi\left(\beta\sqrt{\frac{\mu}{g_0}}\right)\right] \cdot \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} - \frac{1}{(\lambda g_0)^{3/2}} \cdot \beta\sqrt{\frac{\mu}{g_0}}$$

$$= \frac{1}{(n\mu g_0)^{3/2}}\left[\frac{\hat\beta^2 + 1}{h(\hat\beta)} - \hat\beta\right] + o\left(\frac{1}{n^{3/2}}\right)\,.$$

The last equality follows from the definition of $\hat\beta$ and $\lambda \sim n\mu$ $(\lambda, n \to \infty)$.

Finally, in the same way as in **a** and **c**, we can validate the approximation of

$$J_2 = \int_0^\infty x^2 \cdot \exp\left\{\lambda \int_0^x \bar G(u)du - n\mu x\right\} dx$$

by

$$J_{2A} = \int_0^\infty x^2 \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx\,.$$

Proof of Theorem 15.1.

a. Formula (6.78), Lemma 15.1, parts **a** and **b**, and the equivalence $\lambda \sim n\mu$ $(n \to \infty)$, imply that

$$\mathrm{P}\{W > 0\} = \frac{\lambda J}{\mathcal{E} + \lambda J} \sim \frac{\frac{\sqrt{n\mu}}{h(\hat\beta)\sqrt{g_0}}}{\frac{\sqrt{n\mu}}{h(\hat\beta)\sqrt{g_0}} + \frac{\sqrt{n}}{h(-\beta)}} = \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat\beta)}{h(-\beta)}\right]^{-1}\,.$$

Now we must prove the opposite direction: if the probability of wait converges to a constant, then QED staffing prevails. The probability-of-wait function

$$P_{g_0,\mu}(\beta) \triangleq \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat\beta)}{h(-\beta)}\right]^{-1}$$

is monotonically decreasing in β. Hence, the inverse function $P^{-1}_{g_0,\mu}(\alpha)$, $0 < \alpha < 1$, is well-defined.

Assume that

$$\mathrm{P}_{\lambda,n_\lambda}\{W > 0\} \ \to \alpha, \quad 0 < \alpha < 1, \tag{15.117}$$

and take $\beta = P^{-1}_{g_0,\mu}(\alpha)$. We want to show that for all $\epsilon > 0$ and λ large enough,

$$\frac{\lambda}{\mu} + (\beta - \epsilon)\sqrt{\frac{\lambda}{\mu}} \ \leq \ n_\lambda \ \leq \ \frac{\lambda}{\mu} + (\beta + \epsilon)\sqrt{\frac{\lambda}{\mu}}\,. \tag{15.118}$$

Consider the staffing levels

$$n^1_\lambda \ = \ \left\lceil \frac{\lambda}{\mu} + (\beta - \epsilon)\sqrt{\frac{\lambda}{\mu}} \right\rceil \quad \text{and} \quad n^2_\lambda \ = \ \left\lfloor \frac{\lambda}{\mu} + (\beta + \epsilon)\sqrt{\frac{\lambda}{\mu}} \right\rfloor .$$

According to (15.8),

$$\mathrm{P}_{\lambda,n^1_\lambda}\{W > 0\} \ \to \ P_{g_0,\mu}(\beta - \epsilon) \ = \ \alpha + \delta_1, \quad \delta_1 > 0,$$

and

$$\mathrm{P}_{\lambda,n^2_\lambda}\{W > 0\} \ \to \ P_{g_0,\mu}(\beta + \epsilon) \ = \ \alpha - \delta_2, \quad \delta_2 > 0.$$

Therefore, for λ large enough,

$$\mathrm{P}_{\lambda,n^1_\lambda}\{W > 0\} > \alpha + \frac{\delta_1}{2} \quad \text{and} \quad \mathrm{P}_{\lambda,n^2_\lambda}\{W > 0\} < \alpha - \frac{\delta_2}{2}\,.$$

We know that $\mathrm{P}\{W > 0\}$ is monotonically decreasing in the staffing level n. This fact and (15.117) imply that for λ large enough $n^1_\lambda < n_\lambda < n^2_\lambda$, which proves (15.118).

b. Note that the definition of the QED regime implies

$$\lambda - n\mu = -\beta\sqrt{\lambda\mu} + o(\sqrt{\lambda})\,.$$

Applying the above to formula (6.81) and using the approximation for J from Lemma 15.1, we get the expression for the conditional probability to abandon.

c. Direct consequence of (6.84).

d. We must prove that

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[W]}{\mathrm{E}_\lambda[V]} = 1\,,$$

where the performance measures are indexed by the arrival rate in the QED regime. Recall that $W = \min(V, \tau)$, where V and τ are independent.

It can be derived from the proof of Lemma 15.1 (Part **c**) that, $\forall\, \delta > 0$,

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \delta]}{\mathrm{E}_\lambda[V]} \;=\; \frac{\int_\delta^\infty x v_\lambda(x) dx}{\int_0^\infty x v_\lambda(x) dx} \to 0\,. \tag{15.119}$$

(Specifically, formula (15.116) shows that an exponential bound is available for $\int_\delta^\infty$.)
Now,

$$\begin{aligned} & \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \tau]}{E_\lambda[V]} \\ = \;& \lim_{\lambda\to\infty} \left(\frac{\mathrm{E}_\lambda[V; V > \tau; \tau > \delta]}{\mathrm{E}_\lambda[V]} + \frac{\mathrm{E}_\lambda[V; V > \tau; \tau < \delta]}{\mathrm{E}_\lambda[V]} \right) \\ \le \;& \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; \tau < \delta]}{\mathrm{E}_\lambda[V]} \;=\; P\{\tau < \delta\}\,. \end{aligned} \tag{15.120}$$

The first term of (15.120) converges to zero due to (15.119). The last equality follows from the independence between V and τ. The probability $P\{\tau < \delta\}$ can be made arbitrarily small, since τ has no mass at zero. Hence,

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \tau]}{\mathrm{E}_\lambda[V]} = 0 \quad \text{and} \quad \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V \le \tau]}{\mathrm{E}_\lambda[V]} = 1\,. \tag{15.121}$$

Now,

$$\mathrm{E}_\lambda[W] \;=\; \mathrm{E}_\lambda[\min(V; \tau)] \;=\; \mathrm{E}_\lambda[V; V \le \tau] + \mathrm{E}_\lambda[\tau; \tau < V] \;\sim\; \mathrm{E}_\lambda[V]\,.$$

The other two statements of **d** follow from $\mathrm{P}\{W > 0\} \sim \mathrm{P}\{V > 0\}$.

e. Follows from **b**, **c** and **d**.

f. Use formula (6.87) from Background and the QED asymptotics for J and J_1:

$$\begin{aligned} \mathrm{E}[V \mid \mathrm{Ab}] \;&=\; \frac{(\lambda - n\mu)J_1 + J}{(\lambda - n\mu)J + 1} \\ &\sim\; \frac{-\beta\mu\sqrt{n}J_1 + J}{-\beta\mu\sqrt{n}J + 1} \\ &\sim\; \frac{1}{\sqrt{n}} \cdot \frac{-(\beta/g_0) \cdot (1 - \hat\beta/h(\hat\beta)) + 1/(\sqrt{\mu g_0} \cdot h(\hat\beta))}{1 - \hat\beta/h(\hat\beta)} \\ &\sim\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[\frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta \right]. \end{aligned}$$

Now we shall prove formula (15.15). Note that

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \mathrm{E}[\tau|\tau<V] \;=\; \frac{\mathrm{E}[\tau;\tau<V]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty \mathrm{E}[\tau;\tau<x]v(x)dx}{\mathrm{P}\{\mathrm{Ab}\}}$$
$$=\; \frac{\int_0^\infty v(x)\left(\int_0^x tdG(t)\right)dx}{\mathrm{P}\{\mathrm{Ab}\}}\,,$$

where $v(x)$ is the density of the offered wait (recall formula (6.26)).

Due to continuity of the patience density at the origin, $\forall\epsilon>0\ \exists\delta>0$ such that the density exists in $[0,\delta]$ and $g_0-\epsilon\le g(t)\le g_0+\epsilon$, for $t\in[0,\delta]$. Then, for $x\in[0,\delta]$,

$$\frac{x^2}{2}(g_0-\epsilon) \;\le\; \int_0^x tdG(t) \;=\; \int_0^x tg(t)dt \;\le\; \frac{x^2}{2}(g_0+\epsilon)\,.$$

Since $\int_0^x tdG(t)$ is bounded by $\bar\tau$ we can construct an exponential bound for $\int_\delta^\infty v(x)\left(\int_0^x tdG(t)\right)dx$ in the spirit of Lemma 15.1, part **c**. Then, based on the Laplace method, we deduce that

$$\int_0^\infty v(x)\left(\int_0^x tdG(t)\right)dx \;\sim\; \frac{g_0}{2}\int_0^\infty x^2v(x)dx \qquad (\lambda,n\to\infty)\,,$$

and

$$\mathrm{E}[W|\mathrm{Ab}] \;\sim\; \frac{g_0}{2\mathrm{P}\{\mathrm{Ab}\}}\cdot\lambda\pi_{n-1}J_2 \;\sim\; \frac{g_0}{2\mathrm{P}\{\mathrm{Ab}\}}\cdot\frac{\lambda J_2}{\mathcal{E}+\lambda J}\,. \qquad (15.122)$$

From part **d** of Lemma 15.1 we observe that the numerator of (15.122) is equal to

$$\lambda g_0J_2 \;=\; \frac{1}{\sqrt{n\mu g_0}}\cdot\left[\frac{1+\hat\beta^2}{h(\hat\beta)}-\hat\beta\right]+o\left(\frac{1}{\sqrt{n}}\right)\,.$$

For the denominator of (15.122)

$$2\mathrm{P}\{\mathrm{Ab}\}\cdot(\mathcal{E}+\lambda J) \;=\; 2(1+(\lambda-n\mu)J) \;\sim\; 2(1-\beta\mu\sqrt{n}J) \;\sim\; 2\cdot\left(1-\frac{\hat\beta}{h(\hat\beta)}\right)\,.$$

Dividing the numerator of (15.122) by the denominator:

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}}\cdot\frac{1}{2\sqrt{g_0\mu}}\cdot\frac{1+\hat\beta^2-\beta h(\hat\beta)}{h(\hat\beta)-\hat\beta}+o\left(\frac{1}{\sqrt{n}}\right)$$
$$=\; \frac{1}{\sqrt{n}}\cdot\frac{1}{2\sqrt{g_0\mu}}\cdot\left[\frac{1}{h(\hat\beta)-\hat\beta}-\hat\beta\right]+o\left(\frac{1}{\sqrt{n}}\right)\,.$$

Finally, we prove inequalities (15.16). The right one is a consequence of (10.2) and (10.5). The left inequality is equivalent to

$$2h(x) \;>\; x+\frac{1}{h(x)-x}$$

or

$$2h^2(x) - 3xh(x) + x^2 - 1 \;>\; 0\,,$$

which follows from convexity of h and formula (10.3).

g. From formula (6.26) for the density of the virtual offered wait it follows that

$$\mathrm{P}\left\{\frac{V}{\mathrm{E}[S]} > \frac{t\sqrt{\mu}}{\sqrt{\lambda}}\,\middle|\, V>0\right\} \;=\; \frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{\int_0^x \left[\lambda(\bar G(u)-1) - \beta\sqrt{\lambda\mu}\right] du\right\} dx}{J}\,. \quad (15.123)$$

Then using the Laplace method we show that the last expression is equivalent to

$$\frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx}{J} \;=\; \frac{\displaystyle \exp\left\{\frac{\beta^2\mu}{2g_0}\right\}\int_{\frac{t}{\sqrt{\mu\lambda}}+\frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}}^{\infty} \exp\left\{-\frac{\lambda g_0 y^2}{2}\right\} dy}{J}$$

(using the asymptotic expression for J (Lemma 15.1, Part **a**))

$$\sim\; \frac{\bar\Phi\left(\hat\beta + \sqrt{\dfrac{g_0}{\mu}}\cdot t\right)}{\bar\Phi(\hat\beta)}\,.$$

Now, in order to complete the proof, we need to substitute $\sqrt{\dfrac{\lambda}{\mu}}$ by $\sqrt{n}$ and the virtual offered wait V by the waiting time W in the left-hand part of (15.123). The validity of the first substitution can be verified using $\lambda \sim n\mu$.

For the second substitution we must prove

$$\mathrm{P}\{W>0\} \;\sim\; \mathrm{P}\{V>0\} \quad \text{and} \quad \mathrm{P}\left\{\frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \;\sim\; \mathrm{P}\left\{\frac{V}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\}.$$

Both relations directly follow from $W = \min(V,\tau)$ and $V \xrightarrow{p} 0$ $(n\to\infty)$ (see part **d**).

h. Conditional probability to abandon:

$$\mathrm{P}\left\{\mathrm{Ab}\,\middle|\, \frac{V}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \;=\; \frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)(1-\bar G(x))dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)dx}$$

$$\sim\; \frac{g_0\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} xv(x)dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)dx} \;\sim\; \frac{g_0\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} x\exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx}\,. \quad (15.124)$$

Calculating the numerator of (15.124), we get

$$\frac{1}{\lambda}\left[\exp\left\{-\frac{g_0 t^2}{2\mu}-\beta t\right\}-\hat{\beta}\cdot\frac{\bar{\Phi}\left(\hat{\beta}+\sqrt{\frac{g_0}{\mu}}\cdot t\right)}{\bar{\Phi}(\hat{\beta})}\right].$$

The denominator of (15.124) is equal to

$$\frac{1}{\sqrt{\lambda g_0}}\cdot\frac{\bar{\Phi}\left(\hat{\beta}+\sqrt{\frac{g_0}{\mu}}\cdot t\right)}{\bar{\Phi}(\hat{\beta})}.$$

Dividing the numerator by the denominator, we get (15.18).

Proof of Lemma 15.2.

a. $\boldsymbol{\beta > 0}$. Here and in the proofs below we denote $o(\cdot)$ deviation terms in the staffing rules (15.28) and (15.29) by $f(\lambda)$. Apply Lemma 9.1 with

$$k_1=\frac{1}{2};\quad l_1=1;\quad k_2=1;\quad l_2=k+2;\quad m=0; \tag{15.125}$$

(condition $\dfrac{k_1}{l_1}>\dfrac{k_2}{l_2}$ is valid for $k>0$) to derive that

$$J_A \triangleq \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x-f(\lambda)\mu x-\frac{\lambda g_{0k}x^{k+2}}{(k+2)!}\right\}dx \tag{15.126}$$

$$=\int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x-f(\lambda)\mu x\right\}dx-\int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x\right\}\cdot\frac{\lambda g_{0k}x^{k+2}}{(k+2)!}dx+o\left(\frac{1}{\lambda^{(k+1)/2}}\right)$$

$$=\frac{1}{n\mu-\lambda}-\frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+3}}+o\left(\frac{1}{\lambda^{(k+1)/2}}\right).$$

(We use that $n\mu-\lambda=\beta\sqrt{\lambda\mu}+f(\lambda)\mu$.) Now note that

$$J=\int_0^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du-\lambda x-\beta\sqrt{\lambda\mu}x-f(\lambda)\mu x\right\}dx. \tag{15.127}$$

Under the assumptions of Lemma 15.2, $\forall\epsilon>0\ \exists\delta>0$ such that, for $u\in[0,\delta]$,

$$1-\frac{(g_{0k}+\epsilon)u^{k+1}}{(k+1)!}\ \le\ \bar{G}(u)\ \le\ 1-\frac{(g_{0k}-\epsilon)u^{k+1}}{(k+1)!}. \tag{15.128}$$

From Lemma 9.3 $(m=0,n=0,k=1/2,l=1)$, there exists $\nu>0$ such that

$$\int_\delta^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x-f(\lambda)\mu x\right\}dx\ =\ o\left(e^{-\lambda^\nu}\right). \tag{15.129}$$

Formulae (15.128) and (15.129) enable us to apply the Laplace method (see Lemma 15.1) in order to show that J from (15.127) can be approximated by J_A from (15.126).

We use a similar reasoning in order to derive (15.31). (Lemma 9.1 is applied with $m = 1$ and other parameters taken from (15.125).) Specifically,

$$J_{1A} \triangleq \int_0^\infty x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \tag{15.130}$$

$$= \int_0^\infty x\cdot\exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx - \int_0^\infty x\cdot\exp\left\{-\beta\sqrt{\lambda\mu}x\right\}\cdot\frac{\lambda g_{0k}x^{k+2}}{(k+2)!}dx + o\left(\frac{1}{\lambda^{(k+1)/2}}\right)$$

$$= \frac{1}{(n\mu-\lambda)^2} - \frac{(k+3)\cdot\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+4}} + o\left(\frac{1}{\lambda^{(k+2)/2}}\right) \qquad (\lambda\to\infty)$$

and, then substitute

$$J_1 = \int_0^\infty x \cdot \exp\left\{\lambda\int_0^x \bar{G}(u)du - \lambda x - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx$$

instead of (15.130).

b. $\boldsymbol{\beta = 0}$. Using Lemma 9.2 with

$$k_1 = 1, \quad l_1 = k+2, \quad k_2 = \frac{1}{k+2}, \quad l_2 = 1, \quad m = 0,$$

we get

$$J_A \triangleq \int_0^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k}x^{k+2}}{(k+2)!}\right\} dx$$

$$= \frac{1}{k+2}\cdot\left(\frac{(k+2)!}{\lambda g_{0k}}\right)^{1/(k+2)}\cdot\Gamma\left(\frac{1}{k+2}\right) + o\left(\frac{1}{\lambda^{(k+1)/2}}\right), \tag{15.131}$$

and taking $m = 1$,

$$J_{1A} \triangleq \int_0^\infty x \cdot \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k}x^{k+2}}{(k+2)!}\right\} dx$$

$$= \frac{1}{k+2}\cdot\left(\frac{(k+2)!}{\lambda g_{0k}}\right)^{2/(k+2)}\cdot\Gamma\left(\frac{2}{k+2}\right) + o\left(\frac{1}{\lambda^{(k+1)/2}}\right). \tag{15.132}$$

Then we use (15.128), the Laplace method and Lemma 9.3 in order to substitute

$$J = \int_0^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - \lambda x - f(\lambda)\mu x\right\} dx,$$

and

$$J_1 \;=\; \int_0^\infty x \cdot \exp\left\{\lambda \int_0^x \bar{G}(u)du - \lambda x - f(\lambda)\mu x\right\} dx\,,$$

into (15.131) and (15.132), respectively. Note that Lemma 9.3 cannot be applied immediately to get

$$\int_\delta^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \;=\; o\left(e^{-\lambda^\nu}\right)$$

and

$$\int_\delta^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - \lambda x - f(\lambda)\mu x\right\} dx \;=\; o\left(e^{-\lambda^\nu}\right)$$

($-f(\lambda)$ can be positive). However, this problem can be easily solved. For example,

$$\int_\delta^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \;\le\; \int_\delta^\infty \exp\left\{-\frac{1}{2}\cdot\frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx$$

for λ large enough.

c. $\boldsymbol{\beta < 0}$. As in part **a**, we approximate J by

$$J_A \;\triangleq\; \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx\,, \tag{15.133}$$

and, then, apply the Laplace method to show that $J \sim J_A$. However, since $-\beta$ is a positive number, the integrand increases near zero, which requires additional work that involves somewhat cumbersome calculations. Define

$$x^* \;=\; \left(\frac{[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu]\cdot(k+1)!}{\lambda g_{0k}}\right)^{1/(k+1)},$$

to be equal to the point where the integrand of (15.133) reaches a maximum (note that x^* converges to zero at rate $\lambda^{-1/(2k+2)}$). Performing the variable change $y = x - x^*$, we get

$$\begin{aligned} J_A \;&=\; \exp\{[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu]\cdot x^*\} \\ &\cdot \int_{-x^*}^\infty \exp\left\{-\beta\sqrt{\lambda\mu}y - f(\lambda)\mu y - \frac{\lambda g_{0k}(y+x^*)^{k+2}}{(k+2)!}\right\} dy\,. \end{aligned} \tag{15.134}$$

Note that

$$\int_{-\infty}^{-x^*} \exp\{-\beta\sqrt{\lambda\mu}y\}dy \;=\; \frac{1}{-\beta\sqrt{\lambda\mu}}\exp\{\beta\sqrt{\lambda\mu}x^*\}\,.$$

Since β is negative, the integral above decreases at rate $\dfrac{\exp\{-\lambda^{k/(2k+2)}\}}{\sqrt{\lambda}}$ and we can change the integral limits in (15.134) to $\int_{-\infty}^{\infty}$. Now we expand $(y+x^*)^{k+2}$ from (15.134).

The free term $(x^*)^{k+2}$ is taken out of the integral and the $(k+2)y(x^*)^{k+1}$ term is cancelled by $[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu] \cdot y$. We must show now that the quadratic term in the expansion dominates the others. In other words,

$$J_A \sim \exp\left\{\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\} \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{\lambda g_{0k}}{2k!}(x^*)^k y^2\right\} dy$$

$$= \exp\left\{\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\}$$

$$\cdot\sqrt{2\pi k!} \cdot (\lambda g_{0k})^{-1/(2k+2)} \cdot ((k+1)!(\lambda - n\mu))^{-k/(2k+2)}\,.$$

We shall prove that the quadratic term in the integral (15.134) dominates the cubic term, an argument that can be repeated for the terms with larger degrees of y using Remark 9.2. Ignoring cumbersome constants, we must show that

$$\int_{-\infty}^{\infty} \exp\left\{-\lambda^{(k+2)/(2k+2)}y^2 - \lambda^{(k+3)/(2k+2)}y^3\right\} dy$$

$$\sim \int_{-\infty}^{\infty} \exp\left\{-\lambda^{(k+2)/(2k+2)}y^2\right\} dy = \sqrt{2\pi}\lambda^{-(k+2)/(4k+4)}\,.$$

The equivalence above follows from Lemma 9.1 with

$$k_1 = \frac{k+2}{2k+2}\,, \quad l_1 = 2, \quad k_2 = \frac{k+3}{2k+2}\,, \quad l_2 = 3,$$

(Note that condition (9.4) prevails for $k > 0$.)

Formula (15.35) for J_1 is proved via the approximation

$$J_{1A} \sim \int_0^{\infty} x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k}x^{k+2}}{(k+2)!}\right\} dx \sim x^* \cdot J_A\,,$$

where the second equivalence is obtained via the change of variables $y = x - x^*$.

Proof of Theorem 15.2.

a. Probability of wait.

$\boldsymbol{\beta > 0}$. Recall the asymptotic expression for $\mathcal{E}$:

$$\mathcal{E} = \sqrt{\frac{\lambda}{\mu}} \cdot \frac{1}{h(-\beta)}\,,$$

which does not depend on the patience distribution G. Hence,

$$\mathrm{P}\{W > 0\} = \frac{\lambda J}{\mathcal{E} + \lambda J} \sim \frac{\frac{\sqrt{\lambda}}{\beta\sqrt{\mu}}}{\frac{\sqrt{\lambda}}{h(-\beta)\sqrt{\mu}} + \frac{\sqrt{\lambda}}{\beta\sqrt{\mu}}} = \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}.$$

$\boldsymbol{\beta = 0}$. From Lemma 15.2, part **b** and taking into account that $\mathcal{E} \sim \sqrt{\frac{\lambda}{\mu}} \cdot \frac{\pi}{2}$:

$$\mathrm{P}\{W=0\} \;=\; \frac{\mathcal{E}}{\mathcal{E}+\lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; \frac{1}{\lambda^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2\mu}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{(k+2)!}\right]^{\frac{1}{k+2}}$$

$$\sim\; \frac{1}{n^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} .$$

$\boldsymbol{\beta < 0}$. Use that

$$\mathrm{P}\{W=0\} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; \frac{1}{\sqrt{\mu\lambda}J} \cdot \frac{1}{h(-\beta)} .$$

b. Probability of delayed customers to abandon.

$\boldsymbol{\beta > 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;\sim\; \frac{\frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+2}}}{\frac{\sqrt{\lambda}}{\beta\sqrt{\mu}}}$$

$$\sim\; \frac{g_{0k}}{\beta^{k+1}(\lambda\mu)^{(k+1)/2}} \;\sim\; \frac{1}{n^{(k+1)/2}} \cdot \frac{g_{0k}}{(\beta\mu)^{k+1}} .$$

$\boldsymbol{\beta = 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1-(\beta\sqrt{\lambda\mu}+f(\lambda)\mu)J}{\lambda J} \;\sim\; \frac{1}{\lambda J} \;\sim\; \frac{1}{n^{(k+1)/(k+2)}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} .$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1-(\beta\sqrt{\lambda\mu}+f(\lambda)\mu)J}{\lambda J} \;\sim\; -\beta\sqrt{\frac{\mu}{\lambda}} \;\sim\; \frac{-\beta}{\sqrt{n}} .$$

c. Average offered waiting time.

$\boldsymbol{\beta > 0}$.

$$\mathrm{E}[V|V>0] \;=\frac{J_1}{J} \;\sim\; \frac{1}{\beta\sqrt{\lambda\mu}} \;\sim\; \frac{1}{\beta\mu\sqrt{n}} .$$

$\boldsymbol{\beta = 0}$.

$$\mathrm{E}[V|V>0] \;=\frac{J_1}{J} \;\sim\; \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)} .$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim\; x^* \sim \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)} .$$

d. Average waiting time.

Since the survival function $\bar{G}$ is strictly decreasing near zero, the proof from Theorem 15.1, part **d**, can be duplicated.

Proof of Theorem 15.3.

We start with some definitions. Consider the underlying Markov process of the phase type distribution F_T. Let S_0 denote the set of states that correspond to positive values of the initial distribution: $i \in S_0$ iff $q_i > 0$. Then let S_1 be the set of states that can be reached by one jump from some initial state. Formally, $j \in S_1$ iff $j \notin S_0$ and there exists $i \in S_0$ such that $R_{ij} > 0$. Finally, we define recursively the set S_k which comprises states that can be reached by k jumps: $j \in S_k$ iff $j \notin S_0, \ldots, S_{k-1}$ and there exists $i \in S_{k-1}$ such that $R_{ij} > 0$. According to the definition (15.48) of L, the absorbing state $\Delta \in S_L$.

We shall number the states of the underlying Markov process in the following way: first, the states from S_0, then the states that belong to $S_1, \ldots, S_L$, etc. As a start, we prove the case $L = 2$ for illustrative purposes. The left-upper part of the generator matrix will have the form:

	$\boldsymbol{S_0}$	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	$\boldsymbol{\Delta}$
$\boldsymbol{S_0}$	?	+	0	0
$\boldsymbol{S_1}$	?	?	+	+

Here "?" means that the corresponding terms are irrelevant and "0" that all terms in quadrants $S_0 \times S_2$ and $S_0 \times \Delta$ are zero. The sign "+" in quadrants $S_0 \times S_1$ and $S_1 \times S_2$ means that there are no negative elements there and every column in these quadrants contains, at least, one positive term. In the same way, the non-negative column $S_1 \times \Delta$ contains, at least, one positive term.

Define n_i to be the number of states in the set S_i, $i = 0, 1, 2$, and let $N_i \stackrel{\Delta}{=} \sum_{j=0}^{i} n_j$, $i = 0, 1, 2$. Then the initial distribution will have the form

$$\bar{q} \;=\; (q_1, \ldots, q_{N_0}, 0, \ldots, 0), \quad q_i > 0,\ 1 \le i \le N_0.$$

Note that the vector $\bar{r}$ is equal to the column under state Δ. Hence the product $\bar{q} \cdot \bar{r} = 0$ and $f_T(0) = f_T^{(0)}(0) = 0$. Now consider the vector

$$\bar{q}R \;=\; (x_1, \ldots, x_{N_0}, \boldsymbol{x_{N_0+1}, \ldots, x_{N_1}}, 0, \ldots, 0)\,.$$

Note that all elements between x_{N_0+1} and x_{N_1} are positive since they correspond to the product of the positive part of $\bar{q}$ and the columns of $S_0 \times S_1$ quadrant. Then $(\bar{q}R) \cdot \bar{r} > 0$, since the inner product of the vector $(x_{N_0+1}, \ldots, x_{N_1})$ and the $S_1 \times \Delta$ column is positive. Hence, $f_T^{(1)}(0) > 0$ and the theorem is proved for $L = 2$.

Now we prove the general case: assume that the absorbing state Δ can be reached by L jumps. A relevant part of the generator matrix is equal to:

	$\boldsymbol{S_0}$	$\boldsymbol{S_1}$	$\boldsymbol{\ldots}$	$\boldsymbol{S_l}$	$\boldsymbol{S_{l+1}}$	$\boldsymbol{\ldots}$	$\boldsymbol{S_{L-1}}$	$\boldsymbol{S_L}$	$\boldsymbol{\ldots}$	$\boldsymbol{\Delta}$
$\boldsymbol{S_0}$	?	+	0	0	0	0	0	0	0	0
$\boldsymbol{S_1}$	?	?	+ 0	0	0	0	0	0	0	0
$\boldsymbol{\ldots}$	?	?	?	? +	0	0	0	0	0	0
$\boldsymbol{S_l}$	?	?	?	?	+	0	0	0	0	0
$\boldsymbol{S_{l+1}}$	?	?	?	?	?	+ 0	0	0	0	0
$\boldsymbol{\ldots}$	?	?	?	?	?	?	? +	0	0	0
$\boldsymbol{S_{L-1}}$	?	?	?	?	?	?	?	+	0	+

Formula (15.47) implies that we must prove

$$\bar{q}R^l\bar{r} \;=\; 0\,, \qquad 0 \le l \le L-2, \tag{15.135}$$

and

$$\bar{q}R^{L-1}\bar{r} \;>\; 0\,. \tag{15.136}$$

First, we want to show by induction that, for $0 \le l \le L-1$,

$$\bar{q}R^l \;=\; (x_1, \ldots, x_{N_{l-1}}, \boldsymbol{x_{N_{l-1}+1}, \ldots, x_{N_l}}, 0, \ldots, 0)\,, \tag{15.137}$$

where all vector elements between $x_{N_{l-1}+1}$ and x_{N_l} are positive. Statement (15.137) is clearly true for $l = 0$ (if we assign $N_{-1} = 0$). Assume that it is true for some $l \ge 0$. Then

$$\bar{q}R^{l+1} \;=\; (\bar{q}R^l) \cdot R \;=\; (y_1, \ldots, y_{N_l}, \boldsymbol{y_{N_l+1}, \ldots, y_{N_{l+1}}}, 0, \ldots, 0)\,.$$

The elements $y_{N_l+1}, \ldots, y_{N_{l+1}}$ are positive since they are calculated by multiplying the positive vector $(x_{N_{l-1}+1}, \ldots, x_{N_l})$ by the columns of $S_l \times S_{l+1}$ non-negative quadrant (one

positive element in the column, at least). The elements after $y_{N_{l+1}}$ are zero since the upper part of the respective generator columns (right of S_{l+1} columns) is zero. Now note that

$$\bar{r} \;=\; (0,\ldots,0,\boldsymbol{r_{N_{l-2}+1}},\ldots,\boldsymbol{r_{N_{l-1}}},\ldots)'\,,$$

where the vector $(r_{N_{l-2}+1},\ldots,r_{N_{l-1}})$ contains one positive element, at least. Substituting $l=L-1$ in (15.137) and multiplying by $\bar{r}$ we get (15.136). Equality (15.135) is obvious from (15.137), $l<L-1$.

Proof of Lemma 15.3. We proceed with the Laplace method from Lemma 15.1, using that $\bar{G}(u)=1,\ 0\le u\le c$, and approximating $\bar{G}(c+\epsilon)\approx 1-\epsilon g_c$, for small $\epsilon>0$.

a. $\boldsymbol{\beta>0}$.

$$\begin{aligned} J \;&=\; \int_0^\infty \exp\left\{\int_0^x [\lambda\bar{G}(u)-\lambda-\beta\sqrt{\lambda\mu}-f(\lambda)\mu]du\right\}dx \qquad (15.138)\\ &=\; \int_0^c \exp\{-\beta\sqrt{\lambda\mu}x-\mu f(\lambda)x\}dx+\exp\{(-\beta\sqrt{\lambda\mu}-\mu f(\lambda))c\}\\ &\quad\cdot\left[\int_c^\infty \exp\left\{-\beta\sqrt{\lambda\mu}(x-c)-\frac{\lambda g_c(x-c)^2}{2}\right\}dx+o\left(\frac{1}{\sqrt{\lambda}}\right)\right]\\ =\;& \frac{1}{n\mu-\lambda}\cdot\left[1-e^{-c(n\mu-\lambda)}\right]\;+\;e^{-c(n\mu-\lambda)}\cdot\int_c^\infty \exp\left\{-\beta\sqrt{\lambda\mu}(x-c)-\frac{\lambda g_c(x-c)^2}{2}\right\}dx\\ &\qquad +\,o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\right)\\ =\;& \frac{1}{n\mu-\lambda}-\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\cdot\left\{\frac{1}{\beta\sqrt{\mu}}-\frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}\right\}\;+\;o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\right). \end{aligned}$$

$\boldsymbol{\beta=0}$.

$$\begin{aligned} J \;&=\; \int_0^c e^{-\mu a x}dx+\int_c^\infty \exp\left\{[\lambda\bar{G}(u)-\lambda-a\mu]du\right\}dx\\ &\sim\; \int_0^c e^{-\mu a x}dx \;=\; \begin{cases} \dfrac{1}{\mu a}\cdot(1-e^{-\mu a c})\,, & a\neq 0\\ c\,, & a=0 \end{cases}. \end{aligned}$$

$\boldsymbol{\beta<0}$. Define the expression in the exponent of (15.138) by $h_\lambda(x)$. Then, similar to the case $\beta>0$:

$$J \;=\; \int_0^c \exp\{h_\lambda(x)\}dx+\int_c^\infty \exp\{h_\lambda(x)\}dx$$

$$= \frac{1}{\lambda - n\mu} \cdot \left[e^{c(\lambda - n\mu)} - 1 \right] + e^{c(\lambda - n\mu)} \cdot \int_c^\infty \exp\left\{ -\beta\sqrt{\lambda\mu}(x-c) - \frac{\lambda g_c (x-c)^2}{2} \right\} dx$$
$$+ o\left(\frac{e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \right)$$
$$= \frac{e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} + o\left(\frac{e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \right).$$

(The term $\dfrac{1}{\lambda - n\mu}$ is negligible if $\beta < 0$.)

b. $\boldsymbol{\beta > 0}$. Here we need only the first approximation term, ignoring $\int_c^\infty$:

$$J_1 = \int_0^c x \exp\{h_\lambda(x)\} dx + \int_c^\infty x \exp\{h_\lambda(x)\} dx \sim \frac{1}{(n\mu - \lambda)^2} = \frac{1}{\beta^2 \lambda\mu} + o\left(\frac{1}{\lambda}\right).$$

$\boldsymbol{\beta = 0}$.

$$J_1 \sim \int_0^c x e^{-\mu a x} dx = \begin{cases} \dfrac{1}{\mu^2 a^2} \cdot (1 - e^{-\mu a c}) - \dfrac{c e^{-\mu a c}}{\mu a}, & a \neq 0 \\ \dfrac{c^2}{2}, & a = 0 \end{cases}.$$

$\boldsymbol{\beta < 0}$.

$$\int_0^c x \exp\{h_\lambda(x)\} dx = \int_0^c x e^{(\lambda - n\mu)x} dx = \frac{c e^{c(\lambda - n\mu)}}{\lambda - n\mu} - \frac{e^{c(\lambda - n\mu)}}{(\lambda - n\mu)^2}. \tag{15.139}$$

The term $\int_c^\infty x \exp\{h_\lambda(x)\} dx$ can be approximated by

$$\int_c^\infty x \cdot \exp\left\{ (\lambda - n\mu)x - \frac{\lambda g_c (x-c)^2}{2} \right\} dx = e^{c(\lambda - n\mu)} \int_0^\infty (y+c) \cdot \exp\left\{ -\beta\sqrt{\lambda\mu} y - \frac{\lambda g_c y^2}{2} \right\} dx$$
$$= \frac{c e^{c(\lambda - n\mu)}}{h(\hat{\beta}_c)\sqrt{\lambda g_c}} + o\left(\frac{e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \right). \tag{15.140}$$

Now formulae (15.139) and (15.140) imply (15.58).

Proof of Theorem 15.4.

a. Probability of wait.

$\boldsymbol{\beta > 0}$. The asymptotic formula and its proof are the same as in Theorem 15.2.

$\boldsymbol{\beta = 0}$. Substitute (15.52), (15.53) and (15.4) into

$$\mathrm{P}\{W = 0\} = \frac{\mathcal{E}}{\mathcal{E} + \lambda J} \sim \frac{\mathcal{E}}{\lambda J}.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{W}=0\} \;=\; \frac{\mathcal{E}}{\mathcal{E}+\lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; e^{c(\lambda-n\mu)} \cdot \frac{\frac{1}{h(-\beta)\sqrt{\mu}}}{-\frac{1}{\beta\sqrt{\mu}}+\frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}}\,.$$

b. Probability of delayed customers to abandon.

$\boldsymbol{\beta > 0}$. Formula (15.63) is derived by substituting (15.50) into

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J}\,.$$

$\boldsymbol{\beta = 0}$. Substitute (15.52) and (15.53) into

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;=\; \frac{1-a\mu J}{\lambda J}\,.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;\sim\; \frac{\lambda-n\mu}{\lambda} \;\sim\; \frac{-\beta}{\sqrt{n}}\,.$$

c. Average offered waiting time.

$\boldsymbol{\beta > 0}$.

$$\mathrm{E}[V|V>0] \;=\frac{J_1}{J} \;=\; \frac{1}{\beta\mu\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right)\,.$$

$\boldsymbol{\beta = 0}$. Substitute (15.52), (15.53), (15.56) and (15.57) into

$$\mathrm{E}[V|V>0] \;=\frac{J_1}{J}\,.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim\; c\,.$$

d. Average waiting time.

According to the formulation of the theorem, $\tau = c + Y$, where $c > 0$ is a constant and Y is a random variable with a positive density at the origin. First, we prove that for all $\delta > 0$

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V>c+\delta]}{\mathrm{E}_\lambda[V]} \;=\; 0\,.$$

(which turns out equivalent to the proof in Theorem 15.1, part **d**, after the change of variables $y = x - c$). Then we continue along the lines of the proof of Theorem 15.1 via

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V > \tau]}{\mathrm{E}_\lambda[V]} \;\le\; \mathrm{P}\{\tau < c + \delta\}\,;$$

and

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V \le \tau]}{\mathrm{E}_\lambda[V]} \;=\; 1\,.$$

The proof of Theorem 15.5 is very similar to the proof of Theorem 15.4.

Proof of Lemma 15.4.

a. Recall that the patience survival function at the origin is equal to $\bar{G}(0) = 1 - \mathrm{P}\{\mathrm{Blk}\}$ and $g_0 = -\bar{G}'(0)$. Then $\forall \epsilon > 0\ \exists \delta > 0$ such that, for $u \in [0, \delta]$,

$$1 - \mathrm{P}\{\mathrm{Blk}\} - (g_0 + \epsilon)u \;\le\; \bar{G}(u) \;\le\; 1 - \mathrm{P}\{\mathrm{Blk}\} - (g_0 - \epsilon)u\,. \tag{15.141}$$

We shall approximate J by

$$J_A \;=\; \int_0^\infty \exp\left\{-\lambda \mathrm{P}\{\mathrm{Blk}\}x - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_0 x^2}{2}\right\} dx$$

Applying Lemma 9.1 with

$$m = 0, \quad k_1 = 1, \quad l_1 = 1, \quad k_2 = 1, \quad l_2 = 2,$$

we get

$$\begin{aligned} J_A \;&=\; \frac{1}{\lambda \mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\mu\lambda} + \mu f(\lambda)} - \frac{g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right) \\ &=\; \frac{1}{\lambda \mathrm{P}\{\mathrm{Blk}\} + (n\mu - \lambda)} - \frac{g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right)\,. \end{aligned}$$

Now using the Laplace method from Lemma 15.1 and (15.141), we can prove that the same approximation is valid for

$$J \;=\; \int_0^\infty \exp\left\{\lambda \int_0^x (\bar{G}(u) - 1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx\,.$$

b. Similar to **a**. (However, here only one approximation term is needed.)

$$J_1 \;=\; \int_0^\infty x{\cdot}\exp\left\{\lambda \int_0^x (\bar{G}(u) - 1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \;\sim\; \int_0^\infty x{\cdot}\exp\left\{-\lambda \mathrm{P}\{\mathrm{Blk}\}x\right\} dx$$

$$= \frac{1}{\lambda^2 \cdot \mathrm{P}\{\mathrm{Blk}\}^2} \sim \frac{1}{n^2\mu^2\ \mathrm{P}\{\mathrm{Blk}\}^2}\,.$$

Proof of Theorem 15.6.

a. Formula (15.85) from Lemma 15.4 implies that

$$J = \frac{1}{\lambda \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\lambda}\right). \tag{15.142}$$

Now the formula for the probability of positive virtual wait follows from (6.78), (15.142) and $\lambda \sim n\mu \ \ (\lambda, n \to \infty)$. Since

$$\mathrm{P}\{W > 0 | V > 0\} = 1 - \mathrm{P}\{\mathrm{Blk}\}\,,$$

formula (15.88) for the probability of actual wait prevails.

b. The conditional probability to abandon

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

$$= \frac{1 - [\beta\sqrt{\lambda\mu} + \mu f(\lambda)] \cdot \left[\frac{1}{\lambda \mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\lambda\mu} + \mu f(\lambda)} - \frac{g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^2}\right] + o\left(\frac{1}{\lambda^2}\right)}{\frac{\lambda}{\lambda \mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\lambda\mu} + \mu f(\lambda)} - \frac{\lambda g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{\lambda}\right)}$$

$$= \mathrm{P}\{\mathrm{Blk}\} + \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right).$$

Note that

$$\mathrm{P}\{\mathrm{Ab}\} = \mathrm{P}\{\mathrm{Ab}; W = 0\} + \mathrm{P}\{\mathrm{Ab}; W > 0\} = \mathrm{P}\{\mathrm{Blk}\} \cdot \mathrm{P}\{V > 0\} + \mathrm{P}\{\mathrm{Ab}|W > 0\} \cdot \mathrm{P}\{W > 0\}$$

$$= \mathrm{P}\{V > 0\} \cdot (\mathrm{P}\{\mathrm{Blk}\} + (1 - \mathrm{P}\{\mathrm{Blk}\}) \cdot \mathrm{P}\{\mathrm{Ab}|W > 0\})\,.$$

Hence,

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} = (\mathrm{P}\{\mathrm{Ab}|V > 0\} - \mathrm{P}\{\mathrm{Blk}\}) \cdot \frac{1}{1 - \mathrm{P}\{\mathrm{Blk}\}}\,,$$

which implies formula (15.90). Finally, (15.91) follows from formulae (15.87) and (15.89).

c. Statement (15.92) is a consequence of (6.84). Then formula (15.87) implies (15.93).

d. First we derive (15.95).

$$\mathrm{E}[W] = \mathrm{E}[\min(V, \tau)] = \mathrm{E}[\min(V, \tau)|\tau > 0] \cdot (1 - \mathrm{P}\{\mathrm{Blk}\}) \sim \mathrm{E}[V] \cdot (1 - \mathrm{P}\{\mathrm{Blk}\}) \quad (\lambda, n \to \infty)\,,$$

where the last equivalence can be proved using the methods of Theorem 15.1, part **d**. Now formulae (15.93) and (15.88) imply (15.94) and (15.95), respectively.

Proof of Lemma 15.5.

a. The proof is similar to Lemma 15.1, part **a**, and is implied by

$$J = \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - n\mu x\right\} dx = \int_0^\infty \exp\left\{\lambda \int_0^x (\bar{G}(u) - 1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx$$

$$\sim \int_0^\infty \exp\left\{-(\beta + p_b)\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx\,, \qquad (15.143)$$

where (15.143) follows from

$$\mathrm{P}_n\{\mathrm{Blk}\} = \frac{p_b}{\sqrt{n}} \sim p_b\sqrt{\frac{\mu}{\lambda}} \qquad (\lambda, n \to \infty)$$

and the Laplace method from Lemma 15.1.

The proof of Part **c** is identical to the proof of Lemma 15.1 (Part **c**), where β is replaced by $(\beta + p_b)$.

Proof of Theorem 15.7.

a. Direct consequence of Lemma 15.5 (parts **a** and **b**):

$$\mathrm{P}\{V > 0\} = \frac{\lambda J}{\mathcal{E} + \lambda J} \sim \left[1 + \sqrt{\frac{g_0}{\mu}}\right]^{-1}.$$

Since the fraction of balking customers cannot exceed the order $O\left(\frac{1}{\sqrt{n}}\right)$ we get $\mathrm{P}\{V > 0\} \sim \mathrm{P}\{W > 0\}$.

b. From Lemma 15.5,

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} = \frac{1 - (n\mu - \lambda)J}{\lambda J} = \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta\right] + o\left(\frac{1}{\sqrt{n}}\right).$$

Now note (see the proof of Theorem 15.6, Part **b**) that

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} = (\mathrm{P}\{\mathrm{Ab}|\mathrm{V} > 0\} - \mathrm{P}\{\mathrm{Blk}\}) \cdot \frac{1}{1 - \mathrm{P}\{\mathrm{Blk}\}}.$$

Since $1 - \mathrm{P}\{\mathrm{Blk}\} \sim 1$, the last expression is equivalent to

$$\frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta - p_b\right] + o\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}] + o\left(\frac{1}{\sqrt{n}}\right).$$

c.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot [h(\hat{\beta}) - \hat{\beta}]\,.$$

d. As in the proof of Theorem 15.6, part **d**,

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V] \cdot (1 - \mathrm{P}\{\mathrm{Blk}\}) \;\sim\; \mathrm{E}[V]\,.$$

The equivalence between $\mathrm{E}[W|W>0]$ and $\mathrm{E}[V|V>0]$ follows from part **a**.

16 Quality-Driven operational regime

16.1 Formulation of results

The quality-driven (QD) operational regime is defined by

$$n = \frac{\lambda}{\mu} \cdot (1+\gamma) + o(\sqrt{\lambda})\,, \qquad \gamma > 0\,. \tag{16.1}$$

In this regime the offered load per agent

$$\rho \;=\; \frac{\lambda}{n\mu} \;\to\; \frac{1}{1+\gamma} \;<\; 1\,.$$

If $o(\sqrt{\lambda}) \equiv 0$ in (16.1), the connection between ρ and γ is exact and given by:

$$\rho = \frac{1}{1+\gamma} \quad \text{and} \quad \gamma = \frac{1-\rho}{\rho}\,. \tag{16.2}$$

Lemma 16.1 (Building blocks) Assume that the density of the patience time at the origin exists and is positive: $g_0 > 0$. Then

a.

$$J \;=\; \frac{1}{n\mu - \lambda} - \frac{g_0}{\lambda^2\gamma^3} + o\left(\frac{1}{\lambda^2}\right)\,. \tag{16.3}$$

b.

$$\mathcal{E} \;\sim\; \sqrt{2\pi n} \cdot (1+\gamma)^{n-1} \cdot \exp\left\{-\frac{\lambda\gamma}{\mu}\right\}\,. \tag{16.4}$$

c.

$$J_1 \;=\; \frac{1}{(n\mu - \lambda)^2} - \frac{3g_0}{\lambda^3\gamma^4} + o\left(\frac{1}{\lambda^3}\right)\,. \tag{16.5}$$

Theorem 16.1 (Performance measures) Under the assumptions of Lemma 16.1, the performance measures of the M/M/n+G queueing system in the quality-driven regime can be approximated by:

a. The probability of wait decreases exponentially in n. Specifically,

$$\mathrm{P}\{\mathrm{W}>0\} \;\sim\; \frac{1}{\sqrt{2\pi n}}\cdot\frac{1}{\gamma}\cdot\left(\frac{1}{1+\gamma}\right)^{n-1}\cdot\exp\left\{\frac{\lambda\gamma}{\mu}\right\}. \tag{16.6}$$

b. Probability to abandon given wait.

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1}{n}\cdot\frac{1+\gamma}{\gamma}\cdot\frac{g_0}{\mu}+o\left(\frac{1}{n}\right) \;=\; \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{g_0}{\mu}+o\left(\frac{1}{n}\right). \tag{16.7}$$

(Note that if the $o(\sqrt{\lambda})$ deviation term in (16.1) is not equal to zero, the two $o(1/n)$ terms in (16.7) will not be identical.)

c. Average offered waiting time.

$$\mathrm{E}[V\mid V>0] \;=\; \frac{1}{n}\cdot\frac{1+\gamma}{\gamma}\cdot\frac{1}{\mu}+o\left(\frac{1}{n}\right) \;=\; \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{1}{\mu}+o\left(\frac{1}{n}\right). \tag{16.8}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V];\quad \mathrm{E}[W\mid W>0] \;\sim\; \mathrm{E}[V\mid V>0]\,. \tag{16.9}$$

e. Ratio between the probability to abandon and average wait.

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;\sim\; g_0 \qquad (n\to\infty)\,. \tag{16.10}$$

f. Total Service Factor.

$$\mathrm{P}\left\{\frac{W}{\mathrm{E}(S)}>\frac{t}{n}\;\middle|\;W>0\right\} \;\sim\; e^{-(1-\rho)t}\,. \tag{16.11}$$

Remark 16.1 Assume that the staffing level (16.1) is kept exact: $n=\dfrac{\lambda}{\mu}\cdot(1+\gamma)$. Then the asymptotic formula for the probability of wait transforms to:

$$\mathrm{P}\{\mathrm{W}>0\} \;\sim\; \frac{1}{\sqrt{2\pi n}}\cdot\frac{1}{1-\rho}\cdot(\rho e^{1-\rho})^n \qquad (n\to\infty)\,.$$

Remark 16.2 If the deviation in (16.1) is larger than $o(\sqrt{\lambda})$, for example,

$$n=\frac{\lambda}{\mu}\cdot(1+\gamma)+o(\lambda)\,,$$

formulae (16.7)-(16.11) still prevail. However, the approximation (16.6) can be wrong.

16.2 Numerical experiments

Two distributions that were already used in Subsection 15.2 are considered here again: uniform with support $[0, 4]$ and hyperexponential (mixture of two exponentials with means 1 and 3). We do not use the other two distributions from Subsection 15.2, since in the quality-driven regime our approximations are established for $g_0 > 0$ only. (Since the QD operational regime is less important than the QED regime, we did not try to duplicate the extensive set of special cases, analyzed in Theorems 15.1-15.7.)

The experiments are performed according to Subsection 15.2 guidelines. The arrival rate λ changes from 20 to 2000 (it has been from 20 to 1000 in the QED case). The quality-driven staffing rule is

$$n = \left[\frac{\lambda}{\rho\mu}\right], \qquad \rho \le 1\,.$$

Four values of ρ: 0.8, 0.9, 0.95 and 0.98 were chosen. In addition, we calculate the QED regime approximation using

$$\beta \;=\; \frac{n - \lambda/\mu}{\sqrt{\lambda/\mu}}\,.$$

(The service grade β increases with the increase of λ and n.)

Example 1 (Figures 37,38): $\rho = 0.8$.

Figure 37 presents evolution of several performance measures in the format of Subsection 15.2. Figure 38 studies our approximations when performance measures take very small values.

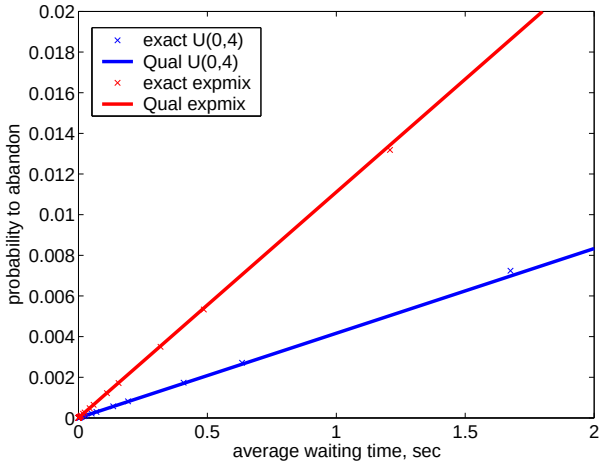
- Here and in all other special cases of Subsection 16.2, we observe an excellent linear fit between the average wait and the probability to abandon. In general, if the offered wait is small (quality-driven and QED regimes) and patience density at the origin is positive, a linear pattern prevails.

- The average wait and the probability of wait decrease exponentially on λ. The approximation does not depend on the specific distribution. The conditional probability to abandon decreases at rate $1/n$ or, the same, $1/\lambda$. For small values of λ the exact values are somewhere between quality-driven and QED approximations.

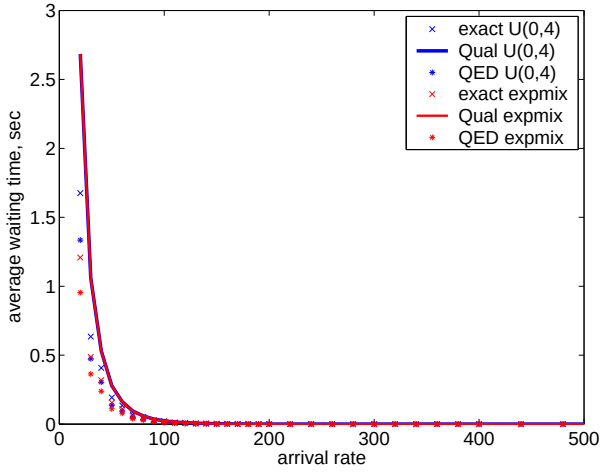
Figure 38 demonstrates that for large values of λ the quality-driven approximations are excellent (and much better than the QED approximations).

Figure 37: **Offered load per server $\rho = 0.8$, performance measures and approximations**

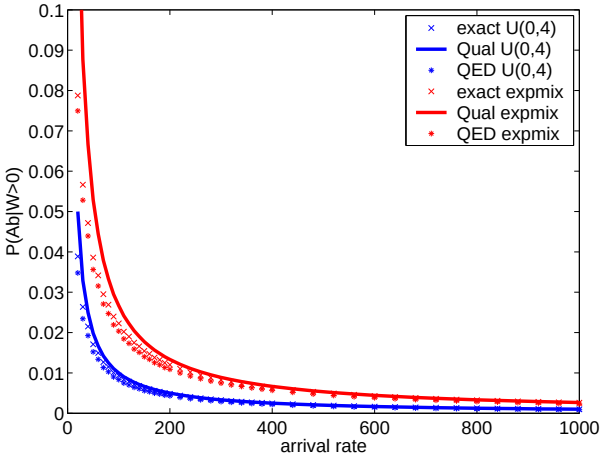
Probability to abandon
vs. average waiting time



Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate



Probability of wait
vs. arrival rate

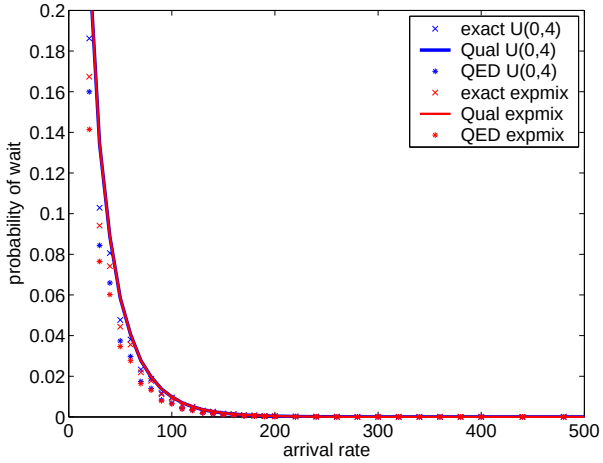
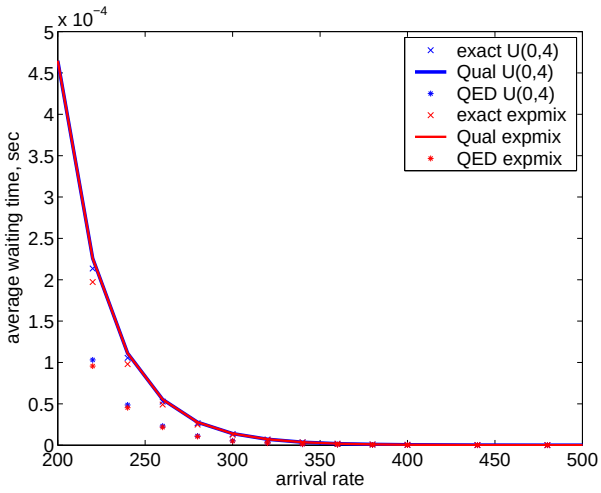
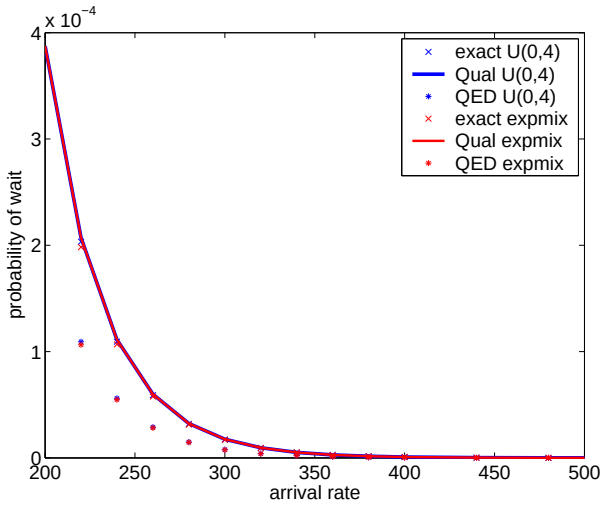


Figure 38: **Offered load per server $\rho = 0.8$, performance measures and approximations. Large values of arrival rate**

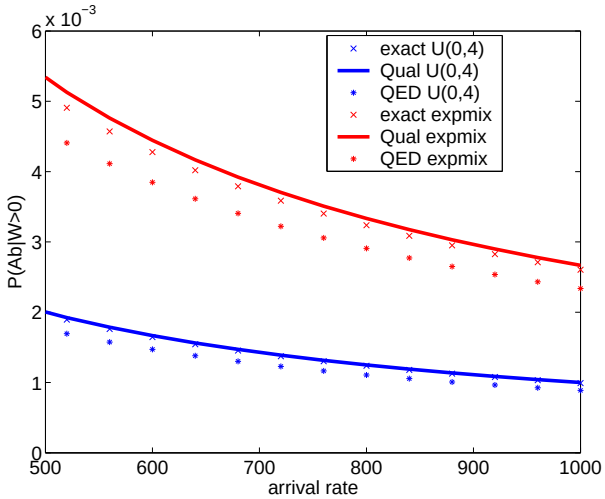
Average waiting time
vs. arrival rate



Probability of wait
vs. arrival rate



Probability to abandon given delay
vs. arrival rate



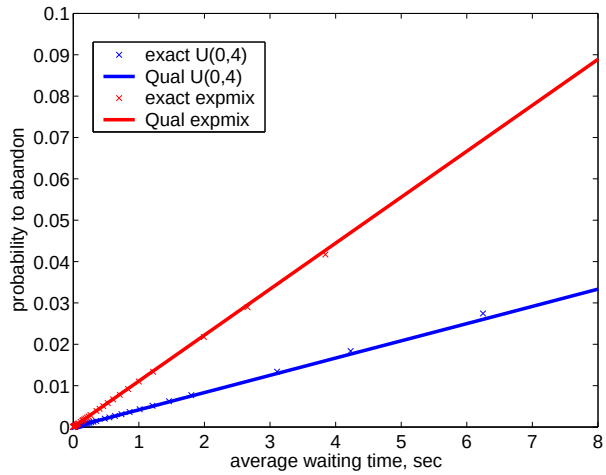
Example 2 (Figure 39): $\rho = 0.9$.

For small values of λ the QED approximations are better than the quality-driven. For larger values both types of approximations are good (and, again, one can check that the quality-driven approximations are excellent for small values).

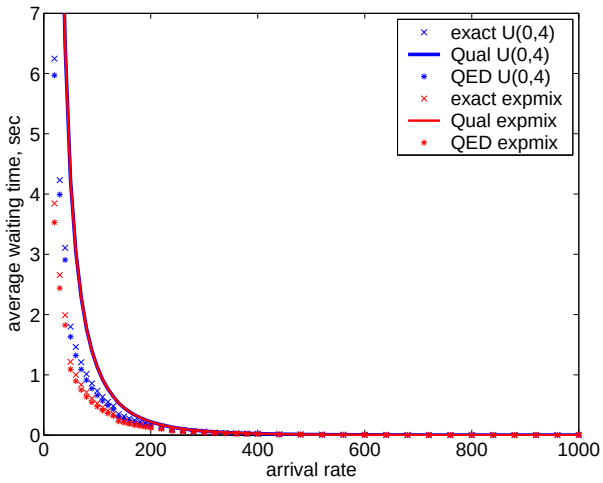
If we consider probability-to-abandon separately, the quality-driven approximation is better for the uniform distribution.

Figure 39: **Offered load per server $\rho = 0.9$, performance measures and approximations.**

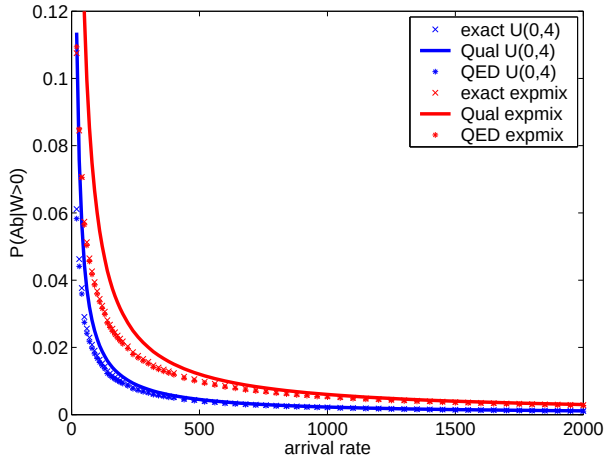
Probability to abandon
vs. average waiting time



Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate



Probability of wait
vs. arrival rate

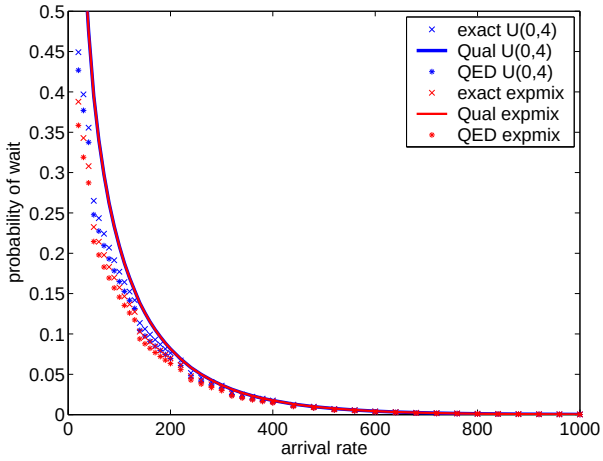
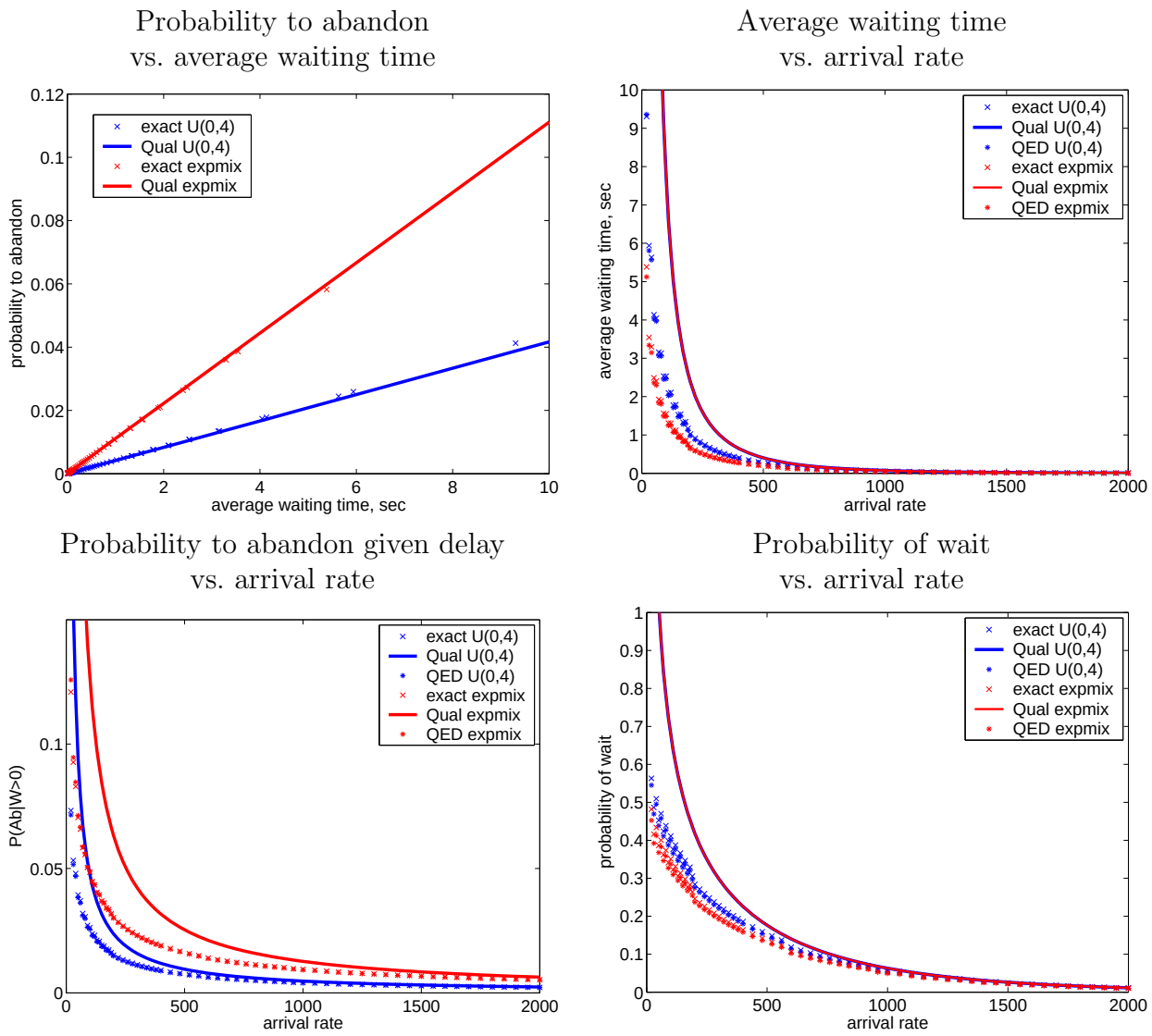


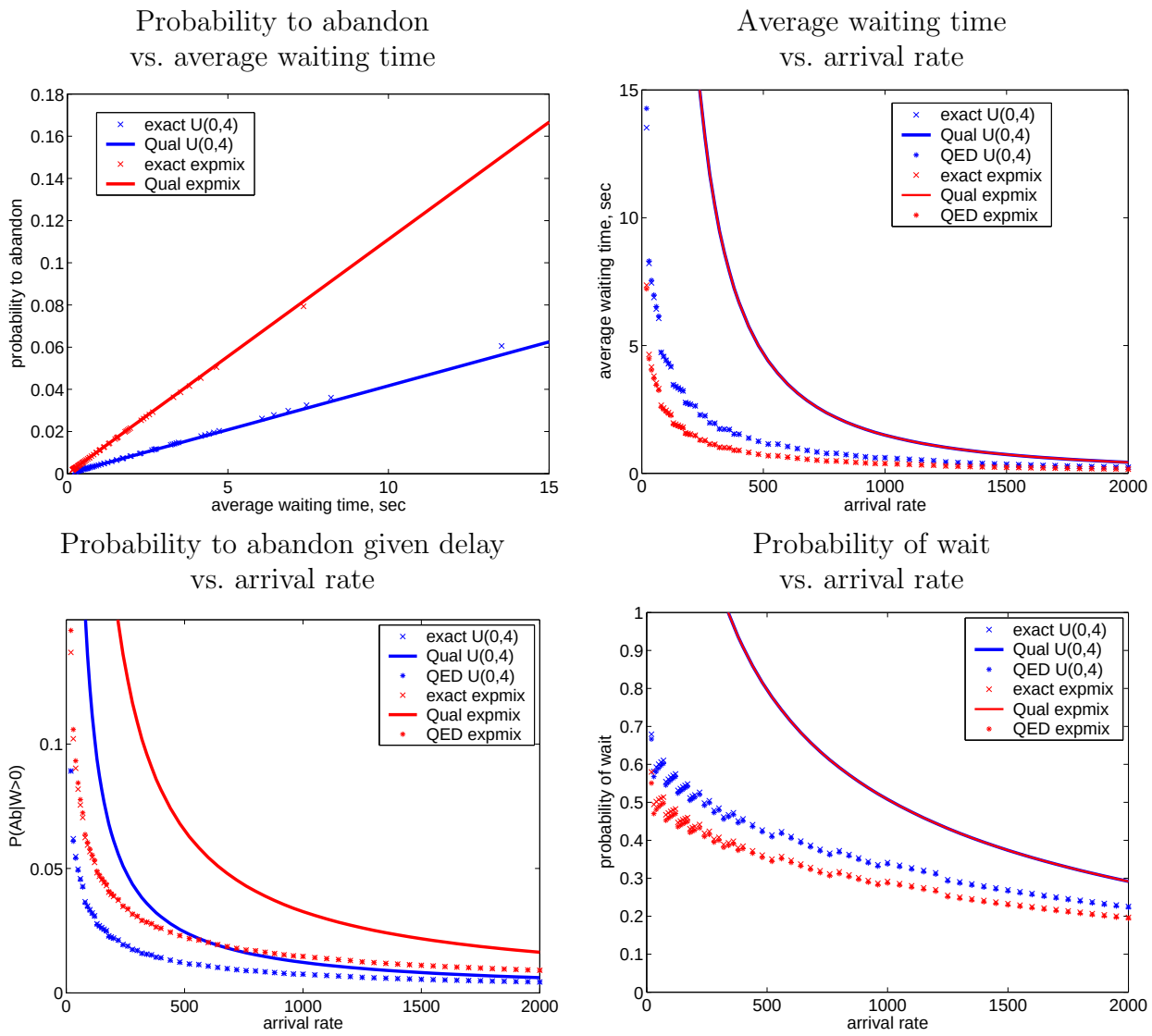
Figure 40: **Offered load per server $\rho = 0.95$, performance measures and approximations**



Example 3 (Figure 40): $\rho = 0.95$.

The quality-driven approximations are good only for $n \geq 500$ (uniform distribution) or $n \geq 1000$ (hyperexponential distribution). The QED approximations are excellent.

Figure 41: **Offered load per server $\rho = 0.98$, performance measures and approximations**



Example 4 (Figure 41): $\rho = 0.98$.

We clearly observe that the QED approximation should be used in this case. Note that the linear P{Ab} / E[W] relation still prevails.

General conclusions.

It is reasonable to use the QD approximations, instead of QED, if the values of the performance measures (probabilities of wait and abandonment, average wait) are small.

For a "rule of thumb", one can take the probability of wait to be less than 0.1 (or even 0.05). The linear relation $\mathrm{P}\{\mathrm{Ab}\} = g_0 \cdot \mathrm{E}[W]$ prevails for all the special cases considered here. The reason is that this relation is asymptotically true both in the QD and the QED regimes.

16.3 Proofs of the QD results

Proof of Lemma 16.1.

a. Lemma 9.1 with

$$m = 0, \quad k_1 = 1, \quad l_1 = 1, \quad k_2 = 1, \quad l_2 = 2,$$

implies that

$$J_A \triangleq \int_0^\infty \exp\left\{-\lambda\gamma x - f(\lambda)\mu x - \frac{\lambda g_0 x^2}{2}\right\} dx$$

$$= \frac{1}{\lambda\gamma + f(\lambda)\mu} - \frac{1}{\lambda^2}\,\frac{g_0}{\gamma^3} + o\left(\frac{1}{\lambda^2}\right) = \frac{1}{n\mu - \lambda} - \frac{1}{\lambda^2}\,\frac{g_0}{\gamma^3} + o\left(\frac{1}{\lambda^2}\right), \tag{16.12}$$

where the relation $n\mu - \lambda = \lambda\gamma + f(\lambda)\mu$ follows from the staffing rule (16.1). (Recall that $f(\lambda)$ denotes the deviation term $o(\sqrt{\lambda})$ from (16.1).)

Standard Laplace arguments from Lemma 15.1 ensure that

$$J = \int_0^\infty \exp\left\{\int_0^x \left[\lambda(\bar{G}(u) - 1)\right] du - \lambda\gamma x - f(\lambda)\mu x\right\} dx$$

can be substituted into (16.12) instead of J_A.

b. The proof is very similar to part **a**. Note that

$$J_1 = \int_0^\infty x \cdot \exp\left\{\int_0^x \lambda(\bar{G}(u) - 1) du - \lambda\gamma x - f(\lambda)\mu x\right\} dx\,.$$

The Laplace method validates the approximation of J_1 by

$$J_{1A} \triangleq \int_0^\infty x \cdot \exp\left\{-\lambda\gamma x - f(\lambda)\mu x - \frac{\lambda g_0 x^2}{2}\right\} dx = \frac{1}{(n\mu - \lambda)^2} - \frac{3g_0}{\lambda^3\gamma^4} + o\left(\frac{1}{\lambda^3}\right).$$

c. Recall that

$$\mathcal{E} = \int_0^\infty e^{-t}\left(1 + \frac{t\mu}{\lambda}\right)^{\frac{\lambda}{\mu}(1+\gamma)+f(\lambda)-1} dt \tag{16.13}$$

$$= \lambda \int_0^\infty e^{-\lambda x}(1 + \mu x)^{\frac{\lambda}{\mu}(1+\gamma)+f(\lambda)-1} dx$$

We perform the change of variables $y = x - \frac{\gamma}{\mu}$. (If we do not take the "$f(\lambda) - 1$" term in the power into account, the expression under the integral (16.13) reaches a maximum at γ/μ.)

$$\begin{aligned} \mathcal{E} &= \lambda \exp\left\{-\lambda\frac{\gamma}{\mu}\right\} \cdot \int_{-\gamma/\mu}^{\infty} e^{-\lambda y}(1+\gamma+\mu y)^{\frac{\lambda}{\mu}(1+\gamma)+f(\lambda)-1} dy \qquad (16.14) \\ &= \lambda \exp\left\{-\lambda\frac{\gamma}{\mu}\right\} \cdot \int_{-\gamma/\mu}^{\infty} \exp\left\{-\lambda y + \left[\frac{\lambda}{\mu}(1+\gamma) + f(\lambda) - 1\right] \cdot \ln(1+\gamma+\mu y)\right\} dy \end{aligned}$$

The Taylor expansion of the logarithm function implies

$$\ln(1+\gamma+\mu y) = \ln(1+\gamma) + \frac{\mu y}{1+\gamma} - \frac{\mu^2 y^2}{2(1+\gamma)^2} + O(y^3) \qquad (y \to 0)$$

We approximate $\mathcal{E}$, replacing the logarithm in (16.14) by the first three terms of the expansion above and changing integral limits to $\int_{-\infty}^{\infty}$:

$$\mathcal{E}_A = \lambda \exp\left\{-\lambda\frac{\gamma}{\mu}\right\} \cdot (1+\gamma)^{n-1} \cdot \int_{-\infty}^{\infty} \exp\left\{-\lambda y + \left[\frac{\lambda}{\mu}(1+\gamma) + f(\lambda) - 1\right] \cdot \left[\frac{\mu y}{1+\gamma} - \frac{\mu^2 y^2}{2(1+\gamma)^2}\right]\right\} dy$$

$$= \lambda \exp\left\{-\lambda\frac{\gamma}{\mu}\right\} \cdot (1+\gamma)^{n-1} \cdot \int_{-\infty}^{\infty} \exp\left\{(f(\lambda)-1) \cdot \left[\frac{\mu y}{1+\gamma} - \frac{\mu^2 y^2}{2(1+\gamma)^2}\right] - \frac{\lambda\mu y^2}{2(1+\gamma)}\right\} dy\,.$$

The last term in the exponent above determines the asymptotic value of the integral. For example, using $f(\lambda) = o(\sqrt{\lambda})$

$$\begin{aligned} &\int_{-\infty}^{\infty} \exp\left\{f(\lambda) \cdot \frac{\mu y}{1+\gamma} - \frac{\lambda\mu y^2}{2(1+\gamma)}\right\} dy \\ &= \int_{-\infty}^{\infty} \exp\left\{-\frac{\lambda\mu}{2(1+\gamma)} \cdot \left[y - \frac{f(\lambda)}{\lambda}\right]^2 + \frac{\mu(f(\lambda))^2}{2\lambda(1+\gamma)}\right\} dy \\ &\sim \sqrt{\frac{2\pi}{\lambda}}\sqrt{\frac{1+\gamma}{\mu}} \qquad (\lambda \to \infty) \end{aligned}$$

Therefore, taking into account $n \sim \frac{\lambda}{\mu} \cdot (1+\gamma), \quad \lambda \to \infty,$

$$\mathcal{E}_A \sim \sqrt{2\pi n} \cdot (1+\gamma)^{n-1} \cdot \exp\left\{-\frac{\lambda\gamma}{\mu}\right\}.$$

In order to validate the same result for $\mathcal{E}$ we apply the Laplace argument, based on the following inequality: $\forall\, \epsilon > 0 \quad \exists\, \delta > 0$ such that

$$\left|\ln(1+\mu y+\gamma) - \ln(1+\gamma) - \frac{\mu y}{1+\gamma} + \frac{\mu^2 y^2}{2(1+\gamma)^2}\right| \le \frac{\epsilon\mu^2 y^2}{2(1+\gamma)^2} \quad \text{for } y \in [-\delta, \delta].$$

Proof of Theorem 16.1. In most cases, the proof is a straightforward application of the formulae in Lemma 16.1.

a. Note that

$$\frac{1}{n\mu-\lambda} \;\sim\; \frac{1}{\lambda\gamma}\,.$$

Then

$$\mathrm{P}\{W>0\} \;=\; \frac{\lambda J}{\mathcal{E}+\lambda J} \;\sim\; \frac{\lambda J}{\mathcal{E}} \;\sim\; \frac{1}{\sqrt{2\pi n}}\cdot\frac{1}{\gamma}\cdot\left(\frac{1}{1+\gamma}\right)^{n-1}\cdot\exp\left\{\frac{\lambda\gamma}{\mu}\right\}.$$

b.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;\sim\; \frac{(n\mu-\lambda)g_0}{\lambda^3\gamma^3 J} \;\sim\; \frac{g_0}{\lambda^2\gamma^2 J} \;\sim\; \frac{g_0}{\lambda\gamma}\,.$$

Recall that $\lambda\sim n\mu\rho$ and $\gamma\sim\dfrac{1-\rho}{\rho}$, $(n\to\infty)$. This implies

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;\sim\; \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{g_0}{\mu}\,.$$

c.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim\; \frac{1}{n\mu-\lambda} \;\sim\; \frac{1}{\lambda\gamma} \;\sim\; \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{1}{\mu}\,.$$

d. The proof is similar to part **d** of Lemma 15.1.

e. A direct consequence of parts **b-d**.

f. The proof can be given along the lines of Theorem 15.1, part **g**. Sketch of the calculations is given by:

$$\frac{\mathrm{P}\left\{W>\dfrac{t}{n\mu}\right\}}{\mathrm{P}\{W>0\}} \;\sim\; \frac{\int_{t/(n\mu)}^{\infty}\exp\{-\lambda\gamma x\}dx}{\int_0^{\infty}\exp\{-\lambda\gamma x\}dx} \;\sim\; \exp\left\{\frac{-\lambda\gamma t}{n\mu}\right\} \;\sim\; e^{-\rho\gamma t} \;\sim\; e^{-(1-\rho)t}\,.$$

17 Efficiency-Driven operational regime

17.1 Formulation of results

In the Efficiency-Driven (ED) operational regime, staffing is determined by:

$$n=\frac{\lambda}{\mu}\cdot(1-\gamma)+o(\sqrt{\lambda})\,,\qquad \gamma>0\,. \tag{17.1}$$

The offered load per agent

$$\rho \;=\; \frac{\lambda}{n\mu} \;\to\; \frac{1}{1-\gamma} \;>\; 1\,.$$

If $o(\sqrt{\lambda}) \equiv 0$ in (17.1), the relation between ρ and γ is given by

$$\rho = \frac{1}{1-\gamma} \quad \text{and} \quad \gamma = \frac{\rho-1}{\rho}\,. \tag{17.2}$$

Lemma 17.1 (Building blocks) Assume that the equation

$$G(x) \;=\; \gamma$$

has a unique solution x^* and that the patience density at x^* is positive: $g(x^*) > 0$. Then

a.

$$J \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}} \cdot \exp\{\lambda k(\gamma)\}\,, \tag{17.3}$$

where

$$k(\gamma) \;\triangleq\; x^* \cdot \left(1 - \frac{n\mu}{\lambda}\right) - \int_0^{x^*} G(u)du\,. \tag{17.4}$$

(Note that the definition of x^* implies that $k(\gamma) > 0$ for λ large enough.)

b.

$$\mathcal{E} \;\sim\; \frac{1}{\gamma}\,. \tag{17.5}$$

c.

$$J_1 \;\sim\; x^* \cdot J \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}} \cdot x^* \cdot \exp\{\lambda k(\gamma)\}\,. \tag{17.6}$$

Theorem 17.1 (Performance measures) Under the assumptions of Lemma 17.1, the performance measures of the M/M/n+G queue in the efficiency-driven operational regime can be approximated by:

a. Probability to get service immediately decreases exponentially:

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; \frac{1}{\gamma} \cdot \sqrt{\frac{g(x^*)}{2\pi\lambda}} \cdot \exp\{-\lambda k(\gamma)\}\,. \tag{17.7}$$

b. Probability to abandon converges to the constant $\gamma \approx 1 - \dfrac{1}{\rho}$.

$$\mathrm{P}\{\mathrm{Ab}\} \;\sim\; \gamma\,. \tag{17.8}$$

c. The average offered wait $\mathrm{E}[V]$ converges to the constant x^*.

$$\mathrm{E}[V] \;\sim\; x^* \,. \tag{17.9}$$

The offered wait also converges to x^* in probability:

$$V \;\xrightarrow{p}\; x^* \,. \tag{17.10}$$

d. Define the distribution $G^* = \{G^*(x),\ x \geq 0\}$ by

$$G^*(x) \;=\; \begin{cases} \frac{G(x)}{G(x^*)} \;=\; \frac{G(x)}{\gamma}, & x \leq x^* \\ 1, & x > x^* \end{cases}$$

(In fact, G^* is the distribution of the random variable $\min(x^*, \tau)$, where τ is the patience time.)

Then the average waiting time W weakly converges to the distribution G^*:

$$W \;\xrightarrow{w}\; G^* \,. \tag{17.11}$$

In addition,

$$\mathrm{E}[W] \;\to\; \mathrm{E}[\min(x^*, \tau)] \;=\; \int_0^{x^*} \bar{G}(u)du \,. \tag{17.12}$$

e. Total Service Factor.

The distribution of wait is given by:

$$\mathrm{P}\{V > t\} \;\sim\; \begin{cases} 1, & t < x^* \\ 0, & t > x^* \end{cases} \,. \tag{17.13}$$

$$\mathrm{P}\{W > t\} \;\sim\; \begin{cases} \bar{G}(t), & t < x^* \\ 0, & t > x^* \end{cases} \,. \tag{17.14}$$

The distribution of wait around x^* can be approximated in the following way.

Let $-\infty < t < \infty$. Then

$$\mathrm{P}\left\{ \frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;\sim\; \bar{\Phi}\left(t\sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) \,. \tag{17.15}$$

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;\sim\; (1-\gamma) \cdot \bar{\Phi}\left(t\sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) \,. \tag{17.16}$$

Remark 17.1 Limits for the probability to abandon and waiting time can be obtained using "fluid" (deterministic) considerations and are sometimes referred to as *fluid limits*. For example, see results in Whitt [72] that are closely related to (17.8) and (17.12).

Remark 17.2 Assume that the staffing level (17.1) is kept exact: $n = (\lambda/\mu) \cdot (1 - \gamma)$. Then we can rewrite definition (17.4) as

$$k(\gamma) \;=\; \gamma x^* - \int_0^{x^*} G(u)du\,.$$

17.2 Numerical Experiments

Three distributions that were considered above in Subsection 15.2 are used in our experiments: uniform, hyperexponential and delayed exponential. Instead of the conditional probability $\mathrm{P}\{\mathrm{Ab}|W > 0\}$, we plot $\mathrm{P}\{\mathrm{Ab}\}$ (the probability of wait is close to one and there are no reasons to distinguish between the two performance measures). Note that, in contrast to the QED regime, the ED approximation formulae are the same for distributions with both positive and zero densities at the origin. (Although the rate of convergence of the approximations can be very different for the two types of distributions.) As in Subsection 16.2, we compare between the ED and QED approximations.

The efficiency-driven staffing rule is

$$n = \left\lceil \frac{\lambda}{\rho\mu} \right\rceil , \qquad \rho \geq 1\,. \tag{17.17}$$

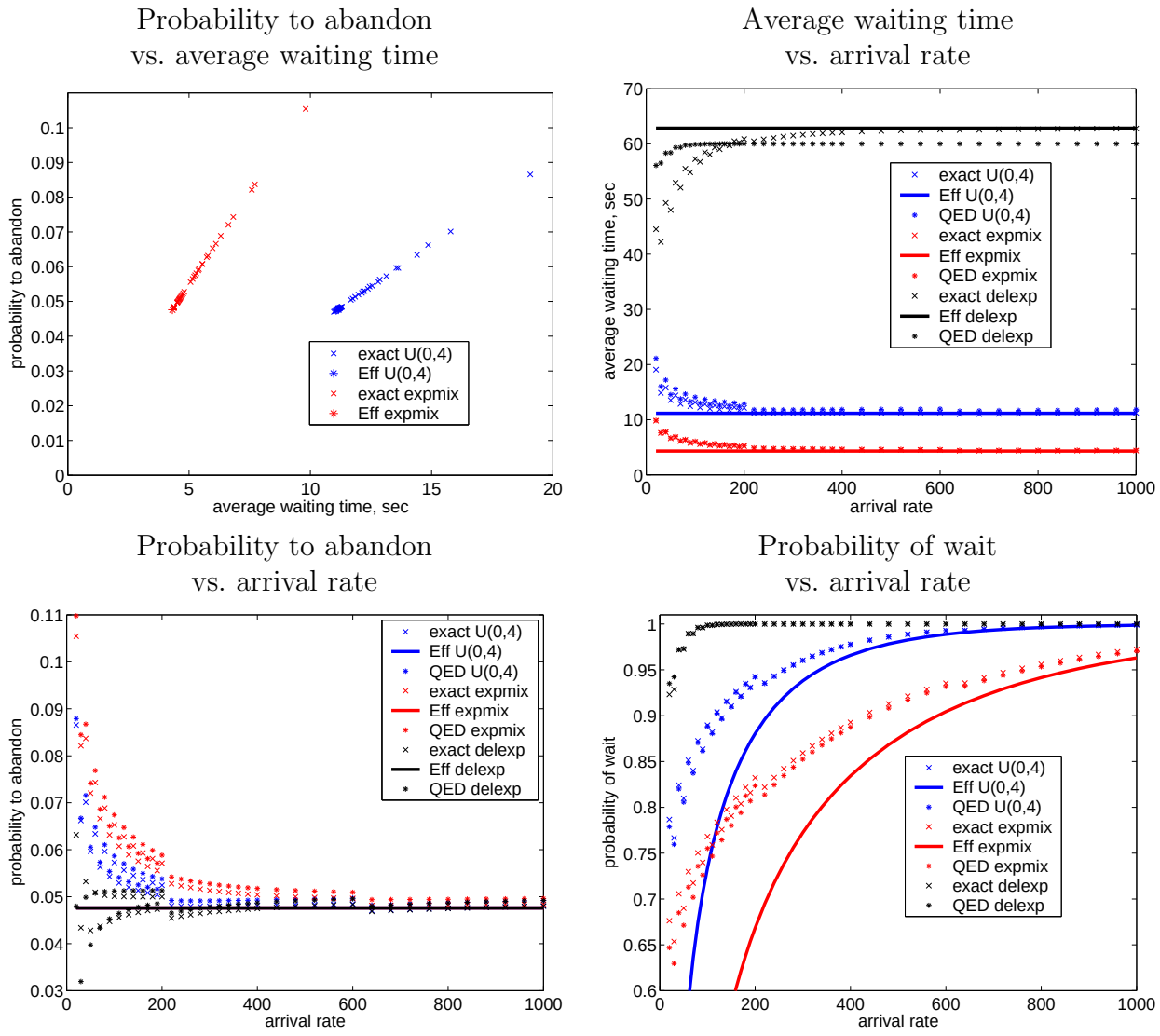
Four values of ρ: 1.05, 1.1, 1.2 and 1.5 were chosen. Other assumptions are the same as in Subsection 15.2.

Example 1 (Figure 42): $\rho = 1.05$.

- We observe that the probability to abandon and the average wait converge to fluid limits. The limit for the probability to abandon is $\gamma = 1 - \dfrac{1}{\rho}$, independently of the patience distribution. The limit for the average wait (17.12) depends on the specific patience-time distribution.

- Surprisingly, for the uniform and hyperexponential distributions we again observe a linear $\mathrm{P}\{\mathrm{Ab}\}/\mathrm{E}[W]$ relation. (The curve for delayed exponential distribution

Figure 42: **Offered load per server $\rho = 1.05$, performance measures and approximations**



does not provide any remarkable pattern and is not plotted.) This relation prevails mainly for small values of λ. Unlike the QED and the QD regimes, points of the scatterplot do not converge to the origin when $\lambda \to \infty$. In addition, there exist no theoretical support for a linear relation if $\lambda \to \infty$. In contrast, one can check that if $n = \frac{\lambda}{\rho\mu}$ prevails exactly, P{Ab} converges to the fluid limit exponentially in λ, and E[W] converges at a polynomial rate. We think that the observed linearity for small λ is due mostly to a rounding effect of the staffing level in formula (17.17). If

n exceeds $\frac{\lambda}{\rho\mu}$, than P{Ab} and E[W] decrease together. Otherwise, they increase. If we choose ρ such that $\dfrac{\lambda}{\rho\mu}$ is an integer number, a linear pattern is not observed.

- The QED approximations are better than the efficiency-driven for small values of λ. However, we observe that QED approximations for P{Ab} and E[W] do not always converge to fluid limits (see the delayed exponential distribution for average wait).

- The situation is somewhat different for the approximation of the probability-of-wait. Here the QED approximation is better almost everywhere. (The ED approximation for delayed exponential has been found bad for all ρ values, therefore it is not plotted.)

Example 2 (Figure 43): $\boldsymbol{\rho = 1.1}$.

- Linear pattern of the relation P{Ab} / E[W] still prevails for small λ.

- In this case, almost all QED approximations for P{Ab} and E[W] do not converge to fluid limits (although, for very small λ they can be superior to ED.) In fact, the QED approximations for P{Ab} are close to $-\beta/\sqrt{n}$, which converges to $\sqrt{\rho}\cdot(1-1/\rho)$. It differs from the proper fluid limit by a factor $\sqrt{\rho}$.

- The quality of the ED approximation for P{$W > 0$} improves, but QED is still better for small values of λ.

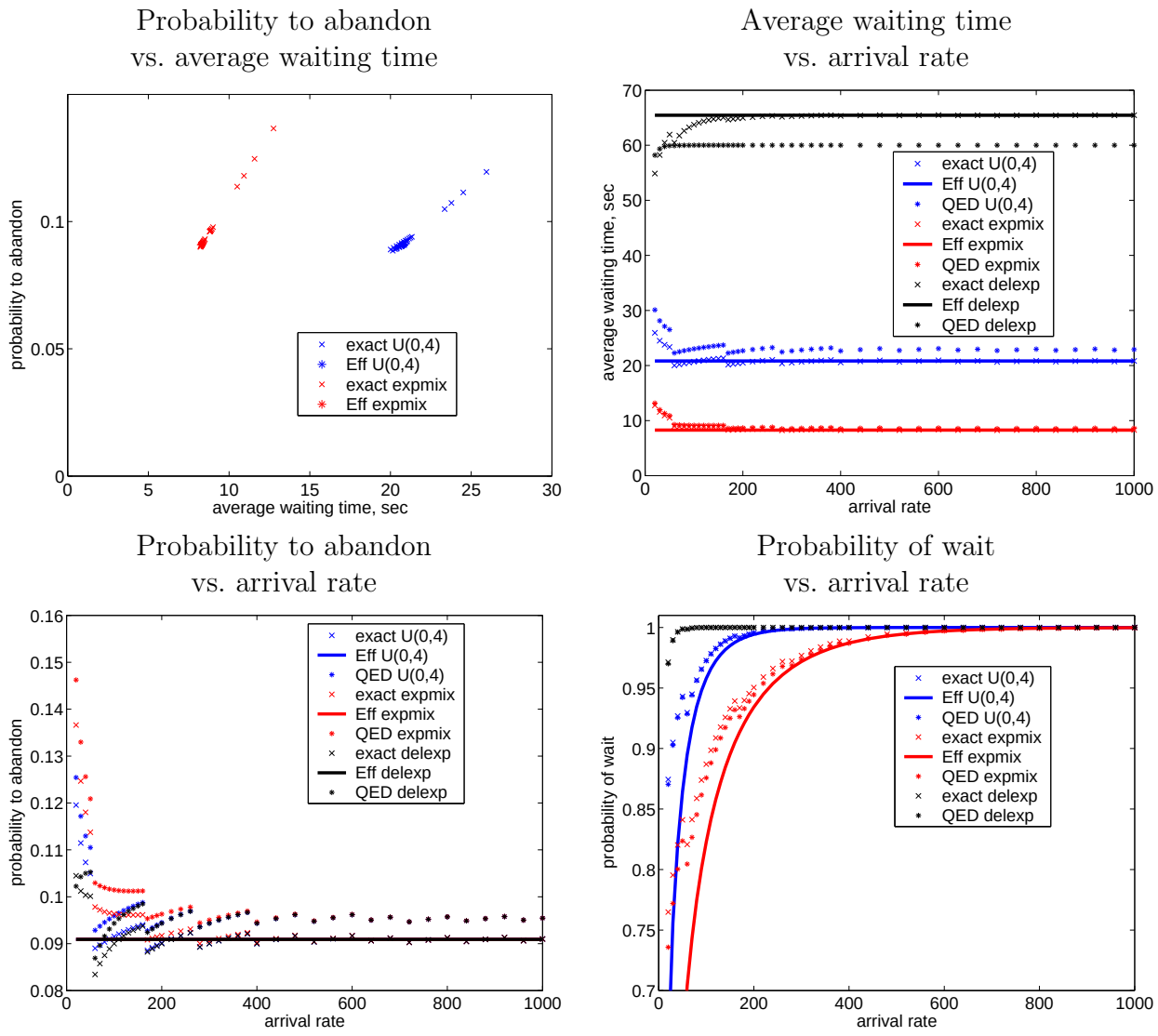
Example 3 (Figure 44): $\boldsymbol{\rho = 1.2}$.

- The convergence rate of P{Ab} and E[W] to the fluid limits increases. QED approximations are bad starting from small λ. Note also that the QED approximations for P{Ab} are indistinguishable for the three distributions.

- Linear pattern of the first plot is much less convincing than in Examples 1 and 2.

- ED approximations for the probability of wait are good now.

Figure 43: **Offered load per server $\rho = 1.1$, performance measures and approximations**



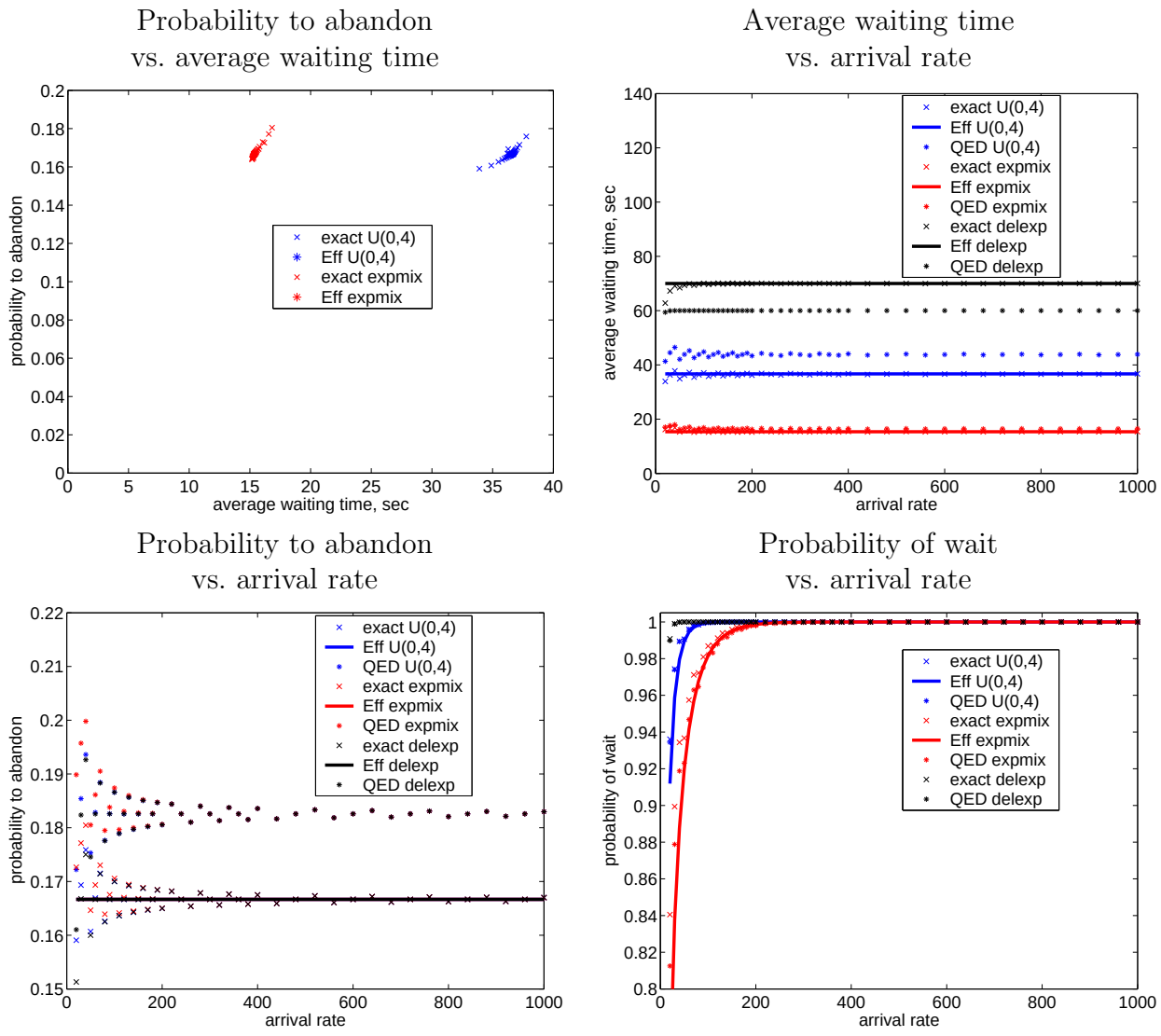
Example 4 (Figure 45): $\rho = 1.5$.

Here the trends observed for Figure 44 are even more pronounced.

General conclusions.

The ED approximations for P{Ab} and E[W] are better than the QED ones if the offered load per agent ρ is significantly larger than 1 (say, $\rho \geq 1.2$). Even if ρ is closer to one (1.1 or 1.05), they are appropriate for large values of λ. In general, comparing with Subsection

Figure 44: **Offered load per server $\rho = 1.2$, performance measures and approximations**

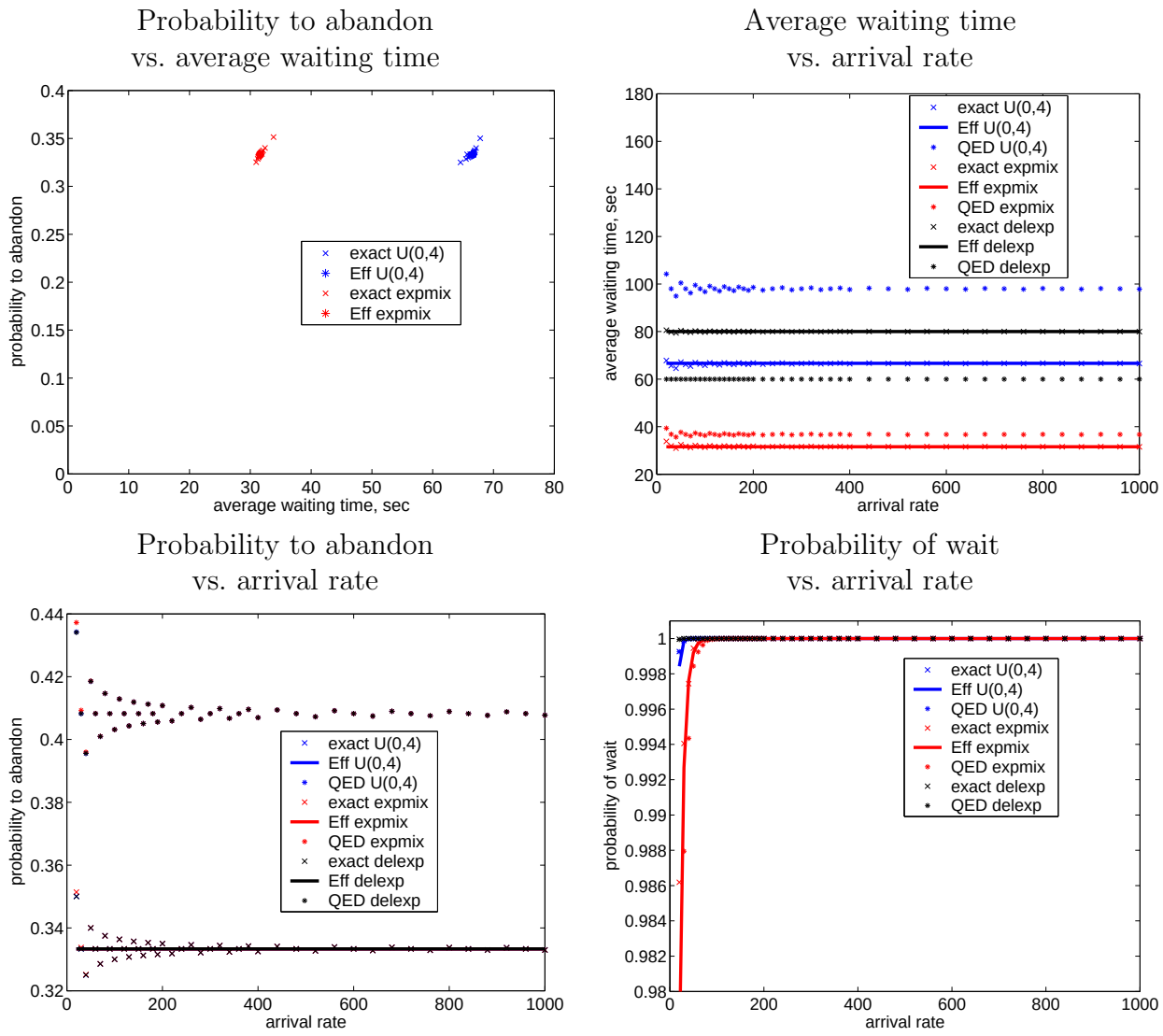


16.2, the ED approximations are "competing" with QED more successfully than the QD approximations.

We observe also that the ED fluid-limit approximations for $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$ are better than the ED approximation for $\mathrm{P}\{W > 0\}$.

A linear pattern, displayed at some of $\mathrm{P}\{\mathrm{Ab}\}$ / $\mathrm{E}[W]$ plots (with ρ close to one, and relatively small values of λ) should be also mentioned. It adds support to the QED and QD results, in order to explain why this pattern is so widespread in the call center

Figure 45: **Offered load per server $\rho = 1.5$, performance measures and approximations**



environment.

17.3 Proofs of the ED results

Proof of Lemma 17.1.

a. In the efficiency-driven operational regime,

$$J \;=\; \int_0^\infty \exp\left\{\lambda\gamma x - f(\lambda)\mu x + \lambda\int_0^x [\bar{G}(u)-1]du\right\} dx\,.$$

It is straightforward to verify that the function

$$h_\lambda(x) \;=\; \lambda\gamma x + \lambda\int_0^x [\bar{G}(u)-1]du$$

reaches a maximum at x^*. Changing variables: $y = x - x^*$, we get

$$J \;=\; \exp\{x^*\cdot(\lambda-n\mu)\}\cdot\int_{-x^*}^{\infty}\exp\left\{\lambda\gamma y - f(\lambda)\mu y + \lambda\int_0^{y+x^*}[\bar{G}(u)-1]du\right\}dx\,. \quad (17.18)$$

The three leading terms of the Taylor expansion for $\int_0^{y+x^*}[\bar{G}(u)-1]du$ at $y=0$ are given by

$$\int_0^{y+x^*}[\bar{G}(u)-1]du \;=\; \int_0^{x^*}[\bar{G}(u)-1]du - \gamma y - \frac{1}{2}g(x^*)y^2 + O(y^3)\,.$$

Hence, $\forall\epsilon>0\;\;\exists\delta>0$ such that for $|y|<\delta$

$$\int_0^{x^*}[\bar{G}(u)-1]du - \frac{1}{2}g(x^*+\epsilon)y^2 \;\le\; \gamma y + \int_0^{y+x^*}[\bar{G}(u)-1]du$$

$$\le\; \int_0^{x^*}[\bar{G}(u)-1]du - \frac{1}{2}g(x^*-\epsilon)y^2\,. \quad (17.19)$$

Define

$$J_A \;=\; \exp\{\lambda k(\gamma)\}\cdot\int_{-\infty}^{\infty}\exp\left\{-f(\lambda)\mu y - \frac{\lambda g(x^*)y^2}{2}\right\}dy\,.$$

For $f(\lambda)=o(\sqrt{\lambda})$,

$$J_A \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}}\cdot\exp\{\lambda k(\gamma)\}\,.$$

Now we perform the Laplace argument, based on inequalities (17.19), obtaining

$$J \;\sim\; J_A\,.$$

The equivalence of the $\int_{-\delta}^{\delta}$ integrals is derived via the Taylor expansion (recall Lemma 15.1, part **a**). In addition, we must construct "exponential bounds" in the spirit of formula (15.112) for

$$\int_\delta^\infty \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_{x^*}^{x^*+y}G(u)du\right\}dy \quad (17.20)$$

and the corresponding integral $\int_{-\infty}^{-\delta}$.

Define

$$\alpha \;=\; G\left(x^*+\frac{\delta}{2}\right) - \gamma \;>\; 0\,.$$

(The value of α is positive since G is strictly increasing at x^*.) Then the integral in (17.20) is less or equal to

$$\int_\delta^\infty \exp\left\{-f(\lambda)\mu y - \lambda\alpha\left(y-\frac{\delta}{2}\right)\right\} dy \;\le\; \exp\left\{\frac{\lambda\alpha\delta}{2}\right\}\cdot\int_\delta^\infty \exp\left\{-\frac{3}{4}\lambda\alpha y\right\} dy \;=$$

$$= \frac{\exp\{-(\lambda\alpha\delta)/4\}}{(3/4)\lambda\alpha} \;=\; o(e^{-\nu\lambda})\,, \qquad \nu>0\,.$$

The bound for $\int_{-\infty}^{-\delta}$ is constructed in a similar way.

b. First we present an approximation

$$\mathcal{E}_A \;=\; \lambda\int_0^\infty \exp\left\{-\lambda\gamma x + \mu(f(\lambda)-1)x\right\} dx \;\sim\; \frac{1}{\gamma}\,. \tag{17.21}$$

Then

$$\begin{aligned}\mathcal{E} \;&=\; \lambda\int_0^\infty e^{-\lambda x}(1+\mu x)^{\frac{\lambda}{\mu}(1-\gamma)+f(\lambda)-1}dx\\ &=\; \lambda\int_0^\infty \exp\left\{-\lambda x + \left[\frac{\lambda}{\mu}(1-\gamma)+f(\lambda)-1\right]\cdot\ln(1+\mu x)\right\} dx\end{aligned}$$

Now using $\ln(1+\mu x)=\mu x+o(x)$ and the Laplace argument from Lemma 15.1, we get that

$$\mathcal{E} \;\sim\; \mathcal{E}_A\,.$$

c. In the ED regime,

$$\begin{aligned}J_1 \;&=\; \int_0^\infty x\cdot\exp\left\{\lambda\gamma x - f(\lambda)\mu x + \lambda\int_0^x[\bar G(u)-1]du\right\} dx\\ &=\; e^{x^*\cdot(\lambda-n\mu)}\cdot\int_{-x^*}^\infty (y+x^*)\cdot\exp\left\{\lambda\gamma y - f(\lambda)\mu y + \lambda\int_0^{y+x^*}[\bar G(u)-1]du\right\} dx\\ &=\; x^*\cdot J + e^{x^*\cdot(\lambda-n\mu)}\cdot\int_{-x^*}^\infty x^*\cdot\exp\left\{\lambda\gamma y - f(\lambda)\mu y + \lambda\int_0^{y+x^*}[\bar G(u)-1]du\right\} dx\,.\end{aligned} \tag{17.22}$$

Using Taylor expansion of $\int_0^{y+x^*}[\bar G(u)-1]du$ (see part **a** of the proof), we get that the second term of (17.22) has a smaller order than the first one, given $\lambda\to\infty$.

Proof of Theorem 17.1.

a. Probability to get service immediately.

$$\mathrm{P}\{\mathrm{W}=0\} \;\sim\; \frac{\mathcal{E}}{\mathcal{E}+\lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; \frac{1}{\gamma}\cdot\sqrt{\frac{g(x^*)}{2\pi\lambda}}\cdot\exp\{-\lambda k(\gamma)\}\,.$$

b. Probability to abandon.

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;\sim\; \frac{\lambda-n\mu}{\lambda} \;\sim\; \gamma\,.$$

c. Offered waiting time.

$$E[V] \;=\; \frac{J_1}{J} \;\sim\; x^*\,.$$

In order to prove $V \xrightarrow{p} x^*$, we must derive

$$\int_{x^*+\delta}^{\infty} v_\lambda(x)dx \;\to\; 0 \quad \text{and} \quad \int_0^{x^*-\delta} v_\lambda(x)dx \;\to\; 0\,.$$

Both statements can be proved using "exponential bounds" (see the proof of (17.20)).

d. We have shown that

$$V_\lambda \xrightarrow{w} x^*\,.$$

Hence, the pair (V_λ, τ) converges weakly to (x^*, τ), as a two-dimensional random vector. Since the minimum function is continuous, the virtual waiting time

$$W_\lambda \;=\; \min(x^*, \tau) \;\xrightarrow{w}\; \min(x^*, \tau)\,.$$

In order to prove convergence of expectations, it is sufficient to demonstrate uniform integrability of $\{W_\lambda,\ \lambda \geq 0\}$. Since $W_\lambda \leq V_\lambda$, the uniform integrability can be shown for $\{V_\lambda,\ \lambda \geq 0\}$. The proof follows the pattern of proving (17.20). For example,

$$\lim_{\lambda\to\infty} \int_{x^*+\delta}^{\infty} x\tilde{v}_\lambda(x)dx \;=\; 0\,.$$

e. Formulae (17.13) and (17.14) follow from parts **c** and **d**. The proof of (17.15) proceeds via

$$P\left\{V > x^* + \frac{t}{\sqrt{n}} \,\middle|\, V > 0\right\} \;=\; \frac{\int_{x^*+t/\sqrt{n}}^{\infty} \exp\left\{\lambda\gamma x - f(\lambda)\mu x - \lambda\int_0^x G(u)du\right\} dx}{\int_0^{\infty} \exp\left\{\lambda\gamma x - f(\lambda)\mu x - \lambda\int_0^x G(u)du\right\} dx}$$

$$= \frac{\int_{t/\sqrt{n}}^{\infty} \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_0^{y+x^*} G(u)du\right\} dy}{\int_{-x^*}^{\infty} \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_0^{y+x^*} G(u)du\right\} dx} \;\sim\; \frac{\int_{t/\sqrt{n}}^{\infty} \exp\left\{-\frac{\lambda g(x^*)y^2}{2}\right\}}{\int_{-\infty}^{\infty} \exp\left\{-\frac{\lambda g(x^*)y^2}{2}\right\} dy} \qquad (17.23)$$

$$= \bar{\Phi}\left(t\sqrt{\frac{\lambda g(x^*)}{n}}\right) \;=\; \bar{\Phi}\left(t\sqrt{\frac{g(x^*)\mu}{(1-\gamma)}}\right)\,.$$

(The equivalence in (17.23) can be proved using the methods from Lemma 17.1, part **a**.) Then

$$\mathrm{P}\left\{ \frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;\sim\; \bar{\Phi}\left(t\sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) .$$

Analyzing the actual waiting time,

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;=\; \bar{G}\left(x^* + \frac{t}{\mu\sqrt{n}} \right) \cdot \mathrm{P}\left\{ \frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\}$$

$$\sim\; (1-\gamma) \cdot \bar{\Phi}\left(t\sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) .$$

18 Economies of scale in the M/M/n+G queue

Consider m iid call centers that are pooled into a single operation. Each call center can be modelled by an M/M/n+G queue with the same characteristics: arrival rate λ, service rate μ, patience distribution G, and n servers, where n is determined by the manager of call center.

Assume that all these call centers were run in one of the operational regimes studied in Sections 15-17. If we sustain that regime in the pooled call center, how will performance change? Will Economies Of Scale (EOS) drive improvements in service level? Tables 3-5 summarize answers to these questions.

In Gans et al. [29] an EOS framework for the Erlang-C queue was developed. In this section, we compare between Erlang-C and M/M/n+G, observing many similar EOS effects. That is somewhat surprising, taking into account significant difference between the two models.

18.1 QED regime

Recall that in the QED operational regime, staffing level is determined by

$$n \;=\; [R + \beta\sqrt{R}] ,$$

where R is the offered load. In Erlang-C , β must be positive, but in M/M/n+G, $-\infty < \beta < \infty$.

Define the *safety staffing* Δ as the difference between the staffing level n and the offered load $R = \lambda/\mu$. Again, for the Erlang-C queue we need $\Delta > 0$ to ensure system stability. In models with abandonment, Δ can go negative.

Table 3: **Economies of scale. QED regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \dfrac{\lambda}{\mu}$	mR	$R = \dfrac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	$\sqrt{m}\Delta$	$-\infty < \Delta < \infty$	$\sqrt{m}\Delta$
Number of agents	$R + \Delta$	$mR + \sqrt{m}\Delta$	$R + \Delta$	$mR + \sqrt{m}\Delta$
Service grade	$\beta = \dfrac{\Delta}{\sqrt{R}}$	$\boxed{\beta}$	$\beta = \dfrac{\Delta}{\sqrt{R}}$	$\boxed{\beta}$
P{W>0}	$\left[1 + \dfrac{\beta}{h(-\beta)}\right]^{-1}$	$\boxed{\text{P\{W>0\}}}$	$\left[1 + \dfrac{h(r\beta)}{rh(-\beta)}\right]^{-1}$	$\boxed{\text{P\{W>0\}}}$
Occupancy	$\dfrac{R}{R+\Delta}$	$\dfrac{R}{R+\frac{\Delta}{\sqrt{m}}}$	$\dfrac{R}{R+\Delta} \cdot (1 - \text{P\{Ab\}})$	$\dfrac{R}{R+\frac{\Delta}{\sqrt{m}}} \cdot \left(1 - \dfrac{\text{P\{Ab\}}}{\sqrt{m}}\right)$
P{Ab|W> 0}	—	—	$\dfrac{\beta}{\Delta r} \cdot (h(r\beta) - r\beta)$	$\dfrac{1}{\sqrt{m}} \cdot \text{P\{Ab\|W> 0\}}$
ASA	$\dfrac{1}{\Delta}$	$\dfrac{1}{\sqrt{m}} \cdot \text{ASA}$	$\dfrac{r\beta}{\Delta} \cdot (h(r\beta) - r\beta)$	$\dfrac{1}{\sqrt{m}} \cdot \text{ASA}$
TSF	$e^{-\beta t}$	$(\text{TSF})^{\sqrt{m}}$	$\dfrac{\bar{\Phi}\left(r\beta + \frac{t}{r}\right)}{\bar{\Phi}(r\beta)}$	$\text{TSF} \cdot \dfrac{\bar{\Phi}\left(r\beta + \frac{t}{r}\sqrt{m}\right)}{\bar{\Phi}\left(r\beta + \frac{t}{r}\right)}$

In order to get simple expressions that are straightforward to compare across regimes, we modify the definitions of average wait (ASA) and Total Service Factor (TSF). Both performance measures will be calculated only for delayed customers and they are measured in units of the average service time.

Formally,

$$\text{ASA} \triangleq \text{E}\left[\frac{W}{\text{E}(S)} \;\middle|\; W > 0\right], \tag{18.1}$$

and

$$\text{TSF} \triangleq \text{P}\left\{ \frac{W}{\text{E}(S)} > \frac{t}{\sqrt{n}} \,\middle|\, W > 0 \right\}$$

Definition (18.1) will be the same for the three regimes. In contrast, the definition of TSF will be modified for each special case.

Table 3 illustrates Economies of Scale for the main case of the QED regime (positive patience density at the origin). We make some changes in notation, in comparison to Theorem 15.1, defining

$$r \triangleq \sqrt{\frac{\mu}{g_0}}\,.$$

It will turn out that in each of the three regimes, one or several performance measures are held constant after pooling. The boxed entries in Table 3, as well as in later tables, highlight those performance measures. Indeed, in the QED case, the probability of wait remains fixed under pooling. In addition, we observe that in both queues the agents' occupancy converges to 100% and ASA decreases to ASA/$\sqrt{m}$. Finally, the probability to abandon the queue decreases at rate $1/\sqrt{m}$.

18.2 QD regime

Recall that the Quality-Driven operational regime of M/M/n+G is characterized by:

$$n = R \cdot (1 + \gamma)\,,$$

where the service grade γ is positive.

The definition of TSF in the QD regime is taken to be

$$\text{TSF} \triangleq \text{P}\left\{ \frac{W}{\text{E}(S)} > \frac{t}{n} \,\middle|\, W > 0 \right\}\,.$$

Since abandonment is exponentially negligible in the QD regime, the M/M/n+G performance measures, presented in Table 4, are identical to the Erlang-C case: ASA decreases to ASA/m, TSF decreases to TSF^m and the probability of wait converges to zero exponentially. In addition, the conditional probability to abandon in M/M/n+G decreases at rate $1/n$.

Table 4: **Economies of scale. QD regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \dfrac{\lambda}{\mu}$	mR	$R = \dfrac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	$m\Delta$	$\Delta > 0$	$m\Delta$
Number of agents	$n = R + \Delta$	$mR + m\Delta$	$n = R + \Delta$	$mR + m\Delta$
Service grade	$\gamma = \dfrac{\Delta}{R}$	$\boxed{\gamma}$	$\gamma = \dfrac{\Delta}{R}$	$\boxed{\gamma}$
P$\{W > 0\}$	$\dfrac{1}{\sqrt{2\pi n}} \cdot \dfrac{(\rho e^{1-\rho})^n}{1-\rho}$	$\dfrac{1}{\sqrt{m}} \cdot (\text{P}\{W > 0\})^m$	$\dfrac{1}{\sqrt{2\pi n}} \cdot \dfrac{(\rho e^{1-\rho})^n}{1-\rho}$	$\dfrac{1}{\sqrt{m}} \cdot (\text{P}\{W > 0\})^m$
Occupancy	$\dfrac{1}{1+\gamma}$	$\boxed{\dfrac{1}{1+\gamma}}$	$\dfrac{1}{1+\gamma}$	$\boxed{\dfrac{1}{1+\gamma}}$
P$\{$Ab$\vert$W$> 0\}$	—	—	$\dfrac{1}{n} \cdot \dfrac{1}{1-\rho} \cdot \dfrac{g_0}{\mu}$	$\dfrac{1}{m} \cdot \text{P}\{\text{Ab}\vert W > 0\}$
ASA	$\dfrac{1}{n} \cdot \dfrac{1}{1-\rho}$	$\dfrac{1}{m} \cdot \text{ASA}$	$\dfrac{1}{n} \cdot \dfrac{1}{1-\rho}$	$\dfrac{1}{m} \cdot \text{ASA}$
TSF	$e^{-(1-\rho)t}$	$(\text{TSF})^m$	$e^{-(1-\rho)t}$	$(\text{TSF})^m$

18.3 ED regime

Recall that the definitions of the ED regime for Erlang-C and M/M/n+G are different:

$$n \;=\; R + \gamma$$

for Erlang-C (assume that n is integer), and

$$n \;=\; R \cdot (1 - \gamma)\,, \qquad \gamma > 0\,,$$

for M/M/n+G.

Let

$$\text{TSF} \;\triangleq\; \text{P}\left\{ \frac{W}{\text{E}(S)} > t \;\middle|\; W > 0 \right\}.$$

Table 5: **Economies of scale. ED regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \frac{\lambda}{\mu}$	mR	$R = \frac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	Δ	$\Delta < 0$	$m\Delta$
Number of agents	$n = R + \Delta$	$mR + \Delta$	$n = R + \Delta$	$mR + m\Delta$
Service grade	$\gamma = \Delta$	$\boxed{\gamma}$	$\gamma = -\frac{\Delta}{R}$	$\boxed{\gamma}$
P$\{W > 0\}$	1	1	1	1
Occupancy	1	1	1	1
P{Ab}	—	—	$1 - \frac{1}{\rho}$	$\boxed{1 - \frac{1}{\rho}}$
ASA	$\frac{1}{\Delta}$	$\boxed{\text{ASA}}$	$\frac{\text{E}[\min(R, x^*)]}{\text{E}(S)}$	$\boxed{\text{ASA}}$
TSF	$e^{-t\Delta}$	$\boxed{\text{TSF}}$	$\begin{cases} \bar{G}(t/\text{E}(S)), & t < x^*/\text{E}(S) \\ 0, & t > x^*/\text{E}(S) \end{cases}$	$\boxed{\text{TSF}}$

The ED results are summarized in Table 5. In both queues, essentially all customers are delayed and agents are nearly 100% utilized. The waiting time remains asymptotically the same after pooling (both the mean and distribution). In addition, from Section 17 the probability to abandon in the M/M/n+G queue converges to the fluid limit $1 - 1/\rho$.

18.4 Economies of Scale: main conclusions

Each operational regime corresponds to one or several performance measures that are held constant under pooling. Therefore, the following rules for the M/M/n+G queue can be deduced:

- If a call center manager would like, after pooling, to maintain agents utilization at a constant level, smaller than 100%, the QD operational regime is appropriate. It implies very high performance level, and essentially all customers get service immediately.

- If the objective is to fix the probability of wait, the QED operational regime should be used. It will combine high performance level (ASA, TSF, probability to abandon) and agents' utilization that is not far from 100%.

- If it is enough to sustain the probability to abandon or/and waiting time, the understaffed ED operational regime enables this goal.

Finally, we observed close relations between EOS effects in the simple Erlang-C system and the much more complicated M/M/n+G.

19 Some statistical applications to call centers

19.1 General description of the data set

The source of our data is a large multi-site call center of a US bank. It has sites in New York, Pennsylvania, Rhode Island, and Massachusetts. The daily volume on a regular day is up to 300,000 calls overall. The majority of these calls end at the VRU, but up to 70,000 are seeking to reach agents. Only the latter will be considered here.

The number of agent positions at peak hours varies from 900-1200 on weekdays to 200-500 on weekends. Working hours are 24 hours a day, 7 days a week.

The call center provides many service types. In our research, we consider two of them. The first one, **Retail**, is by far the most common. The second, **Telesales**, is the most common after Retail, together with Business and Consumer Loans.

Call-by-call data was collected from March 2001 to October 2003. Our sample is taken from the five-month period between September 2002 and January 2003. Since service patterns during weekends are different, we analyze regular days only (Monday-Friday), considering calls that arrive between 7am to 24pm.

Below in Table 6 we provide overall descriptive statistics for the two service types under consideration:

Table 6: **Retail and Telesales service types. Descriptive statistics**

September 2002 - January 2003

	Calls	E[S]	P{$W > 0$}	P{Ab}	E[W]
Retail	3,451,743	224.6 sec	30.6%	1.16%	6.33 sec
Telesales	349,371	453.9 sec	24.3%	1.76%	9.66 sec

We observe that, overall, the system seems to work in the QED regime: the probability of wait is neither close to zero not to one, the probability to abandon and average wait are small. However, there are huge difference between the M/M/n+G model and a large multi-site call center.

One of the most important differences lies in the protocol of customers' service. When a call arrives to the call center, it is sent to agents of a specific site. Only if a call is not served within a deadline (around 10 seconds for Retail, different numbers for other types), it can be sent to agents from other sites. This protocol violates work-conservation assumption: waiting customers and available agents can easily co-exist.

In our work with this data, we used the Data-Mocca software [65], developed in the Statistics Laboratory at the Technion.

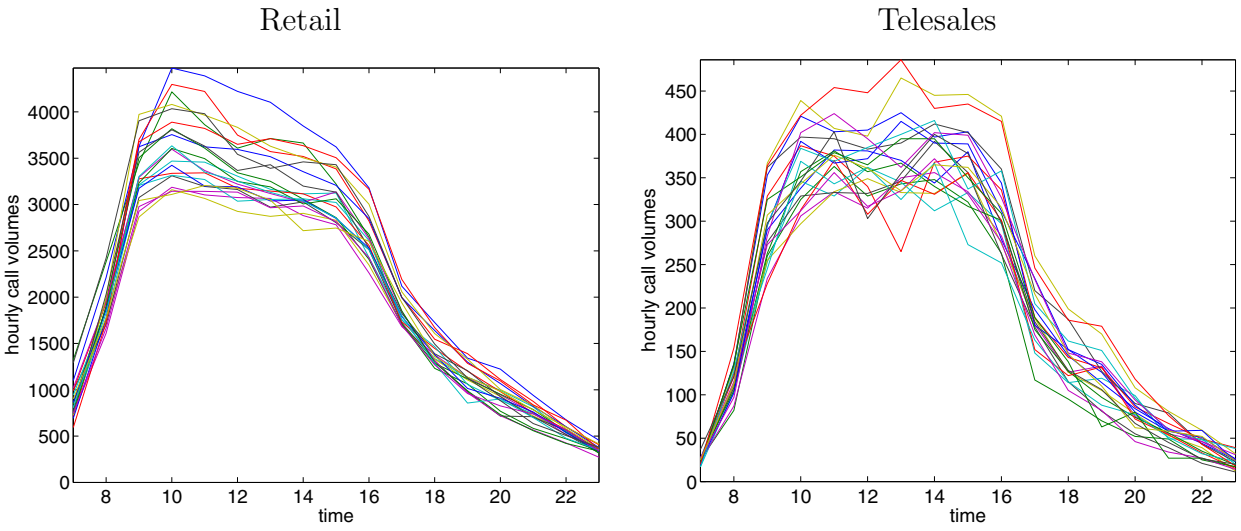
19.2 Model primitives

In order to apply the M/M/n+G model, reliable data for its parameters should be obtained. First, we need an hourly data for λ, μ and n. Second, we must calculate estimates of patience distributions that are appropriate for our methods.

Arrivals. Hourly arrival rates are calculated from call-by-call data. Figure 46 presents them for the 21 January weekdays. Note that we observe a rather stable daily pattern.

Average service times. Hourly average service times were calculated similarly. Wrap-up time (after-call work) was taken into account.

Figure 46: **Hourly arrival rates, January 2003.**



Number of agents. Unfortunately, detailed data on agents is not available in our database. The following estimation algorithm was developed in order to try and fill this gap.

Let $H(t)$ denote number of calls being served at time t (where wrap-up also means that the call is being served). The value of $H(t)$ can be estimated using Data-Mocca with resolution 1 minute. The values of $H(t)$ and the number of customers in queue $Q(t)$ constitute the input of the algorithm. The output is the estimate of the staffing level at time t, denoted by $N(t)$.

1. Use the following *forward* recursive algorithm:
$N_f(7:00) = H(7:00)$;
if $Q(t) > 0$, $N_f(t) = H(t)$ (work-conservation);
if $Q(t) = 0$, $N_f(t) = \max(N_f(t-1), H(t))$.
This algorithm has so far a drawback: If the staffing level decreases and there is no queue (say, at the end of the day), the algorithm will not "notice" the decrease. Therefore, we use also:

2. *Backward* recursive algorithm.
$N_b(24:00) = H(24:00)$;

if $Q(t) > 0$, $N_b(t) = H(t)$ (work-conservation);
if $Q(t) = 0$, $N_b(t) = \max(N_b(t+1), H(t))$.

3. Take the *average* of the last two estimates:

$$N(t) = \frac{N_b(t) + N_f(t)}{2}, \qquad t = 7:00, 7:01, \ldots, 24:00.$$

However our numerical experiments showed that the estimate thus derived underestimates the actual staffing level. The reason could be that the work-conservation assumption is not true in a large multi-site call center, where waiting customers and available agents can co-exist.

4. Divide the working day to three-minutes intervals. Assume that the staffing level is constant for those intervals and take the estimate for each interval to be equal to *maximum* of 3 adjacent estimates from Step 4. This heuristic step is designed to compensate the underestimation, referred to above.

5. Finally, compute *hourly* staffing estimates by averaging out 20 three-minute estimates from Step 5.

Patience times.

Figure 47: **Hazard rates of patience.**

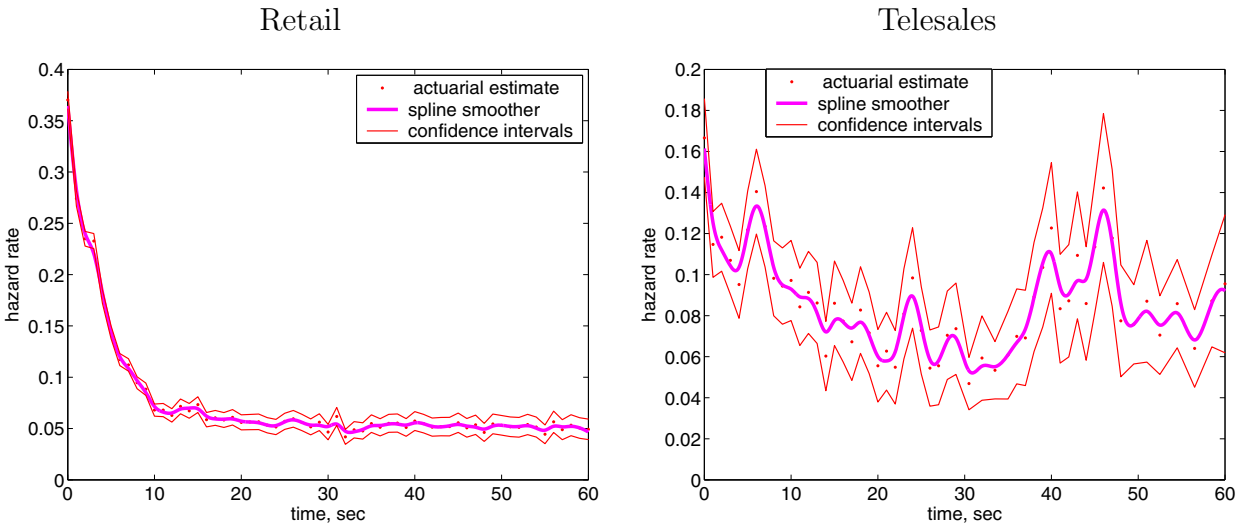
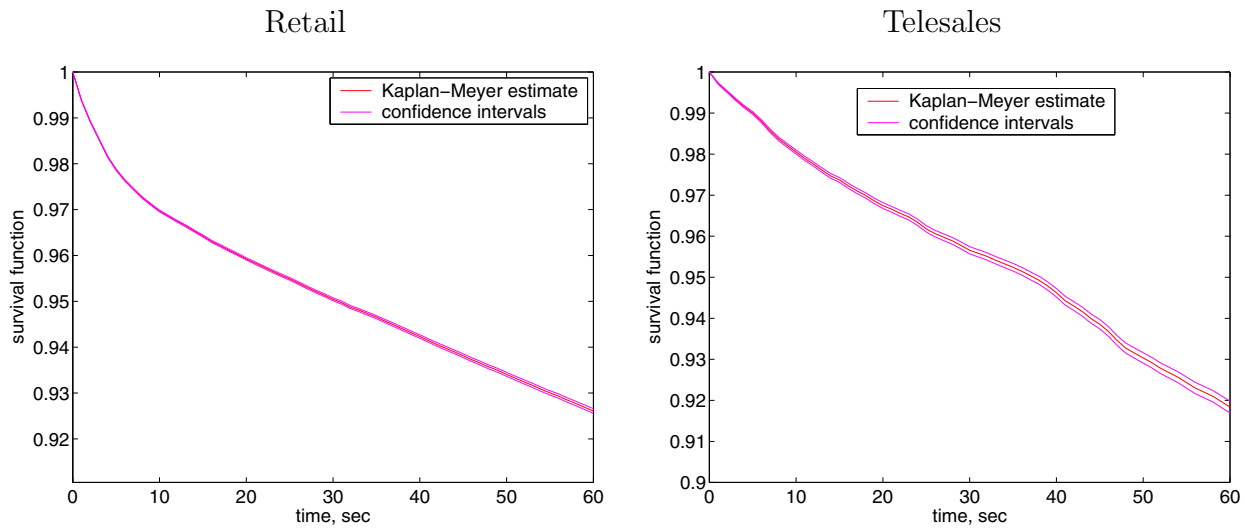


Figure 48: **Survival functions of patience. Kaplan-Meier estimate.**



We produced estimates of the patience hazard and survival-function, based on overall call-by-call data from 5 months, as mentioned before; actuarial estimator, described in Subsection 8.4, was used. Intervals $[a_{j-1}, a_j)$, $j \geq 1$, were chosen in a way, that ensured at least 30 abandonments during each interval.

Figure 47 demonstrates very unstable hazard pattern near the origin, especially for the Retail service type. Figure 48 shows that customers are, overall, very patient: over 90% are willing to wait more than one minute.

We checked that the monthly patience patterns are indeed stable over the five-month period under consideration. However, for months out of this period, the patience hazard function can be very different. A probable reason could be changes in the contents of announcements and their timing. (Recall the second plot of Figure 5, where announcements took place at 15 and 60 seconds of customers' wait, implying peaks of abandonment.)

19.3 Performance measures

Three basic performance measures are considered in this section: $\mathrm{P}\{W > 0\}$, $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$. Below we discuss several issues related to their measurement.

Probability of wait. It turns out that the database contains a very large fraction of waiting times that equal one second. (Around half of the observations! In addition, about 20% of waiting times equal zero.) Since it is unreasonable to assume that 50% of customers experienced *actual positive* wait of one second, the event $\{W = 0\}$ was defined to be equivalent to a wait of 0 or 1 second in the database.

Probability to abandon. Abandonments that took place at 0 or 1 second were discarded. Their meaning is unclear; probably they correspond to customers that decided to leave even before they were sent to queue or service (e.g. at the VRU stage).

Waiting times. Since the event "a customer was served immediately" is equivalent to the wait of 0 or 1 seconds in the database, all waiting times exceeding zero are reduced by 1 second.

19.4 Relation between P{Ab} and E[W]

Figures 49 and 50 display an empirical P{Ab}/E[W] relationship, based on the data of 1649 hours for our five-months period. The aggregated plots were obtained similarly to Figure 7 in the Introduction: 40 points aggregate 41 hours each (except the last point, which aggregates 50 hours). In the process of aggregation, the weight of specific hours was assumed equal to the corresponding hourly number of customers.

We observe a linear curve for Telesales. However, for Retail the pattern is concave with a clear change around 4 seconds of average wait. This phenomenon can be explained by the hazard rates, displayed in Figure 47. The pattern of the Retail hazard rate deviates from the constant exponential hazard more than the Telesales hazard rate. Therefore, the linear relation, which prevails theoretically in the exponential case, is observed practically for Telesales, but not for Retail.

19.5 Fitting QED approximations

Our main approach is the following. First, we estimate the number of agents n via fitting one of our basic performance measures (the probability to abandon was chosen since it

Figure 49: **Retail customers. Probability to abandon vs. average waiting time**

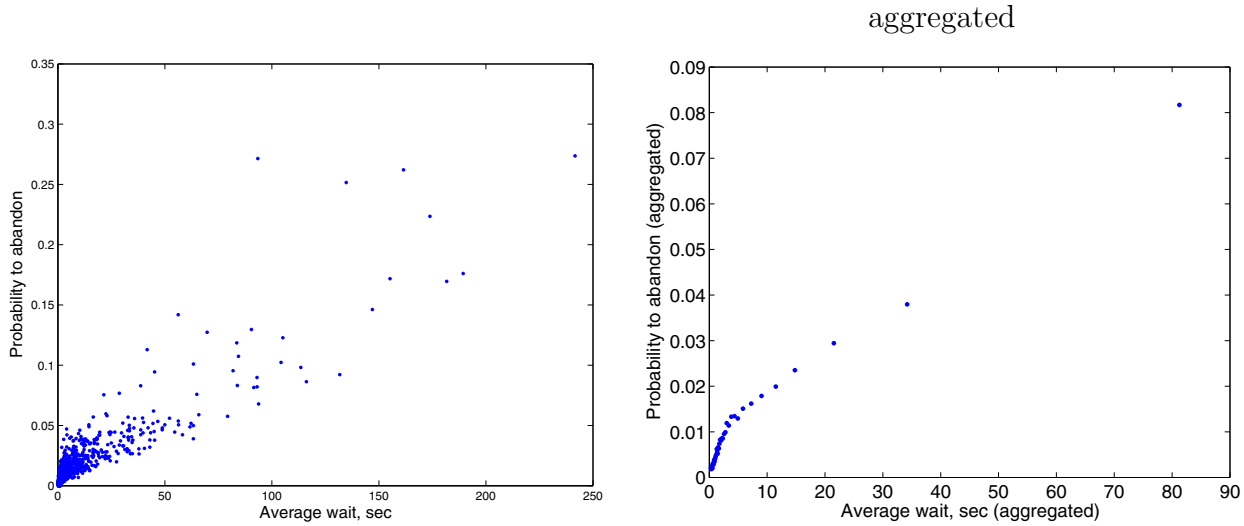
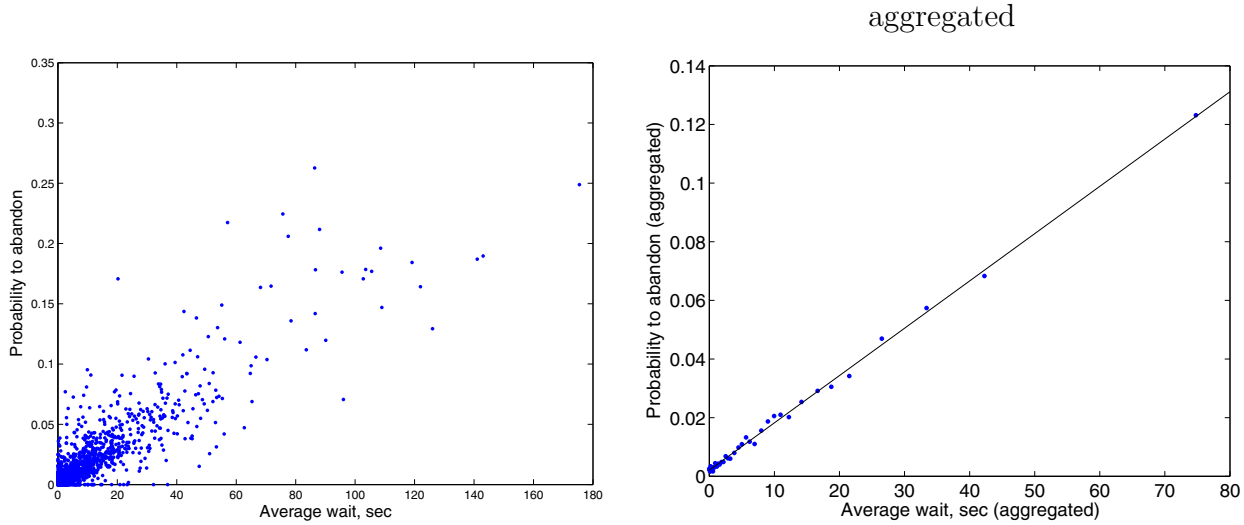


Figure 50: **Telesales customers. Probability to abandon vs. average waiting time**



performed the best). Specifically, we numerically solve the following equations, based on the hourly data:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; f(\lambda, \mu, n, g_0)\,,$$

where n is unknown, f is the formula for the QED estimate from Theorem 15.1, λ and μ are hourly arrival and service rates, respectively, and g_0 is the patience density at the

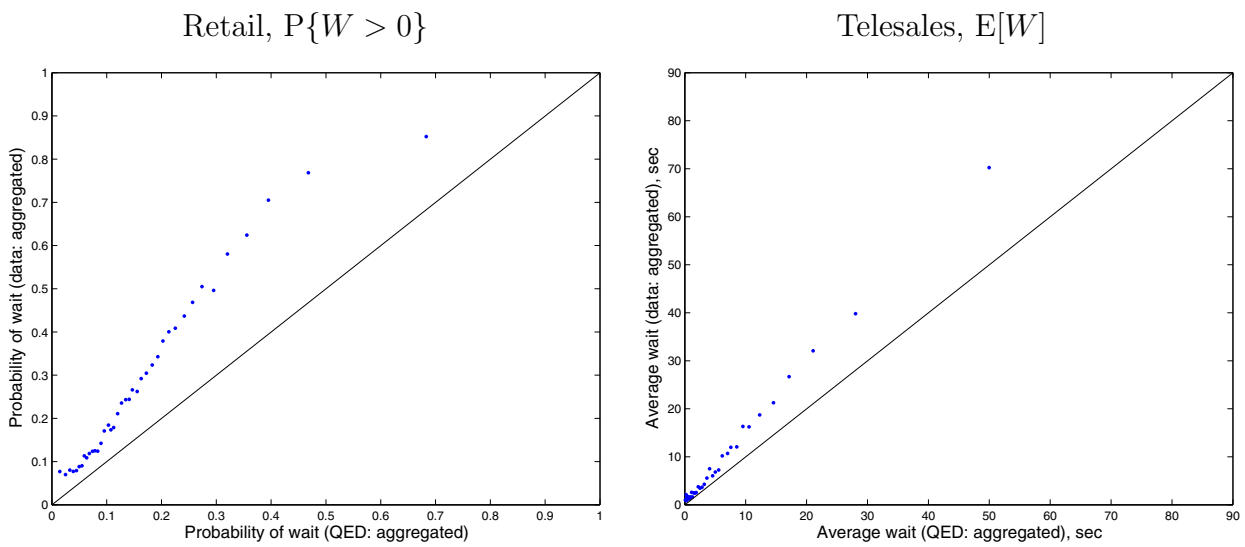
origin.

Then, using this estimate of n, we try to fit other performance measures. Since the estimate of n depends on QED formulae, such an experiment cannot "prove" that QED approximations fit the data. However, a negative result would show that some problems exist in our approach.

An additional important question arises: which value should be substituted into the QED formulae for g_0? The most straightforward way is to substitute the hazard estimate at the origin from Figure 47. Figure 51, which is plotted along the guidelines of Figures 12 and 13 from Section 6, shows the results of this experiment. We observe a very strong bias between data values and model values. In our opinion, the reason for the bias is instability of the two hazard estimates from Figure 47 near the origin.

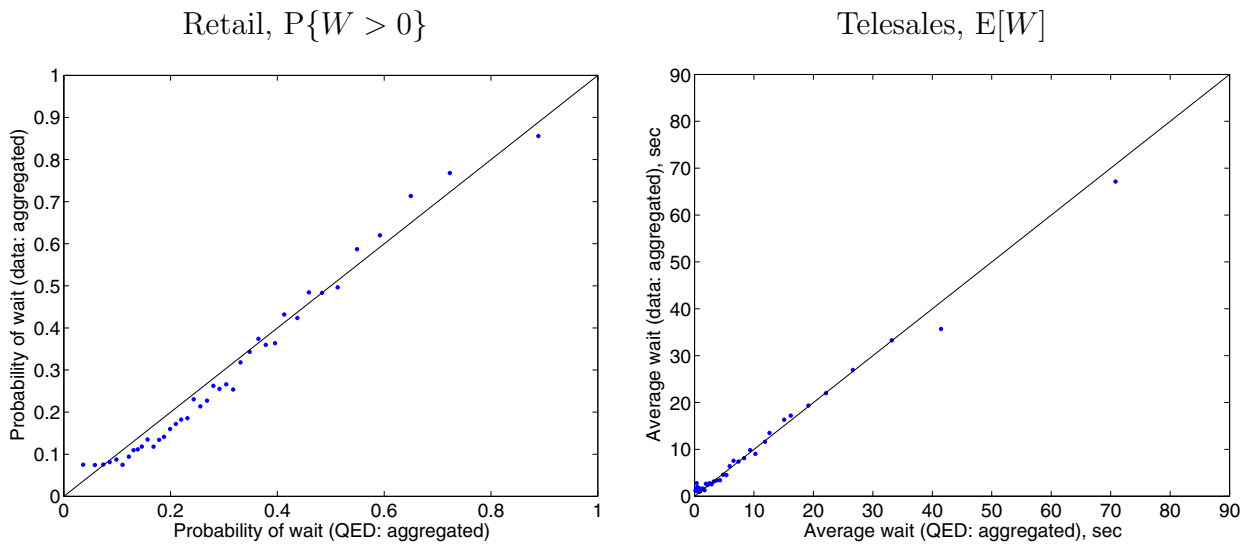
Specifically, the limit statements from Theorem 15.1 prevail in practice, if the patience density (or hazard rate) is more or less stable for typical values of waiting times. In our case, a typical wait is equal to several seconds. (In the QED limit, the wait converges to zero.) However, since the hazard rate oscillates significantly even within the range of several seconds, the limit QED statements do not apply directly.

Figure 51: **Fitting performance measures, g_0:=hazard at zero**



As an answer to this challenge, we suggest to substitute for g_0 the value of the ratio P{Ab}/E[W] into the QED formulae. (See Figure 52.) Now the fit for some performance measures is good. It seems that the value of the ratio gives an appropriate weighted average of the hazard rate near the origin.

Figure 52: **Fitting performance measures, g_0:= P{Ab}/E[W]**



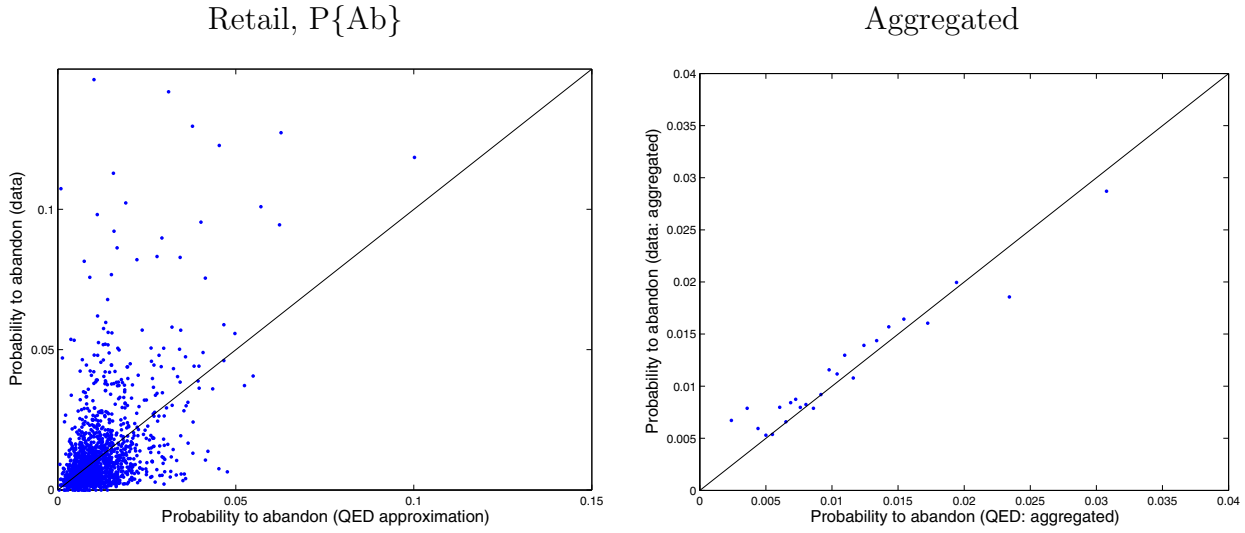
Independent estimate of the number of agents We also performed experiment with the independent estimates of n, presented in Subsection 19.2. Again, the ratio P{Ab}/E[W] was substituted into the QED formulae, instead of g_0. Although, the results were very noisy, a reasonable fit was obtained in some cases. See Figure 53, for example.

19.6 Summary of our data analysis

In some of our experiments, we observed a good fit to the theoretical models. For example, recall Figures 50 and 52. However, several problems and challenges arise that do not enable us to characterize this data research as definite success. These problems that deserve further attention are as follows:

- During data collection, the detailed profiles of agents should be added to call-by-call data. This will enable reliable estimates for the number of agents.

Figure 53: **Fitting performance measures, independent estimate of the number of agents**



- The influence of the volatile customers' behavior during the first few seconds of their wait should be explored. Probably, in addition to the approach from Subsection 19.5, one could try models with balking.

- The reality of a large modern call center is much more complicated than the M/M/n+G model. For example, due to the service protocol described in Subsection 19.2, the FCFS service discipline or work-conservation do not, in general, prevail. Hence, sometimes more complicated models should be applied.

20 Conclusions of Part IV

Here we summarize our conclusions on the three operational regimes, analyzed in Sections 15-17, and on the relation P{Ab}/E[W] that has been explored in different contexts in this research.

QED regime. In contrast to the exact M/M/n+G formulae, the QED approximations can be applied using any software that provides the standard normal distribution (e.g. Excel). In Subsection 15.2 we observed that these approximations work very well for a

wide range of M/M/n+G parameters.

Our rule-of-thumb recommendations for the use of QED formulae are the following:

- Number of servers n from 10's to 1000's;
- Agents highly utilized but not overloaded (∼90-98%);
- Probability of delay 10-90%;
- Probability to abandon: 3-7% for small n, 1-4% for large n.

Finally, in Section 19, where a large and complicated call center was considered, we established that some performance measures are approximated well by the QED formulae.

ED regime. These approximations are relatively simple to apply as well, although they require solving the equation $G(x) = \gamma$, as well as integration (calculating $H(x^*)$). Both can be performed either numerically or analytically, depending on the patience distribution. We suggest to use the ED approximations if:

- Number of servers $n \geq 100$. (One can cautiously use n=10's, if the probability to abandon is large (>10%).)
- Agents very highly utilized (>95%);
- Probability of delay: more than 85%;
- Probability to abandon: more than 5%.

QD regime. In Subsection 16.2 we observed that, unless wait and abandonment are very small, the QED approximations are preferable over the QD ones. The QD formulae should be applied only if the probability of wait is or should be less than 5-10%. (For example, in emergency call centers.)

Linear P{Ab}/E[W] relation. In the QED and QD operational regimes, the linear relation P{Ab}/E[W] prevails. Such a relation was also observed for Telesales data of our US bank, analyzed in Section 19. Summarizing these facts and those established in Part III, we conclude that this phenomenon prevails in a very broad context: exact M/M/n+G performance measures, different approximations and real data.

Part V
Ongoing and future research

Finally, we outline some directions worthy of further research.

Dimensioning the M/M/n+G queue. In our context, the term *dimensioning* was introduced in Borst, Mandelbaum and Reiman [9]. The authors considered an optimization problem for the Erlang-C queue, where the goal is to minimize the sum of *staffing costs* and *waiting costs*. In other words, [9] developed a formal framework for the problem of the *quality-efficiency tradeoff*, discussed in Subsection 1.1 of the Introduction. It turns that if the staffing costs are of the same order as the waiting costs, the QED operational regime optimally arises. Specifically, if c is the hourly cost of an agent, and a is the hourly cost of customers' delay, then the asymptotic optimal staffing $N^* = R+y^*(a/c)\sqrt{R}$, where R is the offered load, and $y^*(\cdot)$ is a function that is easily computable.

In addition, [9] considered a *constraint satisfaction* problem where one chooses the least number of agents that adheres to a given constraint on waiting cost. It turns out that different types of constraints give rise to the three operational regimes from Part IV.

The ongoing research [53] is dedicated to similar problems for the M/M/n+G queue. In addition to staffing and waiting costs, abandonment costs arise in this case. For a wide set of system parameters, a comparison between the asymptotic QED staffing and exact optimal staffing demonstrates that the two staffing rules are almost identical.

Queues with random arrival rate. In Brown et al. [13] it was shown that the Poisson arrival rate in an Israeli call center varies from day to day and its prediction raises statistical and practical challenges. Therefore, it is very interesting to study queueing models, where the Poisson arrival rate Λ is a random variable. (For example, Jongbloed and Koole [43] concentrate mainly on the case where Λ is Gamma-distributed.)

If $\mathrm{E}(\Lambda) \to \infty$ and its standard deviation is of the order $\sqrt{\mathrm{E}(\Lambda)}$, we expect that the QED operational regime and the square-root staffing rule will arise again. However, if $\sigma(\Lambda)$ is of the order $\mathrm{E}(\Lambda)$, the "cruder" ED regime seems to be the most appropriate; see Whitt [75], and Bassamboo, Harrison and Zeevi [5].

Queues with time-inhomogeneous arrival rates. In the ongoing research of Feldman, Mandelbaum, Massey and Whitt [24], a simulation-based algorithm is developed for staffing time-varying queues with abandonment. The algorithm is designed to achieve a given constant probability of delay, generalizing the QED operational regime to the queues with non-homogeneous arrival rates. This work builds on Jennings et al. [42]

Data analysis. Additional studies of customers' patience in tele-service should be performed. Currently data collection in two large banks in the U.S. and Israel, and in an Israeli cellular-phone company is in process. In addition to call-by-call data, we hope to obtain also reliable staffing data, that was lacking in the research in Section 19.

Generally distributed service times: M/G/n+G. In our research, we assumed exponential service times. However, this assumption seems to not apply for many call centers. For example, both in the US bank, studied in Section 19, and in the Israeli bank [13], the lognormal distribution provides an excellent approximation for service times.

Therefore, it is very important to study the M/G/n+G model with generally distributed service times. However, exact analysis of the M/G/n+G queue seems prohibitively difficult, hence one should probably resort to approximations (see Whitt [70]) and simulation (see Mandelbaum and Schwartz [51]).

M/M/n+G: the queue-length distribution. In this research, we presented approximations for the average wait, average offered wait and, sometimes, for their distributions. Approximation for the average queue can be always calculated from the average wait via Little's formula. However, a different technique should be performed in order to derive QED approximations of the queue-length distribution. Here one could try to use some modification of the Distributional Little's Law [33] or apply the exact M/M/n+G formulae from Brandt and Brandt [11].

Process-limit results for M/M/n+G. In this thesis, we focused on steady-state results, both exact and approximate, for many M/M/n+G performance measures. Of interest are also analogous process-limit results, as in Garnett et al. [29] for Erlang-A.

Extremal properties of the patience distribution. The research conducted in Part III can be continued. For example, we conjecture that, given that average patience is fixed, the deterministic distribution minimizes the abandonment rates α_l (recall Subsection 12.3) for all $l > 0$.

References

[1] Armony M. and Mandelbaum A. (2004) Design, staffing and control of large service systems: The case of a single customer class and multiple server types. Working paper. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 7.2

[2] Asmussen S. (1987) *Applied Probability and Queues*, Wiley. 15.1.2

[3] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: F.J.Kylstra (Ed.), *Performance '81*. North-Holland Publishing Company, 159-179. 5.1, 5.3, 5.3.1, 6.2, 6.3, 6.4, 6.5, 14, 15.2

[4] Bain P. and Taylor P. (2002) Consolidation, "Cowboys" and the developing employment relationship in British, Dutch and US call centres. In: Holtgrewe U., Kerst C. and Shire K. (Ed.), *Re-Organising Service Work*. Ashgate Publishing Limited, 42-62. 1.1

[5] Bassamboo A., Harrison J.M. and Zeevi A. (2004) Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method. Submitted for publication. 7.4, V

[6] Bhattacharya P.P. and Ephremides A. (1991) Stochastic monotonicity properties of multiserver queues with impatient customers. *Journal of Applied Probability*, 28, 673-682. 11.1

[7] Bordoloi S.K. (2004) Agent recruitment planning in knowledge-intensive call centers. *Journal of Service Research*, 6(4), 309-323. 1.1

[8] Bittner S., Schietinger M., Schroth J. and Weinkopf C. (2002) Call Centres in Germany: Employment, Training and Job Design. In: Holtgrewe U., Kerst C. and Shire K. (Ed.), *Re-Organising Service Work*. Ashgate Publishing Limited, 63-85. 1.1

[9] Borst S., Mandelbaum A. and Reiman M. (2004). Dimensioning large call centers. *Operations Research*, 52(1), 17-34. 7.2, V

[10] Boxma O.J. and de Waal P.R. (1994) Multiserver queues with impatient customers. *ITC*, 14, 743-756. 6.2, 12.1, 12.4

[11] Brandt A. and Brandt M. (1999) On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35, 1-18. 5.1, 6.2, 6.4, 6.4, 6.4, 12, 14, 14.2, V

[12] Brandt A. and Brandt M. (2002) Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems: Theory and Applications (QUESTA)*, 41, 73-94. 5.1, 5.2, 6.2, 6.4, 6.4, 6.4, 12.3, 14, 14.2

[13] Brown L.D., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (2002) Statistical Analysis of a Telephone Call Center: A Queueing Science Perspective. To be published in *JASA*. 4.1, 4.1, 4.3, 6.2, 8.1, 8.2, 2, 12.1, 12.1, V, V

[14] Call Center Data (2002) Technion, Israel Institute of Technology. Downloadable from http://iew3.technion.ac.il/serveng/callcenterdata/index.html. 4.1

[15] Choi B.D., Kim B. and Zhu D. (2004) $MAP/M/c$ queue with constant impatient time. *Mathematics of Operations Research*, 29(2), 309-325. 6.2

[16] Cleveland B. and Mayben J. (1997) Call Center Management on Fast Forward. Annapolis: Call Center Press. 1, 8.3

[17] Cox D.R. and Oakes D. (1984) Analysis of Survival Data, Chapman and Hall. 4.1, 8.4

[18] Daley D.J. and Servi L.D. (2000) Estimating customer loss rates from transactional data. In: Shanthikumar J.G. and Sumita U. (ed.): *International Series in Operations Research and Management Science*, 313-332. 6.2

[19] Datamonitor. `http://www.datamonitor.com`. 1.1

[20] de Bruijn N.G. (1981) *Asymptotic Methods in Analysis*, Dover. 5.1, 9.1, 15.3, 15.14

[21] Durrett R. (1991) *Probability: Theory and Examples*, Wadsworth. 10

[22] 4CallCenters Software (2002). Available at http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm. 1, 6.1.2, 6.1.4

[23] Erlang A.K. (1948) On the rational determination of the number of circuits. In *The life and works of A.K.Erlang.* Brockmeyer E., Halstrom H.L. and Jensen A., eds. Copenhagen: The Copenhagen Telephone Company. 4.2

[24] Feldman Z. and Mandelbaum A. (2004) Staffing of time-varying queues to achieve time-stable performance. Technical Report, Technion. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. V

[25] Fleming T.R. and Harrington D.P. (1991) *Counting Processes Survival Analysis*, Wiley. 8.4

[26] Friedman J. H. (1984) A variable span scatterplot smoother. Laboratory for Computational Statistics, Stanford University Technical Report No. 5. 4.1

[27] Gans N., Koole G. and Mandelbaum A. (2003) Telephone call centers: a tutorial and literature review. Invited review paper, *Manufacturing and Service Operations Management*, 5 (2), 79-141. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 1.1, 1.1, 6.2

[28] Garnett O. and Mandelbaum A. (2000) An Introduction to Skills-Based Routing and its Operational Complexities. Teaching note, Technion, Israel. Available at `http://iew3.technion.ac.il/serveng2004/Lectures/SBR.pdf`. 1

[29] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a telephone call-center with impatient customers. *Manufacturing and Service Operations Management* 4, 208-227. 1.1, 1, 4.1, 4.2, 5.3.1, 6.1, 6.1.2, 7.3, 12.4, 15.6, 15.6, 15.8, 18, V

[30] Gnedenko B.W. and Kovalenko I.N. (1968) *Introduction to Queueing Theory*, Jerusalem, Israel Program for Scientific Translations. 6.2

[31] Gupta P.L. and Gupta R.C. (1997) On the multivariate normal hazard. *Journal of Multivariate Analysis*, 62(1), 64-73. 10

[32] Gurvich I. (2004) Design and Control of the M/M/N Queue with Multi-Class Customers and Many Servers. M.Sc. Thesis, Technion, 2004. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 7.1, 7.2

[33] Haji R. and Newell G. (1971) A relationship between stationary queue and waiting time distributions. *Journal of Applied Probability*, 8, 617620. V

[34] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588. 4.2, 5.3.1, 7.2

[35] Haugen R.B. and Skogan E. (1980) Queueing systems with stochastic time out. *IEEE Trans. Commun.* COM-28, 1984-1989. 6.2

[36] Help Desk and Customer Support Practice Report; May 1997 survey results. The Help Desk Institute, SOFTBANK Forums, 1997. 1

[37] Iglehart D.L. (1965) Limit diffusion approximations for the many-server queue and the repairman problem. *Journal of Applied Probability*, 2, 429-441. 7.1

[38] Issaev E. (2003) Fitting Phase-Type Distributions to Data from a Telephone Call Center. M.Sc. Thesis, Technion. Available at `http://iew3.technion.ac.il/serveng/References/referencesbody.htmlsupport`. 5.3.1, 15.1.2

[39] Jagerman D.L. (1974) Some properties of the Erlang loss function. *Bell Systems Technical Journal*, 53, 525-551. 7.2, 10

[40] Jagers A.A. and Van Doorn E.A. (1986) On the continued Erlang loss function. *Operations Research Letters*, 5, 43-46. 10

[41] Jelenkovic P., Mandelbaum A. and Momcilovic P. (2004) Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications (QUESTA)*, 47, 53-69. 7.2

[42] Jennings O., Mandelbaum A., Massey W. and Whitt W. (1996) Server staffing to meet time-varying demand. *Management Science*, 42 (10), 1383-1394. V

[43] Jongbloed G. and Koole G.M. (2001) Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17, 307-318. V

[44] Jurkevic O.M. (1971) On many-server systems with stochastic bounds for the waiting time (in Russian), *Izv. Akad. Nauk SSSR Techniceskaja kibernetika*, 4, 39-46. 3, 6.2, 15.12

[45] Kingman J.F.C. (1961) On queues in heavy-traffic. *Journal of the Royal Statistical Society*, Series B, 24, 383-392. 7.1

[46] Kingman J.F.C. (1965) Heavy traffic approximation in the theory of queues. In Smith W. and Wilkinson W., eds. *Proc. of the Symposium on Congestion Theory*, 137-159. 7.1

[47] Kolmogorov A.N. and Fomin S.V. (1999) Elements of the Theory of Functions and Functional Analysis, Dover. 14.1

[48] Kort B.W. (1983) Models and methods for evaluating customer acceptance of telephone connections. *GLOBECOM '83*, IEEE, 706-714. 6.2

[49] Mandelbaum A., Sakov A. and Zeltyn S. (2000) Empirical Analysis of a Call Center. Technical report, Technion. 4.1, 4.1

[50] Mandelbaum A. and Shimkin N. (2000) A model for rational abandonment from invisible queues. *Queueing Systems: Theory and Applications (QUESTA)*, 36, 141-173. 6.2

[51] Mandelbaum A. and Schwartz R. (2002) Simulation experiments with M/G/100 queues in the Halfin-Whitt (QED) regime. Technical Report, Technion. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. V

[52] Mandelbaum A. and Zeltyn S. (2004) The Palm/Erlang-A Queue, with Applications to Call Centers. Teaching note to *Service Engineering* course. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 6.1

[53] Mandelbaum A. and Zeltyn S. (2004) Dimensioning M/M/n+G queue. Working paper. V

[54] Massey A.W. and Wallace B.R. (2004) "An Optimal Design of the M/M/C/K Queue for Call Centers", to appear in *Queueing Systems*. 7.2

[55] Palm C. (1953) Methods of judging the annoyance caused by congestion. *Tele*, 4, 189-208. 6.2

[56] 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.*) 6.1, 6.1.1, 6.1.1, 6.1.2, 6.1.3, 6.2

[57] Peterson T.S. (1950) Elements of Calculus, Harper, New York. 12.5

[58] Puhalskii A.A. and Reiman M.I. (2000) The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32, 564595. 7.2

[59] Rafaeli A. et al. (2004) Call Center Industry. Report on Management of Operations and Human Resources. Research Center for Work Safety and Human Engineering, Technion, Israel (In Hebrew). 1.1

[60] Riordan J. (1962) *Stochastic Service Systems*, Wiley. 6.1, 6.1.2

[61] Roberts J.W. (1979). Recent observations of subscriber behavior. *In Proceedings of the 9th International Tele-traffic Conference*. 6.2

[62] Shimkin N. and Mandelbaum A. (2002) Rational abandonment from tele-queues: non-linear waiting costs with heterogeneous preferences. Submitted to *QUESTA*. 6.2

[63] Sze D.Y. (1984) A queueing model for telephone operator staffing. *Operations Research*, 32, 229249. 7.2

[64] Shaked M. and Shanthikumar J.G. (1994) *Stochastic Orders and Their Applications*, Academic Press, New York. 11.1

[65] Trofimov V., Feigin P., Mandelbaum A. and Ishay E. (2004) DATA-MOCCA: Data Model for Call Center Analysis. Technical Report, Technion, Israel. Downloadable from `http://iew3.technion.ac.il/serveng/References`. 19.1

[66] U.S. Bureau of Labor Statistics. Table B-1:Employees on Nonfarm Payrolls by Major Industry, 1950 to Date. As reported on `www.bls.gov`. 1.1

[67] Ward A.R., Glynn P.W. (2002) A diffusion approximation for a Markovian queue with reneging. To appear in *Queueing Systems: Theory and Applications*. 7.4

[68] Whitt W. (2002) *Stochastic-Process Limits*, Springer-Verlag. 7.1

[69] Whitt W. (2002) Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions, working paper. 1.1

[70] Whitt W. (2003) Engineering Solution of a Basic Call-Center Model. *Management Science*, to appear. 7.4, V

[71] Whitt W. (2003) How Multiserver Queues Scale with Growing Congestion-Dependent Demand. *Operations Research*, 51, No. 4, 531-542. 7.2

[72] Whitt W. (2004) Fluid Models for Many-Server Queues with Abandonments. Submitted to *Operations Research*. 7.4, 17.1

[73] Whitt W. (2004) Two Fluid Approximations for Multi-Server Queues with Abandonments. Submitted to *Operations Research Letters*. 7.4

[74] Whitt W. (2004) Sensitivity of Performance in the Erlang A Model to Changes in the Model Parameters. Submitted to *Operations Research*. 6.1, 6.1.2, 7.4

[75] Whitt W. (2004) Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. Submitted to *Management Science*. 7.4, V

[76] Zohar E., Mandelbaum A. and Shimkin N. (2002) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science*, 48, 566-583. 6.1.3, 6.2