The Impact of Customers' Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/n+G Queue

Avi Mandelbaum and Sergey Zeltyn

Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

emails: avim@tx.technion.ac.il, zeltyn@ie.technion.ac.il

March 30, 2004

Abstract

Our research is motivated by a phenomenon that has been observed in telephone call center data: a clear linear relation between the probability to abandon and average waiting time. Such a relation is theoretically justifiable when customers' patience is memoryless, 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.

In the theoretical part of the paper, 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 the experimental part, 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.

Note to the Reader: a color-version of the present paper is downloadable from `http://iew3.technion.ac.il/serveng/References/references.html`.

Keywords: Abandonment, Call-Centers, Erlang-A, Queues and Queueing.

Acknowledgements. The authors would like to thank Arkadi Nemirovski for valuable comments on the proof of Theorem 1 and Eva Isaev for providing the data to Figure 2.

This work was supported by the ISF (Israeli Science Foundation) Grants 388/99 and 126/02, the Wharton Financial Institutions Center, the Technion funds for the promotion of research and sponsored research, the William Davidson Applied Research Fund and the Niderzaksen Fund.

Contents

1 Introduction

The prevalent model for performance analysis of call centers is the M/M/n queue, frequently referred to as *Erlang-C*. It assumes Poisson arrivals, exponentially distributed service times and n statistically identical agents.

The Erlang-C model is deficient as an accurate depiction of a call center in some major respects: it does not include priorities of customers, it assumes that skills of agents and their service-time distributions are identical, it ignores customers' re-calls, etc. However, in our opinion and based on our experience, the major drawback of the Erlang-C model is that it does not accommodate customers' impatience while waiting, which could culminate in their abandonment (hang-ups).

Garnett et al. [12] analyzed the simplest abandonment model M/M/n+M (*Erlang-A*), 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. [6] 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.

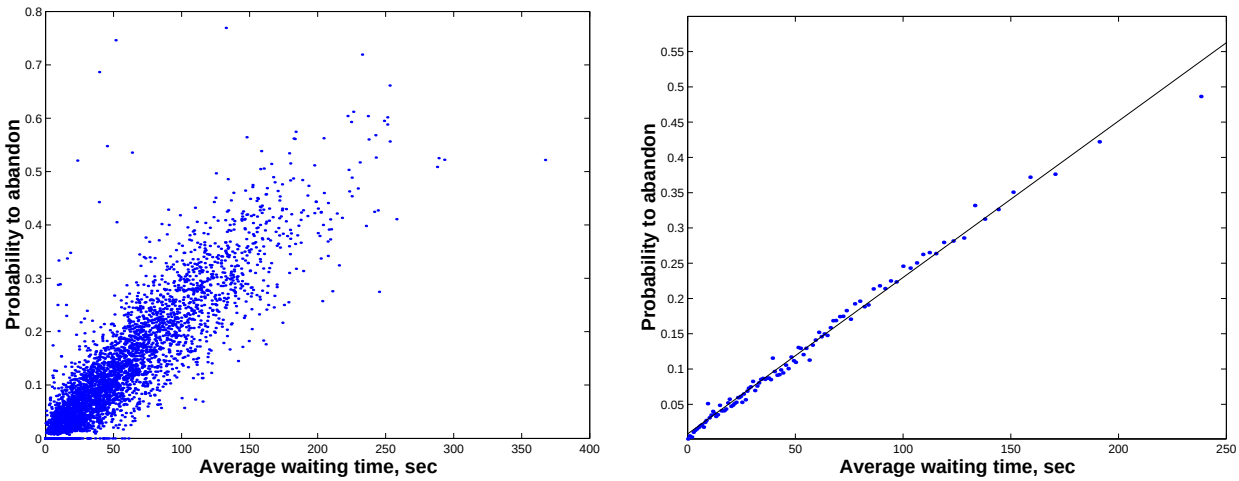
An important question thus arises: how robust is the M/M/n+M model with respect to deviations in its characteristics? In the present paper we answer it for customers (im)patience, which is natural to pursue within the M/M/n+G framework. This model assumes that customers arrive to the queueing system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

In our opinion, the probability to abandon is perhaps the most important operational measure for call center performance. It is one of only few customer-centric measures, through which customers indirectly "inform" the call center on whether the offered service is "worth its wait". Commonly-used measures, such as waiting times, are of course also interesting to customers as they relate to their delay experience, but they are "objective" at the system level. 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 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 1 displays an empirical relationship between the two measures. It was plotted using the yearly data of the Israeli bank call center, analyzed in Brown et al. [6] and Mandelbaum et al. [19]. First, the probability to abandon and average wait were computed for the 4158 hour intervals that constitute the year 1999. The left plot of Figure 1 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

Figure 1: **Probability to abandon vs. average waiting time**



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 1. (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 1 are nearly identical.

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

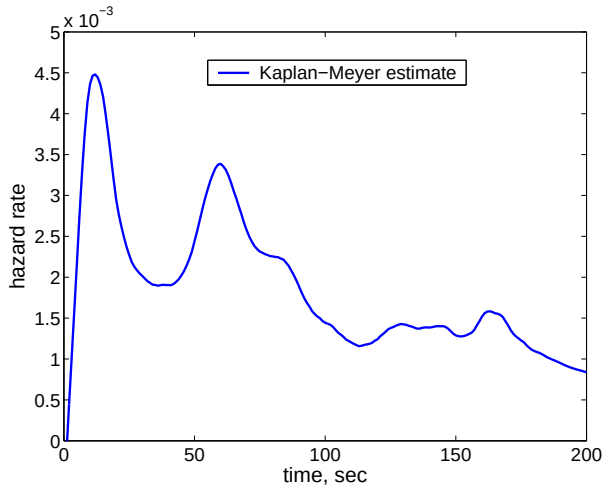
$$\% \text{ Abandonment} = \frac{\mathrm{E}[W]}{\mathrm{E}[R]} \tag{1.1}$$

is provable for models with *exponential* patience (as in [1] or [30], for example). But, as will now be recalled from [6], this obviously is *not* our case: the hazard rate of patience is far from being constant, as it should have been if R was exponential.

Figure 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 [8] or Appendix of [30]) and a smoothing algorithm [10] were performed in order to produce that curve. The hazard pattern is clearly nonlinear, hence the patience distribution pattern is far from exponential. Note the two peaks of the hazard rate: the first peak approximately at 15 seconds and the second peak approximately at 60 seconds. (It turns out that these two surges of abandonment take place after two recorded messages to which customers are exposed: the first one when they enter the queue and the second one after approximately 1 minute.)

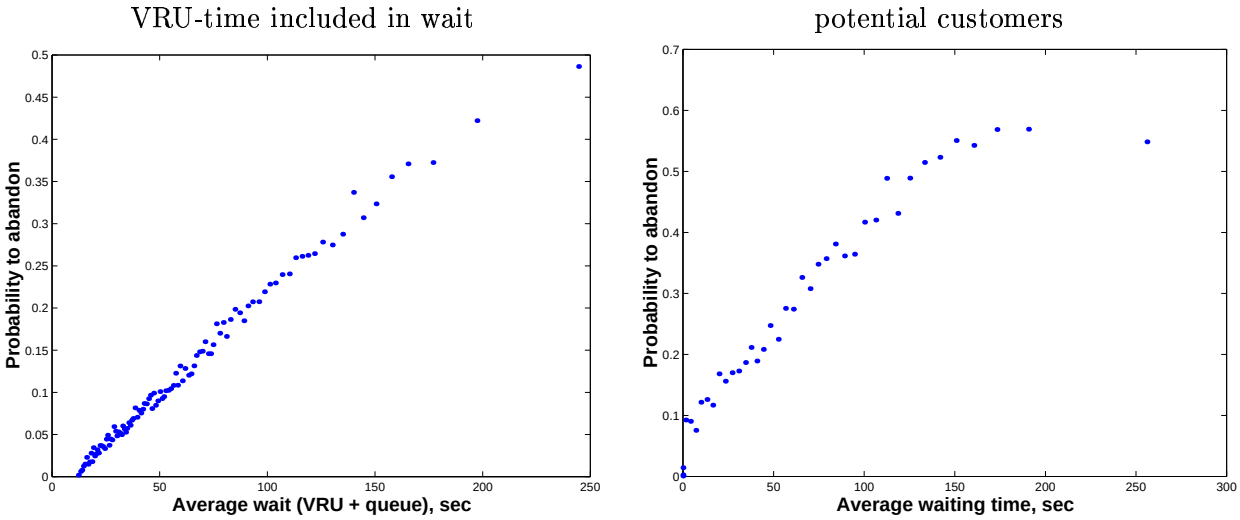
Two other examples of a linear relation between the probability to abandon and average wait are presented in Figure 3. Both are based on the same yearly call center data, referred to above. The first plot takes into account all customers. It differs from Figure 1 by its definition of waiting

Figure 2: **Hazard rate for patience of regular customers**



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 1.

Figure 3: **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 3 can be partially clarified using [19]. 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 2). 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.

2 Summary of results and structure of the paper

The issues that arose in the Introduction 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 4.1). A non-obvious stochastic order relation for patience-time distributions G_1 and G_2 is presented (Lemma 1) that implies order relation 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 1) that, for a fixed average patience $\bar{R}$, the deterministic patience (all customers are willing to wait *exactly* $\bar{R}$) maximizes some performance measures of M/M/n+G (average wait, average queue, probability of wait) while it minimizes the probability to abandon.

We then present some *light-traffic* results (Subsection 4.2). 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 (4.7) and average wait (4.8), both conditional on positive wait.

Subsection 4.3 contains a short presentation of some QED results from Mandelbaum and Zeltyn [22]: in the QED (Quality- and Efficiency-Driven) regime, agents are heavily utilized, which corresponds to the opposite extreme of the light traffic in Subsection 4.2. This QED framework adds useful insights into some problems discussed in our paper. In particular, (4.10) gives rise to additional reasons for linearity of the P{Ab} / E[W] curve.

Then we proceed with some theory-driven experiments (Section 5). 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 4.2 and heavy-traffic operating regime in Subsection 4.3. In particular, we check for patience patterns that imply the relations in Figures 1 and 3. We verify that some non-exponential distributions (uniform, hyperexponential) give rise to a linear P{Ab} / E[W] relation (Figures 5-8) 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 Figure 2 and observe a relatively linear pattern, thus supporting Figure 1. In addition, simulation is used in order to check influence of lognormal service times on system performance (versus exponential services). In our experiments, we observe negligible differences for probability to abandon and very small for the average wait (Figure 9).

On the other hand, some distributions (e.g. deterministic) imply strictly non-linear patterns for the above relation (Figure 10). We also connect the *abandonment rates*, introduced in Brandt and Brandt [5] with linearity or non-linearity of the P{Ab} / E[W] relation (Figure 11). Finally, some theoretical results from Section 4 and Brandt and Brandt [5] are validated via numerical experiments.

Summarizing, the structure of the paper is as follows. Section 3 reviews some existing M/M/n+G results. Section 4 provides formulations of theoretical results with comments. In Section 5 we describe the theory-driven experiments mentioned above and Section 6 has our conclusions. Finally, proofs of our theoretical results are presented in Section 7.

3 Existing theory 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. Let $\bar{G} = \{\bar{G}(x),\ x \geq 0\}$ denote the survival function of the patience time R: $\bar{G}(x) = \mathrm{P}\{R > x\},\ x \geq 0$. We assume that an arriving customer encounters, in steady-state, an offered waiting time V (the time a customer would have to wait given that patience is infinite). Then the actual queueing time of a steady-state customer equals $W = \min(V, R)$.

The main performance measures analyzed here include the probability to abandon, waiting time and queue length. We shall also consider these characteristics conditioned on positive wait.

The seminal work on queueing systems with impatient customers is Palm [23, 24]. 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.

The Erlang-A (M/M/n+M) queueing system with exponential patience times can be analyzed as a birth-and-death Markov process. It was first solved analytically in Palm [24] and Riordan [26].

Gnedenko and Kovalenko [13] analyzed the M/M/n+D queueing system (deterministic patience times). Jurkevic [16] applied their methods to the general M/M/n+G system. Independently, the M/M/n+G queue was analyzed by Baccelli and Hebuterne [1] and Haugen and Skogan [14]. Boxma and de Waal [3] developed several approximations for the probability to abandon in the M/M/n+G queue and checked them via simulation.

The derivation of M/M/n+G performance measures continued in Brandt and Brandt [4, 5].

They considered the more general M(k)/M(k)/n+G system where arrival and service rates can 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 [4, 5] (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 [23] 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 [18] also used the Weibull distribution to model patience while waiting for a dial tone. Baccelli and Hebuterne [1], using data from Roberts [27], fitted it to an Erlang distribution with 3 phases. Brown et al. [6], in research on a bank call center, encountered the patience times in Figure 2. Finally, Daley and Servi [9] 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.

Concerning models of customers' impatience, readers are referred to the papers of Mandelbaum, Shimkin and Zohar [20, 30, 28] 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 [11].

In the present paper, we use the M/M/n+G results from Baccelli and Hebuterne [1] and Brandt and Brandt [4, 5]. Subsections 7.1 and 7.3 survey some material from these references which is relevant in our proofs. In addition, we shall make use of the abandonment rates α_l given l customers in queue, introduced in [5]. (See (7.37) in Subsection 7.2 for the formal definition.) In terms of α_l, the probability to abandon can be represented by

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{\sum_{l=n+1}^{\infty} \alpha_{l-n}\pi_l}{\lambda}\,, \tag{3.2}$$

where π_l are the steady-state number-in-system probabilities.

It is proved in [5] that abandonment rates are asymptotically linear in the sense that

$$\lim_{l\to\infty} \frac{\alpha_l}{l} \;=\; \frac{1}{\mathrm{E}[R]}\,. \tag{3.3}$$

If the m-th moments of the patience-time distribution are finite for all $m > 2$, then

$$\lim_{l\to\infty} (\alpha_l - \alpha_{l-1}) \;=\; \frac{1}{\mathrm{E}[R]}\,. \tag{3.4}$$

Note that for exponential patience, the last two equalities are, in fact, exact for all $l = 1, 2, \ldots$

4 Some new results for the M/M/n+G queue

4.1 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 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

$$\int_0^x \bar{G}_1(\eta)d\eta \;\geq\; \int_0^x \bar{G}_2(\eta)d\eta \tag{4.5}$$

prevails for all $x > 0$, where $\bar{G}_1$ and $\bar{G}_2$ are the corresponding survival functions. Let $\mathrm{P}^i\{\mathrm{Ab}\}$, $\mathrm{P}^i\{\mathrm{Ab}|W > 0\}$ 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. $\mathrm{P}^1\{W > 0\} \;\geq\; \mathrm{P}^2\{W > 0\}$.
b. $\mathrm{P}^1\{\mathrm{Ab}\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}\}$; $\mathrm{P}^1\{\mathrm{Ab}|W > 0\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}|W > 0\}$.

Theorem 1.
Consider the M/M/n+G queue with fixed parameters λ, μ, n and a *fixed* average patience time $\bar{R}$. Then the deterministic distribution of patience G_d (every customer is willing to wait exactly $\bar{R}$) has the following "extremal" properties among all patience-time distributions with average $\bar{R}$:

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

Remarks.

1. It will be shown that Statements **a** and **b** of Theorem 1 are corollaries of Statements **a** and **b** from Lemma 1, respectively. However, inequality (4.5) does not imply order relations for average wait or average queue. An example is provided in Section 5 (see comments adjacent to Figure

12).

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 (4.5) prevails automatically. Bhattacharya and Ephremides [2] 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).

The proofs of Lemma 1 and Theorem 1 are given in Subsection 7.2.

4.2 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 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

$$\lim_{\lambda\to 0} \frac{\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[W]} \;=\; \alpha_1 \;\triangleq\; \frac{1}{\int_0^\infty \bar{G}(x)e^{-n\mu x}dx} - n\mu\,. \tag{4.6}$$

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

$$\lim_{\lambda\to 0} \mathrm{P}_\lambda\{\mathrm{Ab}|W>0\} \;=\; 1 - n\mu\int_0^\infty \bar{G}(x)e^{-n\mu x}dx \;=\; \mathrm{P}\{R < \exp(n\mu)\}\,, \tag{4.7}$$

$$\lim_{\lambda\to 0} \mathrm{E}_\lambda[W|W>0] \;=\; \int_0^\infty \bar{G}(x)e^{-n\mu x}dx \;=\; \mathrm{E}[R \wedge \exp(n\mu)]\,, \tag{4.8}$$

where the patience R is independent of the $\exp(n\mu)$ random variable. See Subsections 7.3 and 7.4 for proofs.

Remark. Formulae (4.7) and (4.8) 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.

4.3 Heavy traffic with many agents

Here we outline some key results from Mandelbaum and Zeltyn [22], where an approach to the M/M/n+G queue analysis is different from this paper.

Assume that μ and G are fixed, and let the arrival rate $\lambda \to \infty$. For reasonable performance, the staffing level clearly must increase with λ. As it turns out, this necessary increase is best understood when measured relative to the mean *offered load* $\frac{\lambda}{\mu}$. More precisely, the following

three operating (staffing) regimes arise. (The notation $A \sim B$ indicates that the ratio $\lim A/B = 1$, as $\lambda \to \infty$.)

Quality-Driven regime: $n \sim \frac{\lambda}{\mu} \cdot (1+\gamma)\,, \quad \gamma > 0, \quad \lambda \to \infty\ .$
In this case, most performance characteristics (probability to abandon, probability of wait, average wait) decrease exponentially in n. Naturally, one must pay for the excellent service by high (most likely excessive) staffing.

Efficiency-Driven regime: $n \sim \frac{\lambda}{\mu} \cdot (1-\gamma)\,, \quad \gamma > 0, \quad \lambda \to \infty.$
Here virtually all customers have to wait, the probability to abandon converges to β and average wait converges to a constant that depends on the patience-time distribution.

QED (Quality- and Efficiency-Driven) regime:

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})\,, \quad -\infty < \beta < \infty, \quad \lambda \to \infty\,. \tag{4.9}$$

This operating regime combines high efficiency and service quality (given the number of agents is large enough). A specific service level is determined by the *service grade* β.

We now quote our main theorem on the QED regime that was proved in [22]. To this end, introduce the *hazard rate* function of the standard normal distribution

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

where $\Phi(x)$ is its cumulative distribution function and $\phi(x) = \Phi'(x)$ is the density.

Theorem 2. Assume that the density of patience exists at the origin and its value g_0 is strictly positive. Then, in the QED operating regime, namely $\lambda \to \infty$ and n as in (4.9), we have
a. The probability of wait converges to a constant that depends on β and $\frac{g_0}{\mu}$:

$$\mathrm{P\{W > 0\}} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}\,,$$

where

$$\hat{\beta} = \beta\sqrt{\frac{\mu}{g_0}}\,.$$

b. The probability of delayed customers to abandon decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P\{Ab}|W > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}}\left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right)\,.$$

c. The average waiting time of delayed customers decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[W|W > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}}\left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right)\,.$$

The above implies that

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}|W>0\}}{\mathrm{E}[W|W>0]} \;\sim\; g_0\,. \tag{4.10}$$

■

Remark. We shall observe in Subsection 5.1 that, practically, g_0 can be very close to the limit in the light-traffic formula (4.6). This can be explained by an asymptotic argument, which can be made rigorous via the methods of de Bruijn [7]. The outline of this argument is as follows:

$$\int_0^\infty \bar{G}(x)e^{-n\mu x}dx \;\approx\; \int_0^\infty (1-g_0x)e^{-n\mu x}dx \;=\; \frac{1}{n\mu}-\frac{g_0}{n^2\mu^2} \qquad (n\to\infty)\,. \tag{4.11}$$

Now substituting (4.11) into the asymptotic expressions (4.7) and (4.8) from Lemma 2 yields

$$\frac{\mathrm{P}_\lambda\{Ab\}}{\mathrm{E}_\lambda[W]} \;=\; \frac{\mathrm{P}_\lambda\{\mathrm{Ab}|W>0\}}{\mathrm{E}_\lambda[W|W>0]} \;\approx\; g_0 \qquad (\lambda\to 0, n\to\infty)\,.$$

Remark. Approximations, analogous to Theorem 2, were also developed in [22] for some types of distributions with $g_0 = 0$ and for models with balking. In general, one gets convergence rates of performance measures that are different from Theorem 2. Relation (4.10) can be uninformative in those cases (the limit is zero or infinity).

Figure 4 illustrates both Theorem 2 and approximations for $g_0 = 0$. The M/M/n+G queue with $n = 10$ and service rate $\mu = 1$ is considered. Four patience distributions were chosen, each with mean 2:

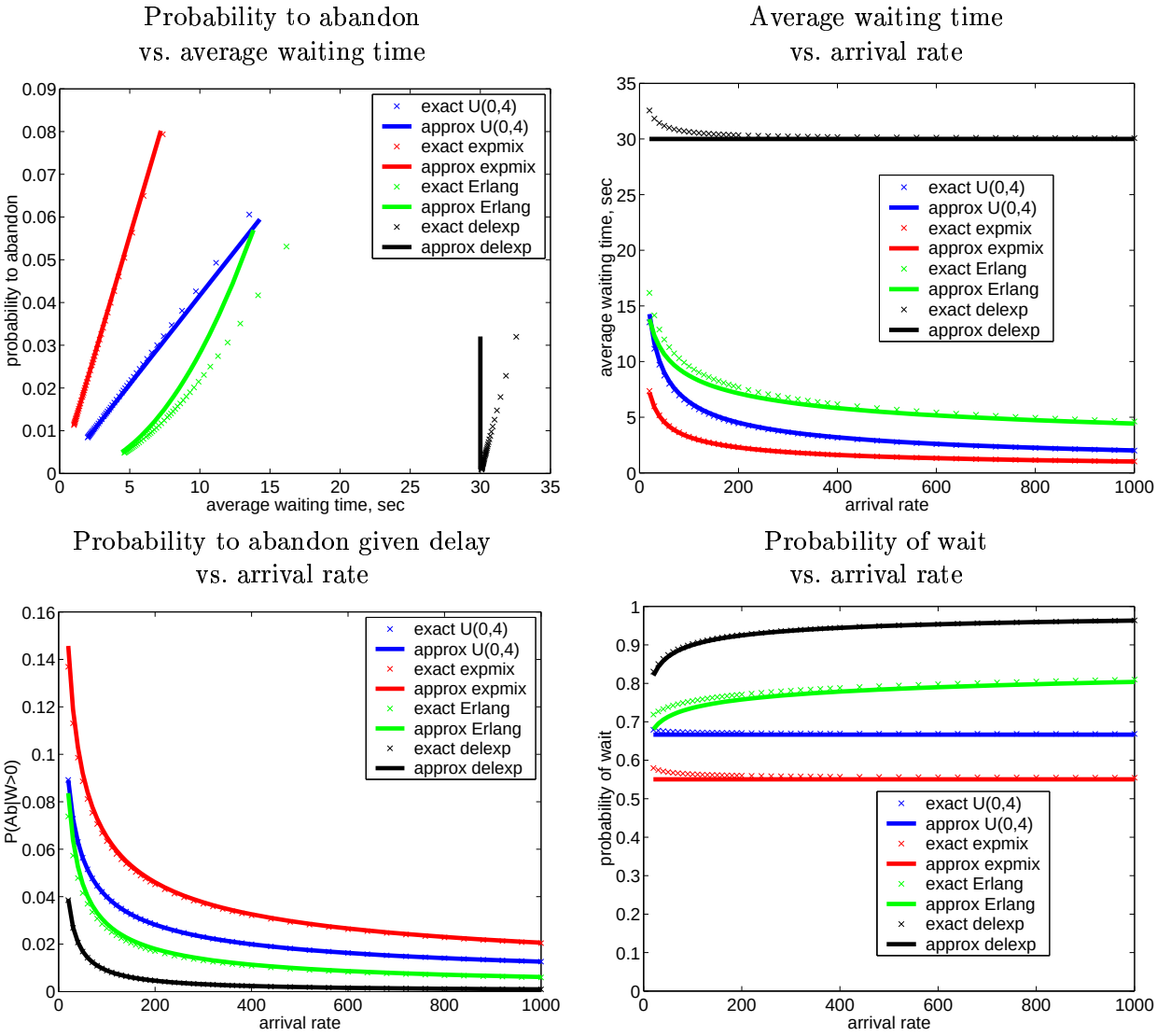
- *Uniform* distribution on (0,4); $g_0 = 0.25$.
- *Hyperexponential* distribution: mixture of exp(mean=1) and exp(mean=3), each one with probability 0.5; $g_0 = 2/3$.
- *Erlang* (Gamma) distribution: two exponential phases, each with mean 1; $g_0 = 0$.
- *Delayed exponential* distribution equal to 1+exp(mean=1); $g_0 = 0$.

The arrival rate λ varies from 20 to 1000 and the number of agents n is chosen according to the QED staffing rule with $\beta = 0$ ($n = \frac{\lambda}{\mu}$). Then exact and approximate values of various steady-steady performance characteristics are compared. (Solid lines are for approximations, and x's are for exact values that were calculated via Brandt and Brandt [4] formulae.)

The first picture is the scatterplot of the probability to abandon versus average wait. Note the straight-line patterns (recall Figure 1) for the two distributions with $g_0 > 0$.

The other three plots show three different performance characteristics, as they change with arrival rate. We observe that the quality of our approximations varies from good to excellent.

Figure 4: **QED operating regime. Service grade $\beta = 0$. Performance measures and approximations**



5 Descriptive results

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 $\bar{R} = 2$. 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 [4], was used for calculations. The full description of our experiments will be published in Mandelbaum and Zeltyn [22]. 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.

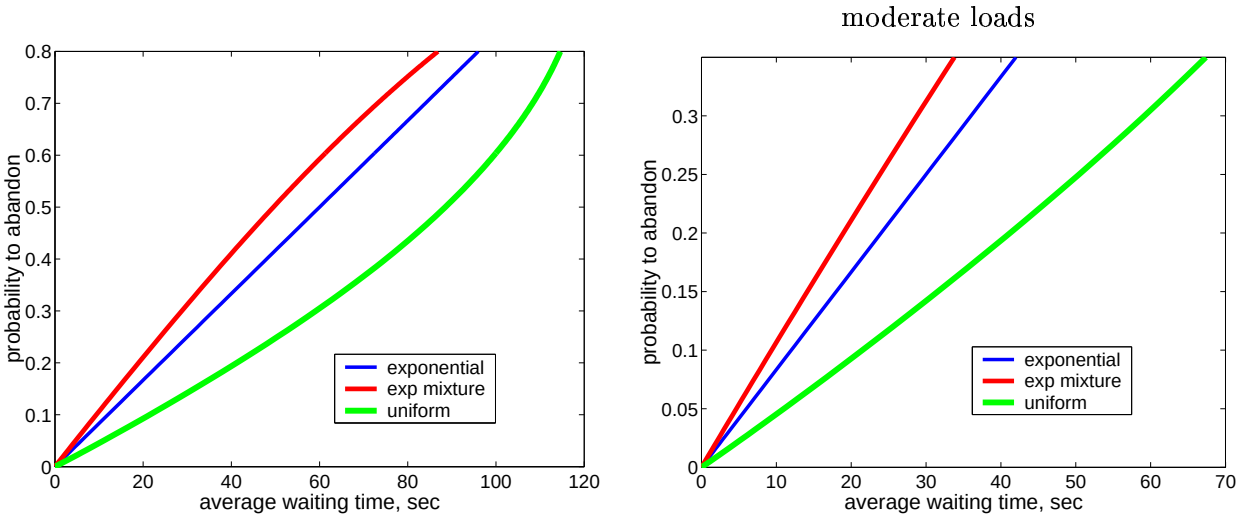
- Explanations of linearity or non-linearity of the above relation.
- Checking some theoretical results for M/M/n+G, exact and asymptotic.
- Exploring the relation between performance measures and the arrival rate, for various patience distributions.

5.1 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 5 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 6 presents the same performance measures but conditioned on positive wait.

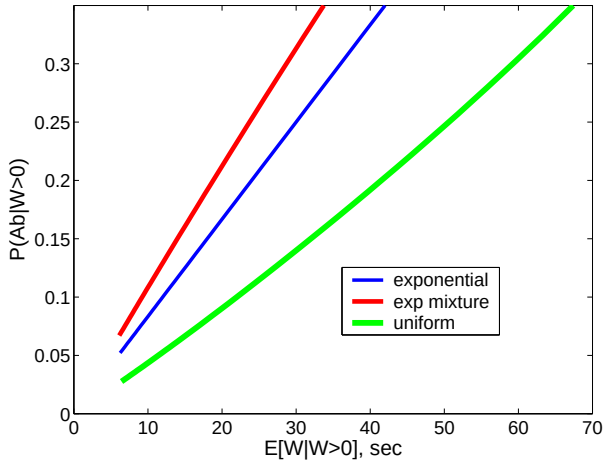
The first plot of Figure 5 illustrates the general form of P{Ab} / E[W] curves, given λ that varies from zero to infinity. Those curves always connect the origin and the point $(\bar{R}, 1)$ ($\bar{R} = 2$ minutes, or 120 seconds, in our case). The reason is that the average wait converges to the average patience, as $\lambda \to \infty$.

Figure 5: **Probability to abandon vs. average wait**



From relation (1.1) we know that exponential patience implies a linear curve, which is supported by Figures 5-6. The curves for other distributions need not be linear. For example, we observe (the first plot of Figure 5) the convex curve for the uniform distribution and the concave curve for the exponential mixture.

Figure 6: **Probability to abandon vs. average wait: delayed customers**



However, if we consider "reasonable loads" (the second plot of Figure 5) the two non-exponential curves are strikingly close to linear patterns. The same phenomenon is observed for conditional values (Figure 6), as well as for uniform distributions with different average [22].

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 (4.7), 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 6.

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.

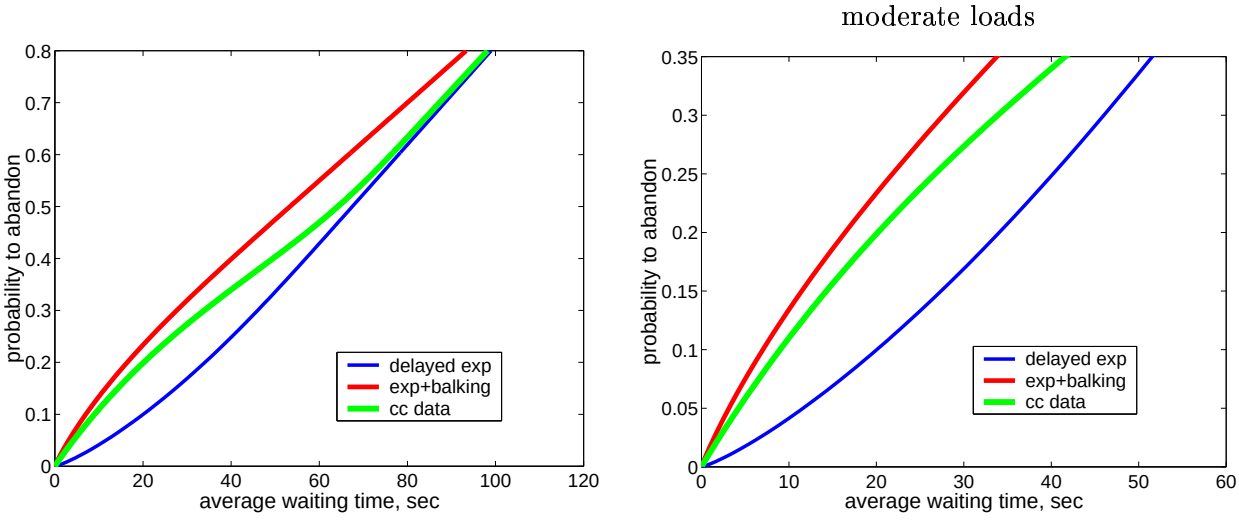
2. Exponential smoothing was performed for the tail of the distribution (for time values larger than 1.2, which is equivalent to 6 minutes in the initial scale). The reason is that estimates of survival functions are very unreliable for large time values (the data is heavily censored, see [6]). Note that the linear pattern of the "cc data" curve for very large loads (in the first plot of Figure 7) is a consequence of the exponential smoothing.

Figures 7-8 were plotted using the same algorithm as in Figures 5-6.

We observe that the cc data curve for moderate loads (the second plot of Figure 7) 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 Figure 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 1 is in fact somewhat concave near zero.

The graphs of the two theoretical distributions in Figure 8 can be used, to some extent, for validation of Figure 3. First, Figure 8 demonstrates the theoretically predicted linear curve with y-positive intercept for the exponential distribution with balking. (Recall the second plot in Figure 3.) The conditional curve of the delayed exponential distribution in Figure 8 is close to a straight line as well, which provides analogy with the first plot of Figure 3. (Note that in both cases, all customers are delayed, either in VRU, or in queue.)

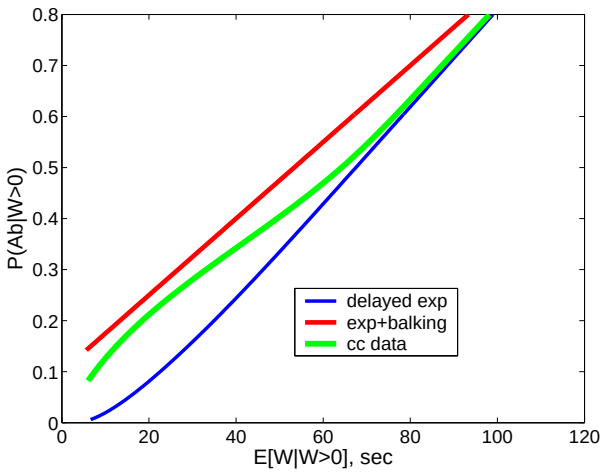
Figure 7: **Probability to abandon vs. average wait**



It was observed in [6] 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 7 and 8, 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

Figure 8: **Probability to abandon vs. average wait: delayed customers**



same as in our call center data). The arrival rate was varied from 3 to 15. Figure 9 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 [3] 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 1.

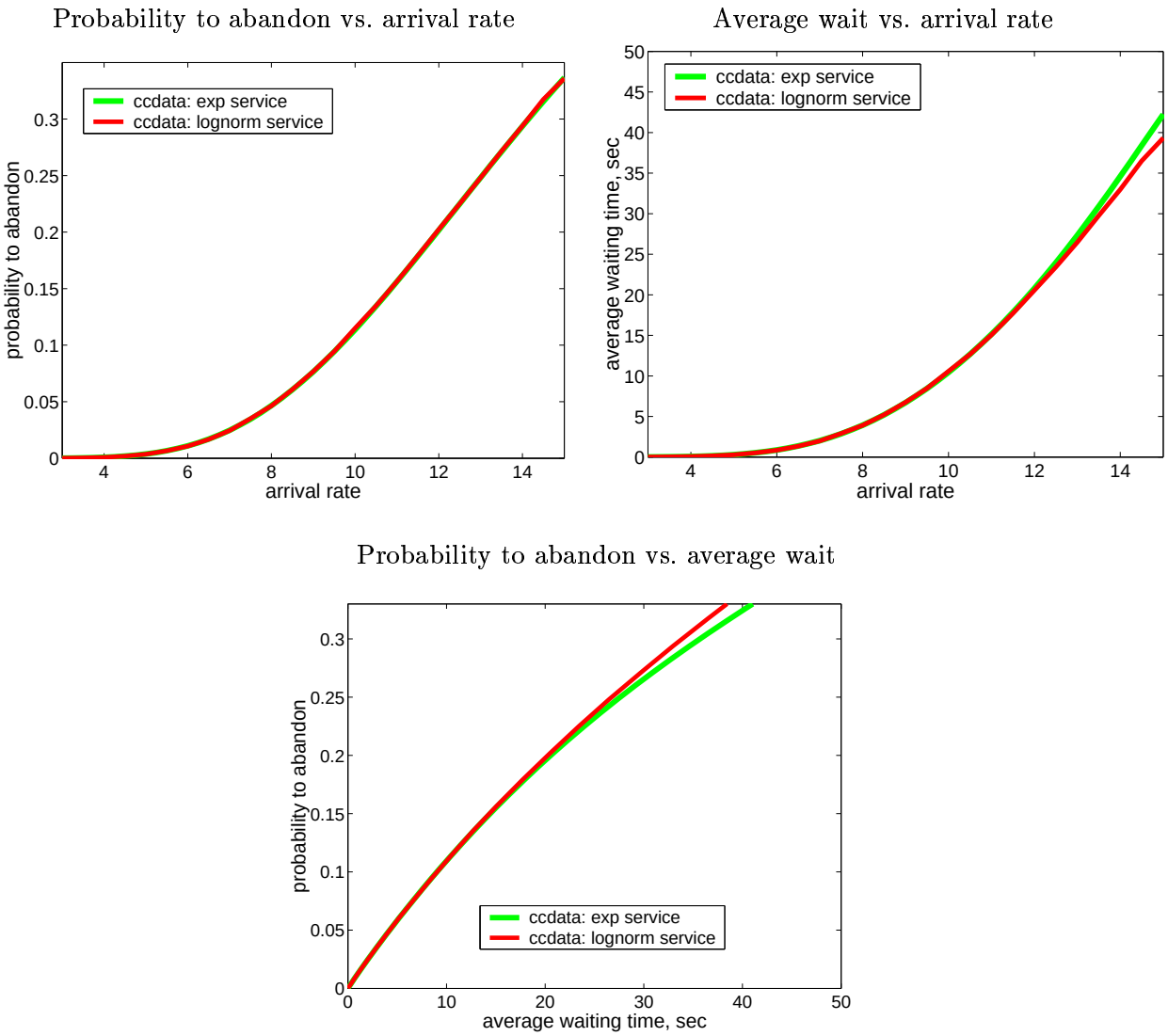
Summarizing, the patterns of Figure 1 and our "cc data" curves are similar. The observed difference can be attributed to various discrepancies between the call center environment and the M/M/n+G model: the number of servers is not constant over the day, the system does not always enter steady-state, priorities and skill-based routing, etc. Yet we believe that simple queueing models, such as Erlang-A (see Chapter 8 of Brown et al. [6]) or M/M/n+G, could turn out very useful in the analysis of complex call centers.

5.2 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.
- The fourth distribution is a 50-50% mixture of two constants: 0.2 and 3.8.

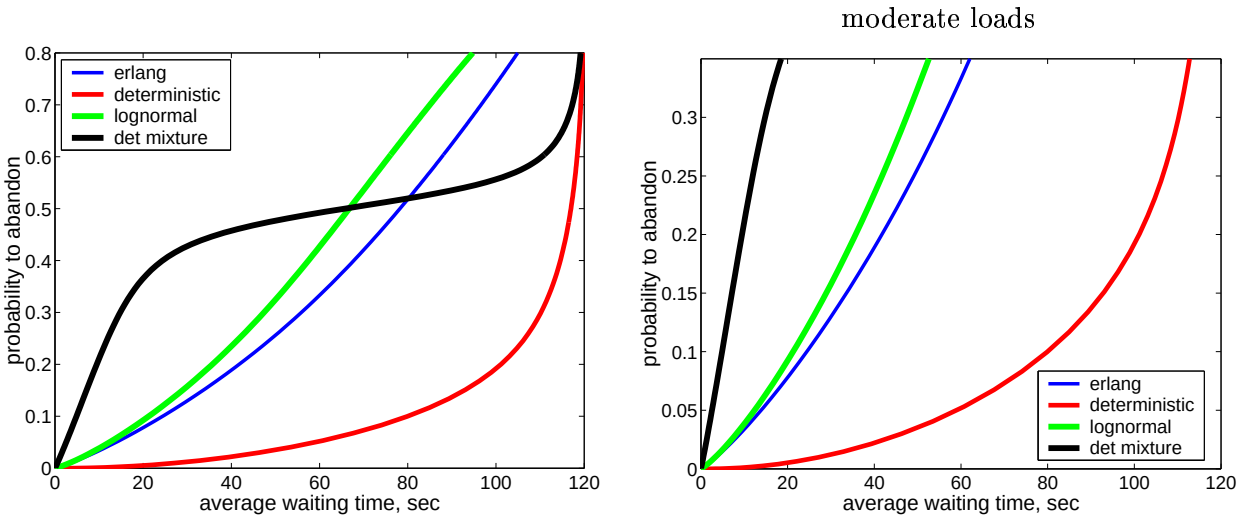
Figure 9: **Comparing M/M/n+G and M/G/n+G (lognormal service)**



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 10, the deterministic curve is strictly convex and lies below all other plots, as could be expected from Theorem 1. The patterns 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.

Figure 10: **Probability to abandon vs. average wait**



5.3 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 (3.2),

$$\frac{\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[W]} \;=\; \frac{\lambda\cdot\mathrm{P}_\lambda\{\mathrm{Ab}\}}{\mathrm{E}_\lambda[L_q]} \;=\; \frac{\sum_{l=1}^{\infty}\alpha_l\cdot\pi_{n+l}(\lambda)}{\sum_{l=1}^{\infty} l\cdot\pi_{n+l}(\lambda)}\,, \tag{5.12}$$

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

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

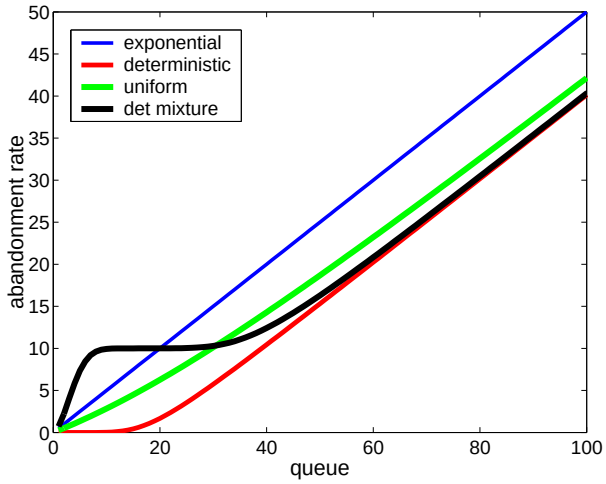
Figure 11 supports the claim that the expression in (5.12) is approximately linear with respect to λ, if the abandonment rates α_l are approximately linear with respect to the queue length l.

We plotted four curves of abandonment rates in Figure 11, using the patience distributions from Examples 1 and 3. One observes a clear connection between the curves from these examples (Figures 5 and 10) and Figure 11: 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.)

The abandonment plot for the deterministic mixture also conforms to the corresponding curve at Figure 10: linear increase for small queues, then a "plateau" (all short-patience customers abandon) and, finally, linear increase again.

Figure 11 provides the opportunity to verify our equations (3.3) and (3.4), both taken from Brandt and Brandt [5]. We observe that the four curves share the same slope $1/\bar{R} = 0.5$, for large values of l.

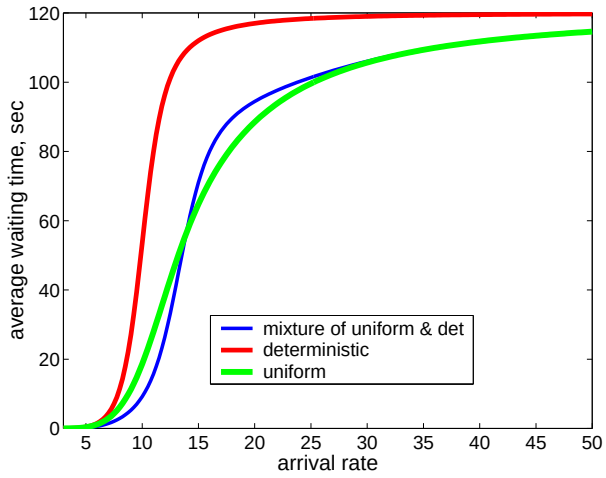
Figure 11: **Abandonment rate given queue**



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

Figure 12 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.

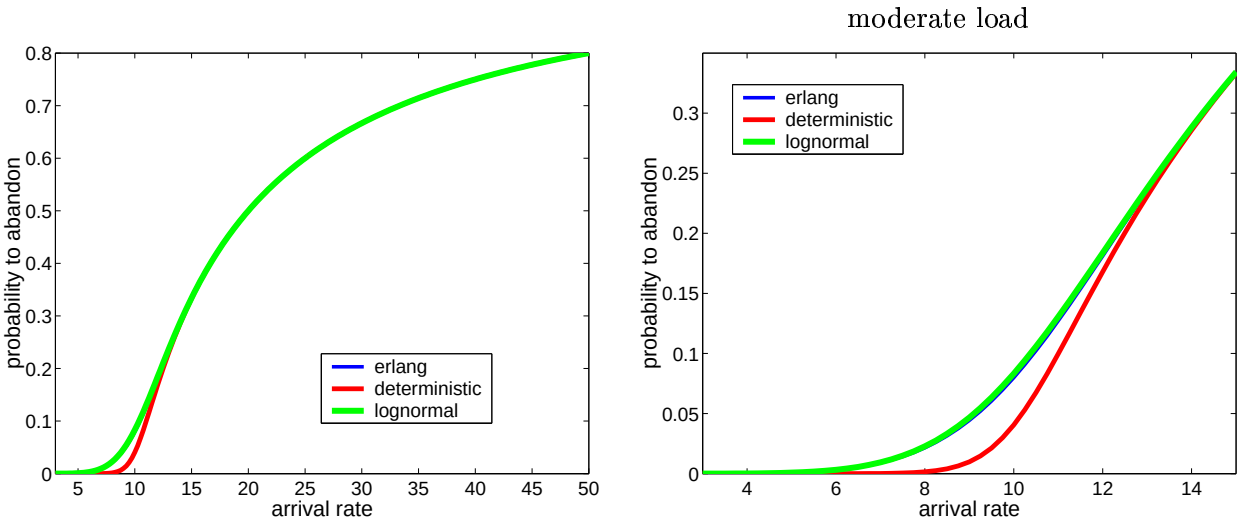
Figure 12: **Average wait vs. arrival rate**



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 12 illustrates two facts. First, as predicted by Theorem 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 (4.5) from Lemma 1. Therefore, this relation does *not* imply any order for average waits.

Figure 13: **Probability to abandon vs. arrival rate**



Finally, we present in Figure 13 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 [12]), 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 13 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 [3].)

5.5 Quantitative verification of linearity: ratio and curvature

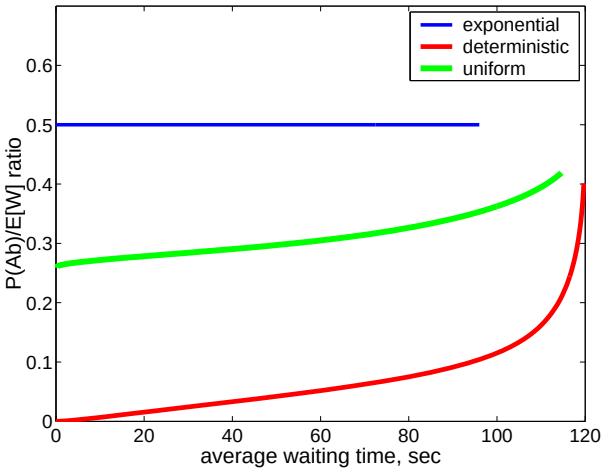
So far we have relied on a visual inspection to identify linearity (or non-linearity) of the relation between the probability to abandon and average wait. Two simple methods, illustrated by Figure 14, can be used to verify this quantitatively.

First, the ratio between the two performance measures can be plotted. The first plot in Figure 14 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.

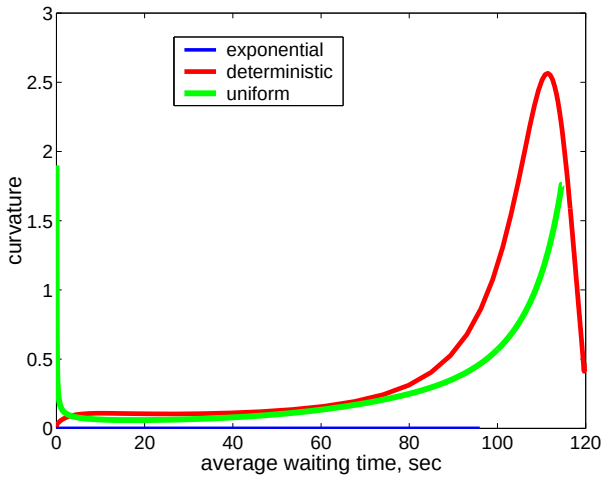
The second plot presents the curvature of the "$\mathrm{P}\{\mathrm{Ab}\}$ vs. $\mathrm{E}[W]$" graphs. (See [25], page 119, for the definition of curvature. Recall that zero curvature corresponds to a straight line.) Consult Figure 12 in order to verify that the curvature of the deterministic distribution increases steeply even for moderate loads and the curvature of the uniform distribution is high for only the very large loads.

Figure 14: **Two methods to evaluate linearity**

Ratio between probability to abandon and average wait



Curvature of P{Ab} vs. E[W]



6 Conclusions

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 Section 1 (e.g. Figure 1), 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 beyond Erlang-A model with exponential patience. To be more specific, an exact linear relation holds theoretically for exponential patience (Figure 5). 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 5). There are exceptions, however, as apparent for the deterministic case (Figure 10).

Finally, we outline some directions worthy of further research. First, empirically, additional studies of customers' patience in tele-service should be performed. Currently data collection in two large banks in the U.S. and Israel is in process. Then, theoretically, it is important to study the M/G/n+G model with non-exponential *service times*. (Recall that in the call center, analyzed in [6] and used here, service times were lognormal.) However, exact analysis of the M/G/n+G queue seems prohibitively difficult, hence one would probably start with approximations (see [29], for example) and simulation (see [21]).

An important motivation for this kind of research is the necessity to evaluate robustness of simple models (Erlang-A, M/M/n+G) in settings with generally distributed service and patience times. For example, it is important to validate the conclusions of Boxma and de Waal [3] which observed insensitivity of the probability to abandon with respect to the service-time distribution and verify if this statement can be extended to other performance measures (average wait,

probability of wait, etc.).

Additional research on extremal properties of patience distributions could be pursued as well. For example, we conjecture that, given that average patience is fixed, the deterministic distribution minimizes the abandonment rates α_l (recall Subsection 5.3) for all $l > 0$.

7 Proof of theoretical results

In Subsection 7.1 we survey some material from Baccelli and Hebuterne [1], which constitutes a necessary theoretical background for Lemma 1 and Theorem 1. Then the proofs of Lemma 1 and Theorem 1 are presented (Subsection 7.2). We proceed in Subsection 7.3 with a review of Brandt and Brandt [4, 5], which is used in the proof of Lemma 2 (Subsection 7.4).

7.1 Summary of relevant results from Baccelli and Hebuterne [1]

Baccelli & Hebuterne [1] assume that arriving customers can calculate their offered wait already at an arrival epoch. (The offered wait is defined as the time that a customer, equipped with infinite patience, would wait until being served.) If the offered wait exceeds their patience time, they 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 directly related to queue-length are concerned.

The analysis in [1] 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:

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

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

$$\pi_j = \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\,\pi_0\,, \quad 0 \leq j \leq n-1 \tag{7.13}$$

$$v(x) = \lambda\,\pi_{n-1} \exp\left\{\lambda \int_0^x \bar{G}(u)du - n\mu x\right\}, \tag{7.14}$$

where

$$\pi_0 = \left[1 + \frac{\lambda}{\mu} + \cdots + \left(\frac{\lambda}{\mu}\right)^{n-1} \frac{1}{(n-1)!}(1+\lambda J)\right]^{-1}, \tag{7.15}$$

$$J \triangleq \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - n\mu x\right\} dx\,. \tag{7.16}$$

Furthermore, the probability to abandon is calculated to be

$$\mathrm{P}\{\mathrm{Ab}\} = \left(1 - \frac{n\mu}{\lambda}\right)\left(1 - \sum_{j=0}^{n-1} \pi_j\right) + \pi_{n-1}\,. \tag{7.17}$$

7.2 Proofs of Lemma 1 and Theorem 1

We start with the proof of Lemma 1, then derive its corollaries: **a** and **b** of Theorem 1 and, finally, proceed to **c** and **d** that require more complicated proofs.

Proof of Lemma 1.

a. Formula (7.16) and inequality (4.5) imply that $J_1 \geq J_2$ (where the values J_1 and J_2 correspond to distributions G_1 and G_2). From (7.13) and (7.15) one gets that

$$\mathrm{P}^i\{W > 0\} \;=\; 1 - \sum_{j=0}^{n-1} \pi_j^i \;=\; 1 - \pi_0^i \cdot \sum_{j=0}^{n-1} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}, \qquad i = 1, 2,$$

is an increasing function of J_i (given λ, μ and n are fixed). Therefore, the relation $\mathrm{P}^1\{W > 0\} \;\geq\; \mathrm{P}^2\{W > 0\}$ prevails.

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 (7.13) and (7.15) imply that $\pi_j^1 \leq \pi_j^2, \;\; 0 \leq j \leq n-1$. Therefore, the inequality for utilizations holds.

Finally, the inequality $\mathrm{P}^1\{\mathrm{Ab}|W > 0\} \;\leq\; \mathrm{P}^2\{\mathrm{Ab}|W > 0\}$ follows from the definition of conditional probability and the two previous statements.

Proof of Theorem 1.

a+b. Define a function

$$H(x) \;=\; \int_0^x \bar{G}(\eta) d\eta \,.$$

Lemma 1 implies that if the deterministic distribution G_d uniformly maximizes $H(x)$ over all possible survival functions with mean $\bar{R}$, then **a+b** holds. Note that $H(x)$ has the following properties:

$$\begin{cases} H(0) \;=\; 0, \\ H(\infty) \;=\; \bar{R}, \\ H(x) \;\text{ is non-decreasing}, \quad x \geq 0, \\ H'(0) \;\leq\; 1\,, \\ H'(x) \;\text{ is non-increasing}, \quad x \geq 0. \end{cases} \tag{7.18}$$

The uniformly maximal survival function that satisfies the constraints (7.18) is:

$$H(x) \;=\; \min(x, \bar{R}), \quad x \geq 0\,, \tag{7.19}$$

which corresponds to the survival function of the deterministic distribution: $\bar{G}(\eta) = I\{\eta < \bar{R}\}$.

c+d. The bulk of the proof is showing that deterministic patience maximizes $\mathrm{E}[W|W>0]$. Then the statement for the average wait will follow from Lemma 1, part **a**. In addition, Little's formula will imply **d**.

We start with deriving an expression for the conditional average wait, which is convenient to analyze:

$$\mathrm{E}[W|W>0] \;\triangleq\; W_0(H) \;=\; \frac{\int_0^\infty H(t)\cdot\exp\{\lambda H(t) - n\mu t\}dt}{\int_0^\infty \exp\{\lambda H(t) - n\mu t\}dt}\,, \tag{7.20}$$

where

$$H(t) = \int_0^t \bar{G}(u)du\,. \tag{7.21}$$

Proof of (7.20). According to formula (7.14), the survival function of the virtual wait is given by

$$\bar{V}(t) \;=\; \lambda\pi_{n-1}\int_t^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}dx$$

Hence, the average wait is equal to

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

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 we note from (7.14) and (7.16) that $\mathrm{P}\{W>0\} = \lambda\pi_{n-1}J$ and derive (7.20). ■

Now we can formulate the optimization problem. Define by S the set of functions that satisfy constraints (7.18). We need to prove that

$$H^D(x) \;\triangleq\; \min(x,\bar{R}) \;=\; \arg\max_{H\in S} W_0(H)$$

where $W_0(H)$ is defined by (7.20). It can be verified that W_0 is continuous on S in the uniform metrics

$$\rho(H_1,H_2) \;=\; \max_{0\le t\le\infty}|H_1(t) - H_2(t)|\,.$$

(Check the numerator and the denominator of (7.20) separately.)

Introduce

$$H_b^D(x) \;\triangleq\; \min(x/b,\bar{R}),\quad b>1, \tag{7.22}$$

which is the family of functions that, as can be easily checked using (7.21), corresponds to patience distributions that take values $b\bar{R}$ 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 $b\bar{R}$.

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:

$$\delta W_0(H, \delta H) \stackrel{\Delta}{=} \frac{\partial}{\partial \alpha} \left[W_0 \left(H(t) + \alpha \delta H(t) \right) \right]_{\alpha = 0} \tag{7.23}$$

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 (7.23): 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

$$\begin{aligned} \delta W_0(H, \delta H) &= \frac{\partial}{\partial \alpha} \left[\frac{\int_0^\infty (H(t) + \alpha \delta H(t)) \cdot \exp\{\lambda (H(t) + \alpha \delta H(t)) - n\mu t\} dt}{\int_0^\infty \exp\{\lambda (H(t) + \alpha \delta H(t)) - n\mu t\} dt} \right]_{\alpha = 0} \\ &= \int_0^\infty \delta H(t) \cdot \exp\{\lambda H(t) - n\mu t\} \cdot r(t) dt \, , \end{aligned} \tag{7.24}$$

where

$$r(t) = \int_0^\infty (1 + \lambda H(t) - \lambda H(x)) \cdot \exp\{\lambda H(x) - n\mu x\} dx \tag{7.25}$$

Note that $r(t)$ is an increasing function (in fact, strictly increasing at points t such that $H(t) \neq \bar{R}$). 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

$$(H + \delta H_1)(t) = \begin{cases} (tH(t^*))/t^*, & 0 \leq t \leq t^* \\ H(t), & t > t^* \end{cases} \tag{7.26}$$

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 (7.24), $\delta W_0(H,\delta H_1)>0$ 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

$$(H+\delta H_2)(t)=\begin{cases} H(t), & 0\le t\le t^* \\ \min(H(t^*)+h\cdot(t-t^*),\bar{R}), & t>t^* \end{cases} \tag{7.27}$$

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 $\bar{R}$) constitute exactly the class H_b^D. ■

Proof of Step 2. We must maximize the functional defined in (7.20) over the one-parameter family of the functions H_b^D, introduced in (7.22). 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 (7.20) and (7.22) shows that, if $\lambda-n\mu b\neq 0$,

$$\begin{aligned}\frac{\partial}{\partial b}W_0(H_b^D) &= \frac{\partial}{\partial b}\frac{\frac{1}{b}\int_0^{b\bar{R}} t\cdot\exp\left\{\frac{(\lambda-n\mu b)t}{b}\right\}dt+\int_{b\bar{R}}^{\infty}\bar{R}\cdot\exp\{\lambda\bar{R}-n\mu t\}dt}{\int_0^{b\bar{R}}\exp\left\{\frac{(\lambda-n\mu b)t}{b}\right\}dt+\int_{b\bar{R}}^{\infty}\exp\{\lambda\bar{R}-n\mu t\}dt} \\ &= \frac{\partial}{\partial b}\frac{\left[\frac{b\bar{R}}{\lambda-n\mu b}\cdot e^{\bar{R}(\lambda-n\mu b)}-\frac{b}{(\lambda-n\mu b)^2}\cdot e^{\bar{R}(\lambda-n\mu b)}+\frac{b}{(\lambda-n\mu b)^2}\right]+\frac{\bar{R}}{n\mu}e^{\bar{R}(\lambda-n\mu b)}}{\frac{b}{\lambda-n\mu b}\cdot\left[e^{\bar{R}(\lambda-n\mu b)}-1\right]+\frac{e^{\bar{R}(\lambda-n\mu b)}}{n\mu}} \\ &\triangleq \frac{\partial}{\partial b}\frac{f(b)}{g(b)},\end{aligned}$$

where

$$f(b) = \left[\frac{\lambda\bar{R}}{(\lambda-n\mu b)n\mu}-\frac{b}{(\lambda-n\mu b)^2}\right]e^{\bar{R}(\lambda-n\mu b)}+\frac{b}{(\lambda-n\mu b)^2} \tag{7.28}$$

and

$$g(b) = \frac{1}{\lambda-n\mu b}\left[\frac{\lambda}{n\mu}e^{\bar{R}(\lambda-n\mu b)}-b\right]. \tag{7.29}$$

We shall need also the derivatives:

$$f^{'}(b) = e^{\bar{R}(\lambda-n\mu b)}\left[\frac{(\lambda+n\mu b)\bar{R}}{(\lambda-n\mu b)^2}-\frac{\lambda\bar{R}^2}{\lambda-n\mu b}-\frac{\lambda+n\mu b}{(\lambda-n\mu b)^3}\right]+\frac{\lambda+n\mu b}{(\lambda-n\mu b)^3}$$

and

$$g^{'}(b) = e^{\bar{R}(\lambda-n\mu b)}\left[\frac{\lambda}{(\lambda-n\mu b)^2}-\frac{\lambda\bar{R}}{\lambda-n\mu b}\right]-\frac{\lambda}{(\lambda-n\mu b)^2}.$$

One must show that

$$f(b)g^{'}(b)-f^{'}(b)g(b) \ge 0, \quad b\ge 1. \tag{7.30}$$

Then some algebra provides us with:

$$(\lambda-n\mu b)^2\cdot[f(b)g^{'}(b)-f^{'}(b)g(b)] =$$

$$\frac{\lambda^2}{n\mu(\lambda-n\mu b)^2}\cdot e^{2\bar{R}(\lambda-n\mu b)}-\left[b\lambda\bar{R}^2+b\bar{R}+\frac{\lambda^2+n^2\mu^2b^2}{n\mu(\lambda-n\mu b)^2}+\frac{\lambda\bar{R}}{n\mu}\right]\cdot e^{\bar{R}(\lambda-n\mu b)}+\frac{n\mu b^2}{(\lambda-n\mu b)^2}$$

Multiply the last expression by $n\mu(\lambda-n\mu b)^2$ and denote $\tilde{\mu}=n\mu b$. We then get

$$\lambda^2e^{2\bar{R}(\lambda-\tilde{\mu})}-[\lambda\tilde{\mu}\bar{R}^2(\lambda-\tilde{\mu})^2+\tilde{\mu}\bar{R}(\lambda-\tilde{\mu})^2+\lambda^2+\tilde{\mu}^2+\lambda\bar{R}(\lambda-\tilde{\mu})^2]\cdot e^{\bar{R}(\lambda-\tilde{\mu})}+\tilde{\mu}^2\,.$$

Now we change variables: $x=\lambda-\tilde{\mu}$ (note that $x>-\tilde{\mu}$) and transform the last expression to

$$(x+\tilde{\mu})^2e^{2\bar{R}x}-[\tilde{\mu}(x+\tilde{\mu})^2\bar{R}^2x^2+\tilde{\mu}\bar{R}x^2+(x+\tilde{\mu})^2+\tilde{\mu}^2+(x+\tilde{\mu})\bar{R}x^2]\cdot e^{\bar{R}x}+\tilde{\mu}^2\,. \tag{7.31}$$

It is easy to check that (7.31) is zero for $x=0$. Differentiating it with respect to x, we get

$$e^{\bar{R}x}(e^{\bar{R}x}\cdot r(x)-h_2(x))\,,$$

where

$$r(x) \;=\; 2\tilde{\mu}+2\tilde{\mu}^2\bar{R}+(2+4\tilde{\mu}\bar{R})x+2\bar{R}x^2$$

and

$$h_2(x) \;=\; 2\tilde{\mu}+2\tilde{\mu}^2\bar{R}+(2+6\tilde{\mu}\bar{R}+2\tilde{\mu}^2\bar{R}^2)x+(\tilde{\mu}^2\bar{R}^3+4\bar{R}+5\tilde{\mu}\bar{R}^2)x^2+(\tilde{\mu}\bar{R}^3+\bar{R}^2)x^3\,.$$

Let $h_1(x)\stackrel{\Delta}{=}e^{\bar{R}x}\cdot r(x)$. Assume $\boldsymbol{x>0}$. Then

$$e^{\bar{R}x} \;>\; 1+\bar{R}x+\frac{\bar{R}^2x^2}{2}$$

and

$$h_1(x) \;>\; r(x)\cdot\left(1+\bar{R}x+\frac{\bar{R}^2x^2}{2}\right) \;=\; h_2(x)+(\tilde{\mu}\bar{R}^3+2\bar{R}^2)x^3+\bar{R}^3x^4 \;>\; h_2(x)\,. \tag{7.32}$$

If $\boldsymbol{x<0}$, then

$$e^{\bar{R}x} \;<\; 1+\bar{R}x+\frac{\bar{R}^2x^2}{2}$$

and

$$h_1(x) \;<\; h_2(x)+2\bar{R}^2x^3+\bar{R}^3x^3(\tilde{\mu}+x)<h_2(x)\,, \tag{7.33}$$

where the last inequality follows from $-\tilde{\mu}<x<0$. According to (7.32) and (7.33), the derivative of expression (7.31) is negative for $x<0$ and positive for $x>0$. Hence (7.31) is non-negative, implying (7.30) and, in turn, Step 2.

The calculations above are not valid for $b=\dfrac{\lambda}{n\mu}$ (due to division by zero in (7.28) and (7.29)). 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

$$H\in\tilde{S} \;\text{ iff }\; \exists\,T \;\text{ s.t. }\; H(T)=\bar{R}\,,$$

and introduce $\tilde{S}_T$ by

$$H \in \tilde{S}_T \text{ iff } H(T) = \bar{R}.$$

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 [17], 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 (7.26) and (7.27), 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 $T \geq \bar{R}$. 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. ■

7.3 Summary of relevant results from Brandt and Brandt [4, 5]

Brandt and Brandt [4, 5] 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. We provide here an M/M/n+G version of [4, 5], adapting their notation to our needs. Since the steady-state distribution of the number-in-system is calculated, the basic Markov process definition in [4, 5] is more complicated than in Baccelli and Hebuterne [1].

Model. Brandt and Brandt assume that patience time has a survival function $\bar{G}$ with continuous density c. (We believe that many results from [4, 5] remain true also for a general patience distribution.) 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 [4, 5] is:

$$(N(t); X_1(t), \ldots, X_{L(t)}(t); I_1(t), \ldots, I_{L(t)}(t)).$$

Its stationary distribution is denoted by

$$P_k(x_1, \ldots, x_l; u_1, \ldots, u_l) \stackrel{\Delta}{=} \tag{7.34}$$

$$\stackrel{\Delta}{=} \lim_{t \to \infty} \mathrm{P}\{N(t) = k; X_1(t) \leq x_1, \ldots, X_{L(t)}(t) \leq x_l; I_1(t) \leq u_1, \ldots, I_{L(t)}(t) \leq u_l\}.$$

Due to the FCFS (First Come First Served) service discipline, the support of distribution (7.34) is contained in

$$\Omega_l \triangleq \{(x_1,\ldots,x_l;u_1,\ldots,u_l)\in R_{2l}^+ : u_1-x_1\geq\ldots\geq u_l-x_l\geq 0\}\,.$$

Define the stationary density (which exists)

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

Relevant results. Introduce the function

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

and the constants

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

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

$$g^{-1} = \sum_{j=0}^{n-1}\frac{n!\cdot\lambda^j\mu^{n-j}}{j!}+\sum_{j=0}^{\infty}\lambda^{n+j}F_j\,. \tag{7.35}$$

Then the steady-state distribution is given by

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

$$\pi_k = g\lambda^k F_{k-n},\quad k>n\,, \tag{7.36}$$

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

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

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

In [4], the density for the waiting time distribution was also derived.

The formal definition of the abandonment rates α_l, introduced in Section 3 and used in Subsection 5.3, is the following:

$$\alpha_l \triangleq \frac{\sum_{i=1}^{l}\int_{R_+^{2l-1}}\pi_{n+l}(x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_l;\,u_1,\ldots,u_l)dx_1\ldots dx_{i-1}dx_{i+1}\ldots dx_l du_1\ldots du_l}{\pi_{n+l}} \tag{7.37}$$

It has been shown in [5] that

$$\alpha_l = \frac{F_{l-1}}{F_l}-n\mu\,.$$

7.4 Proof of Lemma 2

We use the results from Brandt and Brandt [4, 5]. 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 (7.35) 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 (7.36) that $\pi_{n+l} = g \cdot \lambda^{n+l} F_l,\ \ l > 0$. Little's formula implies that the average waiting time is

$$\mathrm{E}[W] \;=\; \frac{\sum_{l=1}^{\infty} l \cdot \pi_{n+l}}{\lambda} \;=\; g\lambda^{n-1} \sum_{l=1}^{\infty} \lambda^l F_l \;=\; \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n F_1 + o(\lambda^n)\,. \tag{7.38}$$

Applying integration by parts:

$$F_1 \;\stackrel{\Delta}{=}\; \int_0^{\infty} F(\xi) e^{-\xi} d\xi \;=\; \int_0^{\infty} \bar{G}(x) e^{-n\mu x} dx\,.$$

From (3.2) and (7.36), the probability to abandon is

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{\sum_{l=1}^{\infty} \alpha_l \cdot \pi_{n+l}}{\lambda} \;=\; \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \alpha_1 F_1 + o(\lambda^n)\,. \tag{7.39}$$

In order to validate the second equality of (7.39), note that from formula (3.3),

$$\alpha_l \;=\; \frac{l}{\mathrm{E}[R]} + o(l)\,, \qquad (l \to \infty)\,.$$

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 (7.39). Now using formulae (7.38) and (7.39), we prove (4.6).

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 (4.7), using (7.39):

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{P}\{W > 0\}} \;\sim\; \alpha_1 F_1 \;=\; 1 - n\mu F_1 \;=\; \mathrm{P}\{R < \exp(n\mu)\}\,, \qquad (\lambda \to 0)\,,$$

and one gets (4.8) using (7.38)

$$\mathrm{E}[W|W>0] \;=\; \frac{\mathrm{E}[W]}{\mathrm{P}\{W>0\}} \;\sim\; F_1 \;=\; \mathrm{E}[R \wedge \exp(n\mu)] \qquad (\lambda \to 0)\,.$$

The last two equations imply

$$\lim_{\lambda\to 0} \frac{\mathrm{P}_\lambda\{\mathrm{Ab}|W>0\}}{\mathrm{E}_\lambda[W|W>0]} \;=\; \alpha_1\,.$$

References

[1] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: Kylstra F.J. (Ed.) Performance '81. North-Holland Publishing Company, pp 159-179.

[2] Bhattacharya P.P. and Ephremides A. (1991) Stochastic monotonicity properties of multiserver queues with impatient customers. Journal of Applied Probability 28:673-682.

[3] Boxma O.J. and de Waal P.R. (1994) Multiserver queues with impatient customers. ITC 14:743-756.

[4] Brandt A. and Brandt M. (1999) On the $M(n)/M(n)/s$ queue with impatient calls. Performance Evaluation 35:1-18.

[5] 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.

[6] 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. Submitted to JASA. Available at
`http://iew3.technion.ac.il/serveng/References/references.html`.

[7] de Bruijn N.G. (1981) Asymptotic Methods in Analysis, Dover.

[8] Cox D.R. and Oakes D. (1984) Analysis of Survival Data, Chapman and Hall.

[9] Daley D.J. and Servi L.D. (2000) Estimating customer loss rates from transactional data. In: Shanthikumar J.G. and Sumita U. (eds) International Series in Operations Research and Management Science, pp 313-332.

[10] Friedman J. H. (1984) A variable span scatterplot smoother. Laboratory for Computational Statistics, Stanford University Technical Report No. 5.

[11] 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`.

[12] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a telephone call-center with impatient customers. Manufacturing and Service Operations Management 4:208-227.

[13] Gnedenko B.W. and Kovalenko I.N. (1968) Introduction to Queueing Theory, Jerusalem, Israel Program for Scientific Translations.

[14] Haugen R.B. and Skogan E. (1980) Queueing systems with stochastic time out. IEEE Trans. Commun. COM-28:1984-1989.

[15] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. Operations Research 29:567-588.

[16] 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.

[17] Kolmogorov A.N. and Fomin S.V. (1999) Elements of the Theory of Functions and Functional Analysis, Dover.

[18] Kort B.W. (1983) Models and methods for evaluating customer acceptance of telephone connections. GLOBECOM '83, IEEE:706-714.

[19] Mandelbaum A., Sakov A. and Zeltyn S. (2000) Empirical analysis of a call center. Technical report, Technion. Available at
`http://iew3.technion.ac.il/serveng/References/references.html`.

[20] Mandelbaum A. and Shimkin N. (2000) A model for rational abandonment from invisible queues. Queueing Systems: Theory and Applications (QUESTA) 36:141-173.

[21] 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`.

[22] Mandelbaum A. and Zeltyn S. (2004) Call centers with impatient customers: empirical and asymptotic analysis, working paper.

[23] Palm C. (1943) Intensitatsschwankungen im fernsprechverkehr. Ericsson Technics 44:1-189.

[24] Palm C. (1953) Methods of judging the annoyance caused by congestion. Tele 4:189-208.

[25] Peterson T.S. (1950) Elements of Calculus, Harper, New York.

[26] Riordan J. (1962) Stochastic Service Systems, Wiley.

[27] Roberts J.W. (1979) Recent observations of subscriber behavior. In Proceedings of the 9th International Tele-traffic Conference.

[28] Shimkin N. and Mandelbaum A. (2002) Rational abandonment from tele-queues: non-linear waiting costs with heterogeneous preferences. Submitted to QUESTA. Available at `http://iew3.technion.ac.il/serveng/References/references.html`.

[29] Whitt W. (2003) A diffusion approximation for the G/GI/n/m queue. To appear in Operations Research.

[30] 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.