DESIGNING A CALL CENTER

WITH AN IVR

(INTERACTIVE VOICE RESPONSE)

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE IN OPERATIONS
RESEARCH AND SYSTEMS ANALYSIS

KHUDYAKOV POLINA

Submitted to the Senate of the Technion - Israel Institute of Technology
Shvat 5766 Haifa February 2006

Contents

List of Tables

List of Figures

The research was carry out under supervision of Prof. Avishai Mandelbaum in the Faculty of Industrial Engineering Management.

The generous financial help of the Technion Graduate School is gratefully acknowledged.

Abstract

A call center is a popular term for a service operation that handles telephone calls of customers. A call center typically consists of agents that handle incoming calls, telephone trunk lines, an Interactive Voice Response (IVR) unit, and a switch that routes calls to agents.

The subject of this thesis is a Markovian model for a call center with an IVR. We calculate operational performance measures, such as the probability for a busy signal and average wait for an agent. The calculations of these measures are cumbersome and they lack insight. We thus approximate the measures in an asymptotic regime known as QED (Quality Efficiency Regime), which is suitable for moderate to large call centers. The approximations are both insightful and easy to calculate (for up to 1000's of agents). They yield, as special cases, known approximations for the Erlang-B, Erlang-C and M/M/S/N queue.

Finally, we develop an algorithm for optimal staffing and trunk level. The algorithm is then used to analyze ways for reducing the operational costs of a call center, to understand the effect of a call center's size on its service level, and to investigate the effect of changes in system parameters on performance - for example, increasing IVR functionality (which would reasonably imply fewer but longer agent calls).

List of Symbols

λ arrival rate

θ IVR service rate

μ agent service rate

S number of agents

N number of trunk lines

ρ offered load per agent in the system for Call Center with IVR $\quad (\rho = \lambda p/(S\mu))$

R offered load (often $R = \lambda/\mu$ in Markovian queues)

$Q_1(t)$ the number of calls at the IVR

$Q_2(t)$ the number of calls at the agents pool (getting service and in queue)

$\pi(i, j)$ the stationary probabilities of having i calls at the IVR and j calls at the agents pool

$\chi(k, j)$ the probability that the system is in state (k, j), $(0 \leq j < k \leq N)$, when a call (among the $k - j$ customers) is about to finish its IVR service. Here, k is the total number of calls in the system, and j is the number of calls in the agents' pool (waiting or served); hence, $k - j$ is the number of calls at the IVR

$\Phi(\cdot)$ the standard normal distribution function

$\varphi(\cdot)$ the standard normal density function

E expectation

P probability measure

W waiting time after the IVR service, for a customer seeking service

$W(t)$ distribution function of the waiting time: $W(t) = P(W \leq t)$

$f_W(t)$ density function of the waiting time

$\approx$ $\qquad a_n \approx b_n$ if $a_n/b_n \to 1$, as $n \to \infty$

$\sim$ distributed as (for example, $X \sim Pois(\lambda)$ mean that X is a random variable that is Poisson distibuted with parameter λ)

o $a_n = o(b_n)$ if $a_n/b_n \to 0$, as $n \to \infty$

$L_W(x)$ Laplace transform of $f_W(t)$

$L_W^{-1}(t)$ inverse Laplace transform for the function $L_W(x)$

List of Acronyms

IVR Interactive Voice Response

ACD Automatic Call Distributor

ASA Average Speed of Answer

CTI Computer-Telephone Integration

FCFS First Come First Served

PASTA Poisson Arrivals See Time Averages

QED Quality and Efficiency Driven

Chapter 1

Introduction

[1]A call center is a popular term for a service operation that handles telephone calls of customers. Today all Fortune 500 companies have at least one call center [14]. Each employs an average of 4,500 agents across their sites. More than $300 billion is spent annually on call centers around the world [14]. Currently, call center operators represent 0.5% of the workforce in France, 1.5% in Great Britain and 4% in the USA [1]. Call centers are widely used, for the purpose of sales, service, and many other specialized transactions.

Rafaeli [27], in a recent research, reports that in Israel there are approximately 500 call centers. The number of employees in the 200 largest call centers is near 11,000 (about 1.8% of the workforce). Most Israeli call centers operate in banking, medical care, insurance, communication, tourism, transport, emergency services and the food industry.

Telephone calls are often characterized as either outbound and inbound. Inbound calls are initiated by the customer to obtain information, report a malfunction, ask for help or perform a business transaction. This is substantially different from outbound calls where the agent initiates the call to a customer, mostly with the aim to sell a product or a service to that customer, but sometimes also to provide information or to return a previous call by the customer.

Call centers have been aided by a range of telecommunications and computer technologies, including automatic call distribution (ACD), interactive voice response (IVR), and computer telephony integration (CTI), the latter allowing the actions of the computer to be synchronized with what is happening on the phone. In addition, customer relationship management (CRM) technologies, and other database systems, are heavily employed in call centers. Readers are referred to Gans, Koole and Mandelbaum [11] for explanations of call center terminology.

[1]Parts of Chapter 1 are adapted from [3].

A recent Purdue University study [8] revealed that 92% of US consumers form their image of a company based on their experience using the company's call center. More strikingly, the study found that 63% of the consumers stop using a company's products based on a negative call center experience. That number rises to almost 100% for consumers between ages 18 and 25.

Increased competition, deregulation and rising customer acquisition costs highlight the importance of both high-quality customer service and effective management of operating costs. To achieve both, most leading companies are deploying new technologies, such us enhanced interactive voice response (IVR), natural speech self-service options and others. The latest internet technologies allow "virtual" call centers to be established across a company's telecommunications network without physically having all the people in one office.

1.1 Call Centers with "self-service"

Technological progress has significantly affected the development of the call-center industry. Computer Telephony Integration (CTI) provides numerous opportunities for combining telephone service with e-mail and Internet services. Consequently, many call centers have evolved to so-called "contact centers", which serve customers through multi-media channels.

Automated Speech Recognition techniques help extend the number of tasks that Interactive Voice Response (IVR) units, traditionally touch tone, have been able to perform. Thus, IVR and Internet applications are becoming essential to the success of any contact center operation. When working properly, they can handle most inbound telephone transactions with a level of speed, efficiency and user-friendliness that is getting closer to matching human agents - and an IVR/CTI call costs only about 1/16 the costs of an agent-assisted call [3].

We now elaborate on the significance and scope that IVR and Internet applications play in contact centers. To repeat, IVR and Internet systems are specialized technologies designed to provide callers with verbal, fax or on-line answers to inquiries, without the assistance of people. This and other "self-service" technologies provide account information, fulfill requests for mailable items, pre-screen callers for script customization, interact with host systems (read and write), and produce reports.

Many organizations would benefit greatly from adding or enhancing "self-service" environments, while other companies' service quality would improve if their IVR/Internet system were removed. For example, one service provider could not afford the cost of staffing a 24-hour service environment or the cost of

outsourcing the off-hour service. However, it could afford to implement a "self-service" application for off-hour service. While the IVR/CTI could not answer 100% of the off-hour questions, it was able to handle 60% of these calls, decreasing call-back requests by 60%. The savings realized from the reduced call-backs paid for the system already during the first year [3].

Some consumers prefer to not deal with a human. eBay says that most of its customers are that way [18]: "Our members are very comfortable on the Internet, and an e-mail option or chat are sometimes preferred," quoting eBay's director of Customer Experience [18].

However, many customers get to hate "self-service" systems. They are the topic of angry anecdotes told around the coffee maker and the copy machine, and all too often, they are the reason that customers shift their loyalty from one company to another.

But it is not really the "self-service" systems themselves that the customers hate - it is the applications some companies develop for them. Often the scripts companies use are confusing to the customers and make it difficult for the customers to complete their interactions with the system. For example, according to [23], the top ten common errors in script design of IVR are the following:

1. *Too many choices.* Just because there are 12 buttons on the touch-tone pad does not mean they should all be used.

2. *Too many layers.* A layer is a set of menus that is connected to additional sets of choices. For example, if the first set of choices asks if the call is about appliances or furniture and the caller selects appliances, then the second layer of menus might list the various appliances and ask the caller to select one of them.

3. *Endless loops.* An endless loop occurs when the choices provided do not include an escape option that would take the caller to the agent (or an after-hours recording). So if callers do not hear a choice that corresponds to their issue, they just keep hearing that list over and over with no escape.

4. *Disconnect of caller.* If the caller does not make a selection the first time through a menu, many systems replay the options again. But in some cases, if the caller does not make a selection, the system simply disconnects. This is incredibly rude and frustrating for the caller.

5. *Use of industry jargon.* It is unreasonable to assume that callers will understand all of the unique terms and acronyms of one's businesses. So, if the caller is presented with a list of choices such as for HMO, press 1; for PPO, press 2; or for Indemnity, press 3 - it is likely that many will end up in the wrong system or agent group.

6. *Constantly changing menus.* Repeat callers become familiar with the numbers that correspond to their common choices and move through the menus with-

out listening to the lists. But if the options are changed, callers need to spend more time listening, which frustrates them and costs money for call center.

7. *Menu choices do not have expected results.* An example of this is the menu that asks the caller which language is preferred, but then connects the caller to an agent group that does not speak that language.

8. *Number first, menu item second.* The script should provide the description of the choice first and then tell the caller which digit to press. If the number is given first, the caller may forget which number it was by the time the right description is heard. This results in the caller having to repeat the menu, increasing both caller frustration and cost to the company.

9. *Unprofessional voices.* The scripts that are read to the caller should all be in the same voice and should be a voice that is easy to listen to with a neutral accent. Many companies have employees do these recordings and then mix up male and female voices as employees come and go.

10. *Unprofessional scripts.* This is more of a problem with speech recognition systems than with IVRs, but the "rule of thumb" is: do not get cute with the script. It might work in some companies with a fun-loving brand image, but it can come off as unprofessional.

The self-service technology itself is neutral - just as capable of making life easier as it is of causing frustration. And many companies use successfully voice "self-service" to create positive experiences that build customer loyalty.

If "self-service" systems seem to frustrate customers, why are so many companies using them? The answer is simple. Properly applied, "self-service" systems yield a range of business benefits that make them worthwhile. And properly applied, they increase rather then reduce customer loyalty. Here are three sound reasons for using "self-service" technology in the contact center [3].

•Improved customer satisfaction

Properly used, "self-service" applications can contribute to customer satisfaction is several ways. For one, they can reduce queue times. If there's anything customers like less than dealing with "self-service" prompts, it's waiting on hold until an agent becomes available. Studies [3] show that 37% of all typical contact center transactions are routine inquiries that can be easily automated. By using "self-service" systems to handle these transactions, the manager frees up agents to handle more of the kinds of transactions that require human service. Customers spend less time on hold and abandon fewer attempts to get in touch with the company.

Another advantage "self-service" applications offer to customers is extended service hours. Most contact centers can't afford to staff around the clock, but with self-service applications, the company can deliver cost-effective 24×7 service.

"Self-service" applications also offer privacy. There are some transactions that customers would prefer not to discuss with an agent. Customers who want to check if they have overdrawn their checking account might not want to ask an agent about it. A health care customer who is calling to get medical test result might be more comfortable hearing the results read by a text-to-speech application rather than by a human agent.

•Increased revenue

Customer "self-service" applications can also be revenue generators. Extended hours of service, for instance, also mean extended business hours. By letting customers use "self-service" systems to order products and services, companies create around-the-clock revenue streams.

The "self-service" systems can also become around-the-world revenue streams, because self-service applications extend companies' market reach. If the business depends entirely on human-service transactions, then to do business in other times zones a company either must maintain a 24×7 operation in one location or it must build and staff contact centers in other time zones.

Last but not least, by using a "self-service" system for routine information requests and simple service transactions, companies free their trained agents to concentrate on more complex calls, such as closing sales, cross-selling and up-selling.

•Reduced cost

The most often cited reason for using self-service applications is reduced costs. Staffing expenses typically account for between 60 and 70 percent of contact center costs [28]. The typical service phone call involving a real person costs a company $7. An Internet transaction, with a person responding, costs $2.25. But a "self-service" phone call with no human interaction costs less than 50 cents, according to the marketing director at TelephonyAtWork, a call center vendor [18].

Salaries are not the only expense associated with staffing. If agents do not have all the skills they need or don't meet a company's service expectations, they are costing money. If a company must compete with other contact centers to hire competent agents, and then raise salaries frequently to keep them, that costs money too. If a company's training costs amount to two or three months salary per agent and the turnover rate is 25% per year, it is spending significant sums without getting any return.

In summary, the use of "self-service" technologies has been increasing in a variety of industries, including banking, brokerage, insurance, sales and catalog houses. The "self-service" technology enables call centers to keep costs from rising (and sometimes to reduce costs), while improving service levels, revenue

and hence profits.

Chapter 2

Literature Review

2.1 The QED regime

The main mathematical framework considered in this thesis is the many-server heavy-traffic asymptotic regime identified already by Erlang [9] and Jagerman [19], but ultimately introduced and formalized by Halfin and Whitt [17]. We refer to this regime as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy a rare combination of very high efficiency together with very high quality of service, as surveyed in Gans, Koole and Mandelbaum [11]. This will be demonstrated in the present thesis as well.

Consider a sequence of S-server queues, indexed by n. Let the arrival-rates $\lambda_n \to \infty$, as $n \uparrow \infty$, and fix μ the service-rate. Define the offered load by $R_n = \frac{\lambda_n}{\mu}$. The QED regime is achieved by choosing λ_n and S_n so that $\sqrt{S_n}(1-\rho_n) \to \beta$, as $n \uparrow \infty$, for some finite β. Here $\rho_n = \frac{R_n}{S_n}$. When customers have infinite patience, ρ_n may be interpreted a the long-run servers' utilization and then one must have $0 < \beta < \infty$. Otherwise, ρ_n is the offered load per server and $-\infty < \beta < \infty$ is allowed. Equivalently, the staffing level is approximately given by

$$S_n \approx R_n + \beta\sqrt{R_n}, \qquad -\infty < \beta < \infty. \tag{2.1}$$

Another equivalent characterization of the QED regime is a non-trivial limit (within (0,1)) of the fraction of delayed customers. The latter equivalence was established for GI/M/S [17], GI/D/S [20] and M/M/S with exponential patience [13]. The staffing rule which appears in (2.1) has been called the *square-root staffing principle* (or sometimes "safety-staffing principle").

As mentioned, the QED regime was explicitly recognized already in Erlang's 1923 paper (that appeared in [9]) which addresses both Erlang-B (M/M/S/S) and Erlang-C (M/M/S) models. Later on, extensive related work took place in

various telecom companies but little has been openly documented, as in Sze [30] (who was actually motivated by AT&T call centers operating in the QED regime). A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given in Jagerman [13]; see also Whitt [31], and then Massey and Wallace [26] for the analysis of finite buffers. The balancing of "service and economy" via a non-trivial delay probability is the operational significance of the QED regime. It was first discovered and formalized by Halfin and Whitt [17]: within the GI/M/S framework, they analyzed the scaled number of customers, both in steady state and as a stochastic process. Puhalskii and Reiman in [31] established convergence of the scaled queueing process in the more general GI/PH/S setting, allowing also priorities, but not covering steady-state.

2.2 The square-root safety staffing principle

The square-root safety staffing principle has been part of the queueing-theory folklore for a long time. This is well documented by Grassmann [15, 16], and recently revisited by Green & Kolesar [24], where both its accuracy and applicability have been convincingly confirmed. The principle was substantiated by Whitt [31], then adapted in Jennings *et.al.* [21] to non-stationary models. All of this work applies infinite-server heuristics, grounded in the fact that the steady-state number of customers in the $M/M/\infty$ queue, say Q^{∞}, is Poisson distributed with mean $R = \frac{\lambda}{\mu}$, the offered load. It follows that Q^{∞} is approximately normally distributed, with mean R and standard deviation $\sqrt{R}$, when R is not too small [5].

The square-root principle has two parts to it: first, the conceptual observation that the safety staffing level is proportional to the square-root of the offered load; and second, the explicit calculation of the proportionality coefficient. Borst, Mandelbaum and Reiman [5] develop a framework that accommodates both of these needs. More important, however, is the fact that their approach and framework allow an arbitrary cost structure, having the potential to generalize beyond Erlang-C. For a concrete example, Garnett *et.al.* [13] accommodate impatient customers: in their main result, the square-root rule arises conceptually, but the determination of the value of the proportionality coefficient is left open. The square-root safety principle also arises in Massey and Wallace [26] for the M/M/S/N queue.

2.3 Analytical models of call center performance

In the detailed introduction to call centers by Gans, Koole and Mandelbaum [11], it is explained how call centers can be modeled by queueing systems of various characteristics. Many results and models with references are surveyed in that paper. The authors examine models of single type customers and single skill agents; models with busy signals and abandonment; skills-based routing; call blending and multi-media; and geographically dispersed call centers.

Figure 2.1 depicts a schematic model of a simple inbound call center with S agents serving one class of customers. A call at either the IVR or within the servers' pool occupies a trunk line. There are N trunk lines in this call center. As shown, the waiting room is limited to $N - S$ waiting positions and waiting customers may leave the system due to impatience. A blocked or abandoning customer might try to call again later (retrial). A queueing model of such an inbound call center is characterized by customer profiles, agent characteristics, queue discipline, and system capacity.

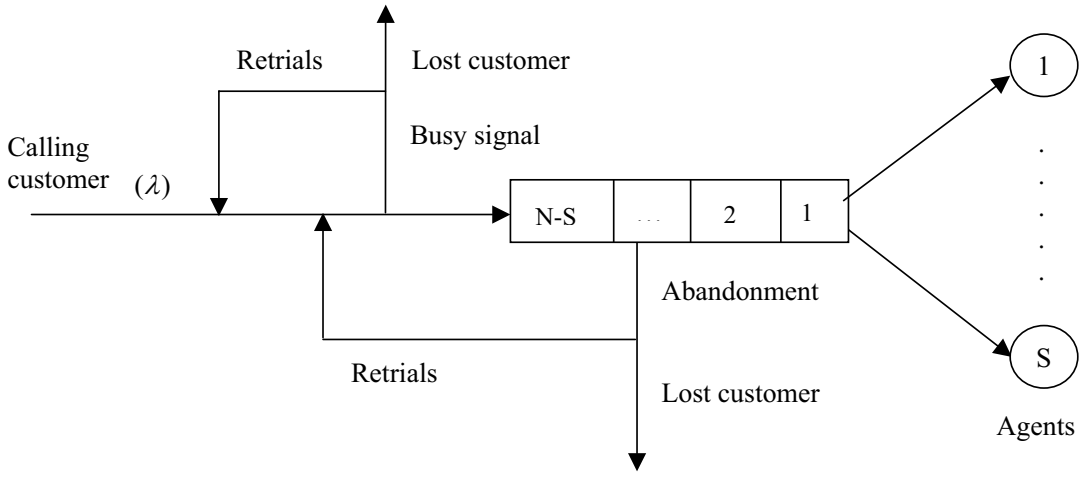

Figure 2.1: Schematic model of a call center with one class of impatient customers, busy signals, retrials and identical agents.

The simplest case with homogeneous customers and homogeneous agents is analytically tractable only if one assumes Poisson arrivals, exponential service times and no retrials. With these assumptions, the underlying stochastic processes are one-dimensional Markov processes, i.e., the future behavior is conditionally independent of the past, given the present state. Figure 2.2 depicts a

schematic model of such a queueing system.

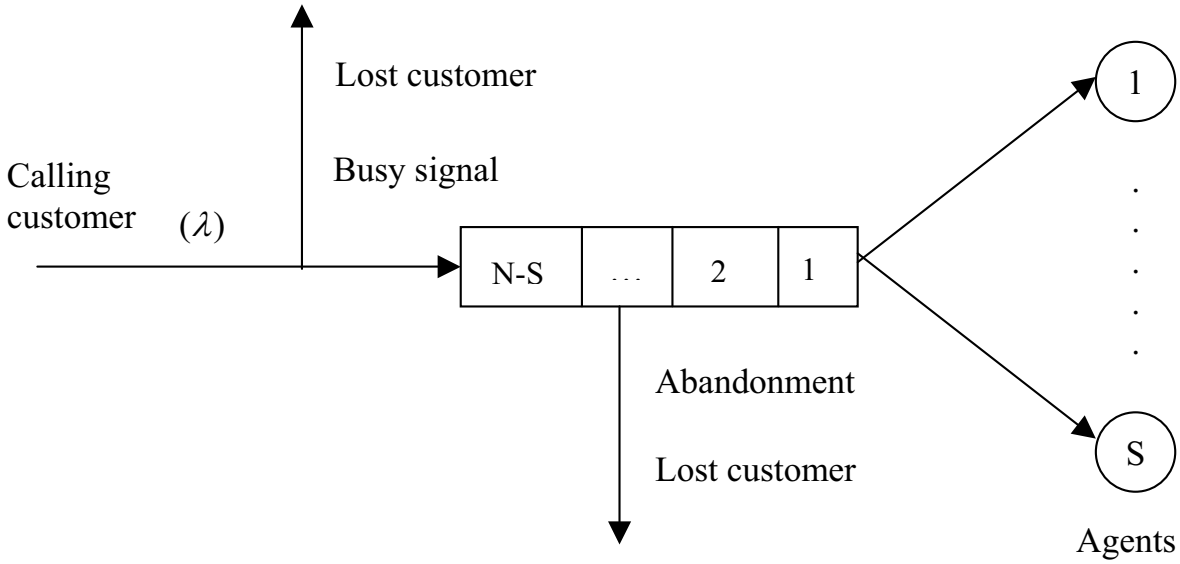

Figure 2.2: Schematic model of a queueing system with busy signals, impatient customers but without retrials.

The basic operational question in the design of call centers is: "How to provide an acceptable quality of service with the least costs?", or "How many agents and trunk lines do we need in order to provide a given service level?" In general: "How to trade off service quality with operational efficiency?"

Frequently used measures which support decision-making are average of the waiting time in queue, the probability to encounter a busy signal, the probability of wait, agents' occupancy, etc. In order to analyze the staffing problem, analytical models have been developed in order to help find the answer. The most widely used model is M/M/S, which is also known as Erlang-C. In this model, the arrival process is Poisson, the service time distribution is exponential and there are S independent, statistically identical agents. It is the simplest yet most prevalent model that supports call center staffing. Borst, Mandelbaum and Reiman [5] determined the asymptotically optimal staffing level S that trades off agents' costs with service quality. They developed this rule for three regimes of operation: quality-driven, where the focus is on service quality; efficiency-driven, which emphasizes agents' costs; and a rationalized regime that balances, and in fact unifies, the other two.

The M/M/S model allows an unbounded number of customers in the system, but in practice this number is bounded by the number of trunk lines. This

gives rise to the model M/M/S/N (when S=N it is called Erlang-B). Massey and Wallace [26] proposed a procedure for determining the appropriate number of agents S and telephone trunk lines N needed by call centers. They construct a new efficient search method for the optimal S and N-S that satisfy a given set of SLA (Service Level Agreement) metrics. Moreover, they develop a second approximate algorithm using steady-state, QED-based asymptotic analysis that in practice is much faster than the search method. The asymptotically derived number of agents and the number of waiting spaces in the buffer are found by iteratively solving a fixed point equation.

Analytical models of a Call Center with an IVR were developed by Brandt, Brandt, Spahl and Weber [6]. They show, and we shall use this fact later on, that it is possible to replace the open network of their model with a closed Jackson network. This latter model has the well known product form solution for its stationary distribution. Such a product-form distribution was used by Srinivasan, Talim and Wang in [29] in order to find expressions for the probability to find the system busy and the conditional distribution function of the waiting time before service.

Chapter 3

Call Centers with an IVR

3.1 Model description

As mentioned already, a call center typically consists of telephone trunk lines, a switching machine known as the automatic call distributor (ACD), an interactive voice response (IVR) unit and agents to handle the incoming calls.

We consider the following model of a call center: The arrival process is a Poisson process with rate λ. There are N trunk lines and S agents in the system ($S \leq N$). First the customer is served by the IVR processor. We assume that the IVR processing times are independent and identically distributed exponential random variables with rate θ. After finishing the IVR process, a call may leave the system with probability $1-p$ or request service from an agent with probability p.

We assume that there is no abandonment in our model. Agents' service times are considered as independent identically distributed exponential random variables with rate μ, which are independent of the arrival times and IVR processing times. If the call finds the system full, i.e. all N trunk lines are busy, it is lost. So we consider our model as a system with two multi-server queues connected in series. The first one represents the IVR processor. This processor can handle at most N jobs at a time, where N represents the total number of trunk lines available. The second queue represents the agents pool which can handle at most S incoming calls at a time. The number of agents is naturally less than the number of trunk lines available, i.e. $S \leq N$. Moreover, N is also an upper bound for the total number of customers in the system: IVR plus waiting to be served or are served by the agents.

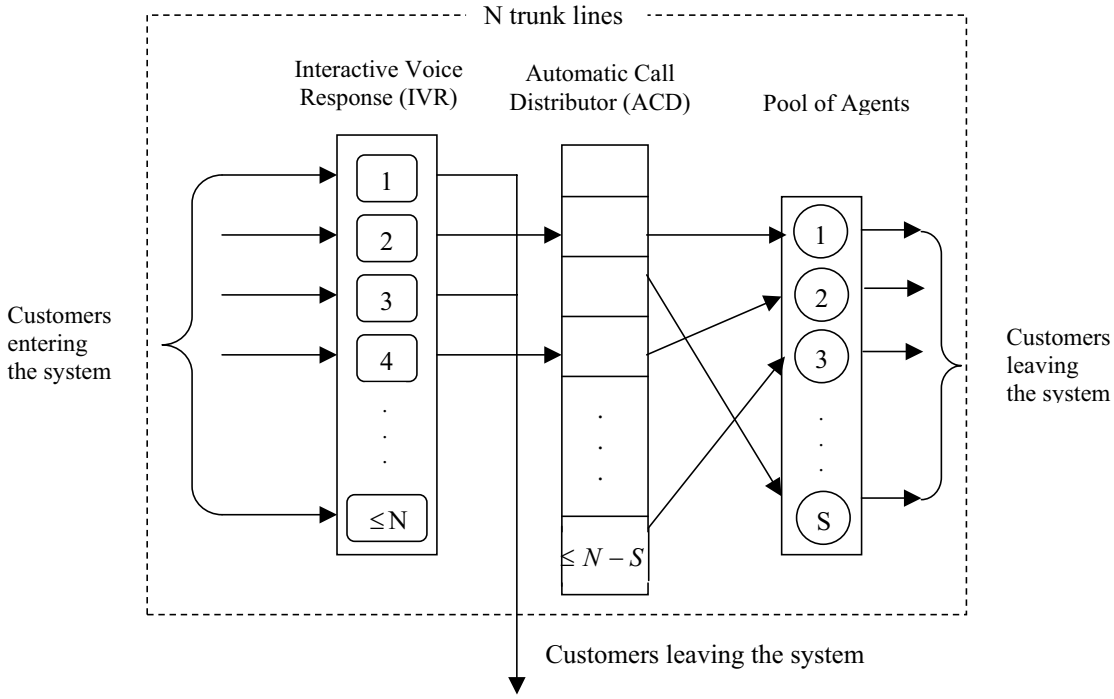

Figure 3.1: Schematic model of a call center with an interactive voice response, S agents and N trunk lines.

Let $Q(t) = (Q_1(t), Q_2(t))$ represent the number of calls at the IVR processor and the agents pool, respectively. Since there are only N trunk lines then $Q_1(t) + Q_2(t) \leq N$, for all $t \geq 0$. Note that the stochastic process Q is a finite-state continuous-time Markov chain. We shall denote its states by the pairs $\{(i,j)|i + j \leq N,\ \ i,j \geq 0\}$.

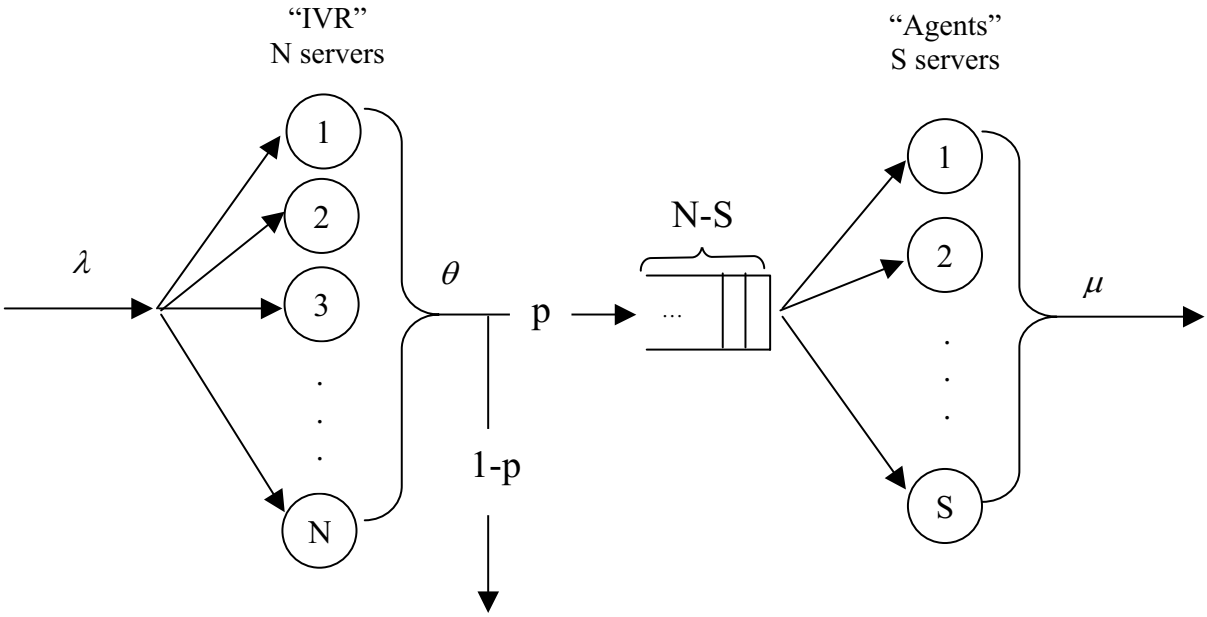

Figure 3.2: Schematic model of a queueing system with an interactive voice response, S agents and N trunk lines.

As shown in [6], one can consider the model as a closed Jackson network, by introducing a fictitious state-dependent queue. Let us look more carefully at this procedure. We can describe our original model as an open queueing network with two nodes and state-dependent arrival and service rates. The first node is an infinite-server node that models the IVR, whereas the second node models the S servers and the queue in front of them. The service times in the first and the second nodes are exponential with rates θ and μ respectively. The system accepts new arrivals only if $Q_1(t) + Q_2(t) < N$. This inequality holds if and only if there is at least one trunk line available. Note that we will say that the system is in the state $(i, j), 0 \leq i, j \leq N$ and $i + j \leq N$, when it contains exactly i calls in the IVR and j in the agents pool. So, the process of accepted customers by the system has the intensity:

$$\lambda(i,j) = \begin{cases} \lambda, & for \;\; i+j < N; \\ 0, & otherwise. \end{cases} \tag{3.1}$$

Since the second node models the customers that are waiting or being served, we can define the service rate $\mu(j)$ in node 2 as being dependent on the number

j of customers in node 2:

$$\mu(j) = min(j, S)\mu, \tag{3.2}$$

where $min(j, S)\mu$ is the departure rate of the actually served customers.

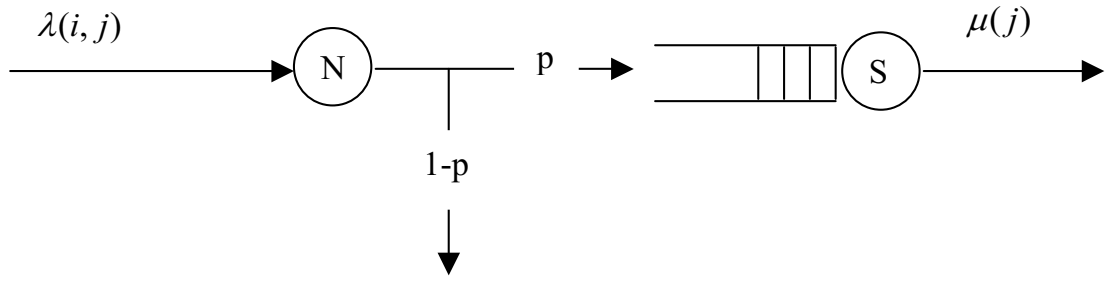

Figure 3.3: Schematic model of a queueing system with an interactive voice response, S agents and N trunk lines.

For the open network model, exact calculation of performance measures is difficult because the network is not of a product form type. Therefore, we replace the open network with the following closed three-node network:

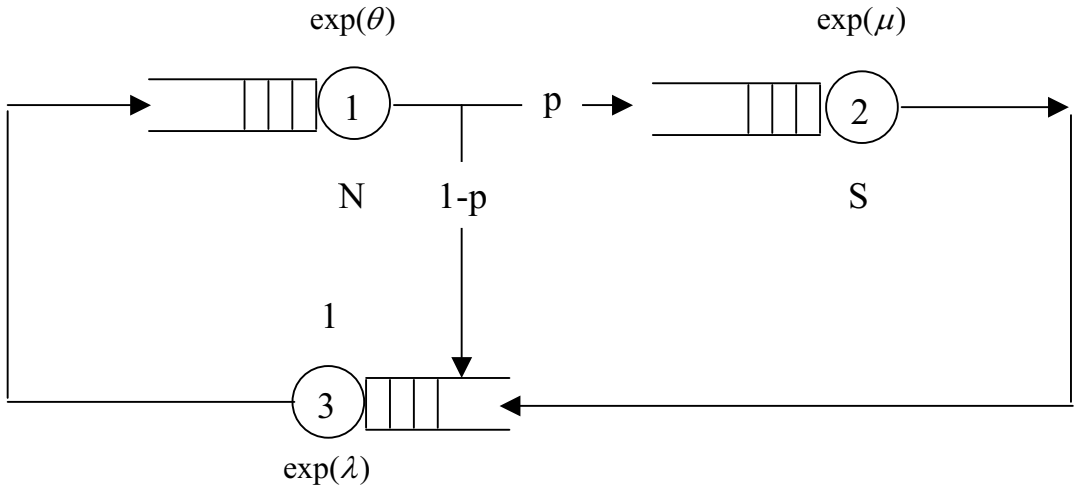

Figure 3.4: Schematic model of a corresponding closed Jackson network.

The state-dependent arrival rate is simply modeled by having only N entities circulating in the network. Service time in the first, second and third nodes are

exponential with rates θ, μ and λ respectively, and the numbers of servers are N, S and 1 in these nodes. So we can consider our model as a three node closed Jackson network, which is well known to have the following product form solution for its stationary distribution:

$$\pi^0(i,j,k) = \begin{cases} \dfrac{\pi^1(\alpha)\pi^2(\beta)\pi^3(\gamma)}{\sum\limits_{i+j+k=N} \pi^1(i)\pi^2(j)\pi^3(k)}, & \alpha+\beta+\gamma = N; \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

where $\pi^l(i)$ is the steady state probability for node l, $l = 1, 2, 3$ (M/M/N, M/M/S and M/M/1 respectively).

So, the stationary probabilities $\pi(i,j)$ of having i calls at the IVR and j calls at the agents pool can be written in a product form as follows:

$$\pi(i,j) = \begin{cases} \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{j!}\left(\dfrac{p\lambda}{\mu}\right)^j, & j \le S,\ 0 \le i+j \le N; \\ \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{S!S^{j-S}}\left(\dfrac{p\lambda}{\mu}\right)^j & j \ge S,\ 0 \le i+j \le N; \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

where

$$\pi_0 = \left(\sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{i+j\le N, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j \right)^{-1} \tag{3.5}$$

3.2 Description and Derivation of Performance Measures

3.2.1 Distribution of the waiting time

An important dimension of the service quality of an inbound call center is the waiting time of its customers. Define the waiting time W as the time spent by customers, who opt for service, from just after they finish the IVR process until they start service by an agent. Now we calculate the density function of W, following the partially heuristic derivation of Srinivasan, Talim and Wang [29].

We say that the system is in state (k, j), $0 \le j \le k \le N$, when it contains exactly k calls, and j is the number of calls in the agents' pool (waiting or served);

hence, $k-j$ is the number of calls in the IVR. Let $\chi(k,j)$, $0 \leq j < k \leq N$, be the probability that the system is in state (k,j), given that a call (among the $k-j$ customers) is about to finish its IVR service. Let A denote the event "a call is about to leave the IVR". Then using Bayes Theorem we get:

$$\begin{aligned}\chi(k,j) :&= \lim_{t\to\infty} P(\ Q(t) = (k-j,j)|A) = \\ &= \lim_{t\to\infty} \frac{P(A|\ Q(t) = (k-j,j))\, P(\ Q(t) = (k-j,j))}{\sum_{l=0}^{N}\sum_{m=0}^{l} P(\ A|\ Q(t) = (l-m,m))\, P(\ Q(t) = (l-m,m))}\end{aligned} \tag{3.6}$$

Each term in (3.6) can be rewritten as

$$\begin{aligned}&\lim_{t\to\infty} P(\ Q(t) = (k-j,j)|A) \\ &\quad= \lim_{t\to\infty,\varepsilon\to 0} \frac{1}{\varepsilon} P\left(Q(t) = (k-j,j) \,\middle|\, \begin{array}{l} Q(t+\varepsilon) = (k-j-1,j)\, or \\ Q(t+\varepsilon) = (k-j-1,j+1) \end{array}\right) \\ &\quad= \lim_{\varepsilon\to 0} \frac{(\varepsilon(\theta(k-j)(1-p)+o(\varepsilon)) + \varepsilon(\theta p(k-j)+o(\varepsilon)))\,\pi(k-j,j)}{\varepsilon} \\ &\quad= (k-j)\,\pi(k-j,j).\end{aligned} \tag{3.7}$$

It follows from (3.6) and (3.7) that

$$\chi(k,j) = \frac{(k-j)\,\pi(k-j,j)}{\sum_{l=0}^{N}\sum_{m=0}^{l}(l-m)\,\pi(l-m,m)}. \tag{3.8}$$

Let $W(0)$ be the probability that a call starts its service immediately after leaving the IVR. Then:

$$W(0) = \sum_{k=1}^{N}\sum_{j=0}^{min(k,S)-1} \chi(k,j). \tag{3.9}$$

The distribution function of the waiting time was found by Srinivasan, Talim and Wang in [29] and it is given by:

$$W(t) = 1 - \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)\sum_{l=0}^{j-S}\frac{(\mu St)^l e^{-\mu St}}{l!}. \tag{3.10}$$

The *expected waiting time* $E[W]$ can be derived via the tail's formula, i.e.,

$$\begin{aligned} E[W] &= \int_0^\infty (1 - W(t))dt = \int_0^\infty \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!} dt = \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \int_0^\infty \frac{(\mu S t)^l e^{-\mu S t}}{l!} dt = \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{1}{\mu S} = \\ &= \frac{1}{\mu S} \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j)(j - S + 1). \end{aligned} \tag{3.11}$$

The average waiting time for answered calls is often called *average speed of answer (ASA)*.

3.2.2 Probability of delay

One out of several measures of the waiting time is chosen to measure the *service level*. This measure is defined as the percentage of calls answered within a given waiting time limit. The service level is a widely used performance measure in call centers. The so-called "80/20 - standard service level" means that 80% of the customers should wait no more that 20 seconds. This 80% can also be interpreted as the individual probability of a randomly selected customers to wait at most 20 seconds. Hence, if W represents the waiting time of a customer, an X/Y service level can be interpreted as the probability $P(W \leq Y) = X$.

But the service level contains only information about upper bound of the waiting time for X percent of all customers. The remaining customers may have significantly larger waiting times than Y. Especially in cases of low service levels, we must analyze other performance measures.

In many cases, a very interesting measure is the expectation of waiting time $E[W]$, which we saw in the previous section. Also we wish to know the fraction of the customers that wait in the queue. We call this fraction *the probability of delay*, which is denoted and given by

$$P(W > 0) = \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \chi(i,j) \tag{3.12}$$

Equation (3.12) gives the conditional probability that a calling customer does not immediately reach an agent, given that the calling customer is not blocked, i.e., $P(W > 0)$ is the *probability of delay for served customers*.

Note the following interesting property of this probability.

Theorem 3.2.1 *For the system with N trunk lines and S agents, the fraction of customers, which are required to wait after their IVR service, coincides with the fraction that in a system with $N-1$ trunk lines and S agents, all agents are busy. Formally,*

$$P_N(W>0) = P_{N-1}(Q_2(\infty) \geq S) \tag{3.13}$$

PROOF. The probability of wait, following (3.12), is:

$$\begin{aligned}
P_N(W>0) &= \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j) = \sum_{i=1}^{N-S}\sum_{j=S}^{N-i}\chi(i,j) \\
&= \frac{\sum_{i=1}^{N-S}\sum_{j=S}^{N-i}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j}{\sum_{i+j\leq N,i\geq 1,j<S}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{i=1}^{N-S}\sum_{j=S}^{N-i}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \\
&= \frac{\frac{\lambda}{\theta}\sum_{i=1}^{N-S}\sum_{j=S}^{N-i}\frac{1}{(i-1)!}\left(\frac{\lambda}{\theta}\right)^{i-1}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j}{\frac{\lambda}{\theta}\left(\sum_{i+j\leq N,i\geq 1,j<S}\frac{1}{(i-1)!}\left(\frac{\lambda}{\theta}\right)^{i-1}\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{i=1}^{N-S}\sum_{j=S}^{N-i}\frac{1}{(i-1)!}\left(\frac{\lambda}{\theta}\right)^{i-1}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j\right)} \\
&= \frac{\sum_{m=0}^{N-S-1}\sum_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j}{\sum_{m+j\leq N-1,j<S}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{m=0}^{N-S-1}\sum_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \\
&= \sum_{m=0}^{N-S-1}\sum_{j=S}^{N-m-1}\pi(m,j),
\end{aligned}$$

where $\pi(m,j)$ is the stationary probability given in (3.4). Thus, it is now easy to see that

$$P(W>0) = P_{N-1}(Q_2(\infty)\geq S), \tag{3.14}$$

which proves Theorem 3.2.1. □

Thus our conditional probabilities can be reduced to unconditional probabilities.

3.2.3 Probability to find the system busy

The service level of a call center is often defined also in terms of the probability that an arriving call finds all trunk lines busy. Let us look first at the probability

P_k, $0 \leq k \leq N$, that there are exactly k calls in the system (processed by IVR, waiting for service or being serviced by an agent); it was found in [29] and has the following form:

$$P_k = \pi_0 \left(\frac{\lambda^k}{k!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^k + \sum_{j=S+1}^{k} \frac{1}{(k-j)!} \left(\frac{1}{S! S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{k-j} \left(\frac{p\lambda}{\mu} \right)^j I_{\{k \geq S+1\}} \right) \tag{3.15}$$

where $I_{\{k>S\}}$ is the indicator function.

One can apply these formula to derive P_N, the probability of having all trunk lines busy. This is also the loss probability due to the PASTA (Poisson Arrivals See Time Averages) property, see [32]. The probability that there are exactly N calls in the system takes the form

$$P_N = \pi_0 \left(\frac{\lambda^N}{N!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^N + \sum_{j=S+1}^{N} \frac{1}{(N-j)!} \left(\frac{1}{S! S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{N-j} \left(\frac{p\lambda}{\mu} \right)^j \right) \tag{3.16}$$

3.3 Other performance measures

In many cases, it is interesting to know the *expected queue length* $E[L]$, which can be derived via Little's formula, i.e.,

$$E[L] = \lambda_{eff} E[W] = p\lambda E[W](1 - P(block)). \tag{3.17}$$

The operating costs in call centers are mainly driven by the costs of the agents. Therefore, the *utilization* of the agents is often used as a technical measure to approximate the economic (efficiency) performance of an inbound call center. The expected utilization of the agents, say u, is the ratio of the effective arrival rate λ_{eff} for the pool of agents and the maximal service rate, i.e.,

$$u = \frac{\lambda_{eff}}{S\mu} = \frac{p\lambda(1 - P(block))}{S\mu}. \tag{3.18}$$

Chapter 4

Heavy traffic limits and asymptotic analysis

The ultimate goal of this chapter is to derive rules of thumb for solving the staffing and trunking problems for a call center with an IVR. This will be done analogously to Halfin and Whitt [17] and Massey and Wallace [26].

4.1 Our domain for asymptotic analysis

All the following approximations will be found for the case when the arrival rate λ tends to ∞. In order for the system not to be overwhelmed, assume that the number of agents S and the number of trunk lines N tend to infinity as well. But now, look more carefully at this problem. Which conditions do we need to assume further, in order to find an approximations for performance measures?

Let us consider the model of a call center with an IVR as an expanded model of the M/M/S/N queue. As mentioned before, this latter model was investigated by Massey and Wallace [26]. They found approximations of the performance measures of the M/M/S/N queue when λ, S and N tend to ∞ simultaneously and under the following assumptions:

$$\begin{aligned} &(i)\ \ N - S = \eta\sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), \quad 0 < \eta < \infty; \\ &(ii)\ \ S = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), \quad 0 < \beta < \infty; \end{aligned} \tag{4.1}$$

In these conditions, β is assumed positive because [26] used the M/M/S queue for finding the approximations. We shall dispose of this assumption momentarily.

For our call center with an IVR, according to assumptions (4.1), we need $S + \eta\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right)$ trunk lines for the queue in the agents' pool and $\frac{\lambda}{\theta} + \eta_2\sqrt{\frac{\lambda}{\theta}} + o\left(\sqrt{\lambda}\right)$ trunk lines for the IVR service pool. So, we can formulate the following conditions for our system. Let λ, S and N tend to ∞ simultaneously so that:

$$\begin{aligned} &(i)\ \ N - S = \eta_1\sqrt{\frac{\lambda p}{\mu}} + \frac{\lambda}{\theta} + \eta_2\sqrt{\frac{\lambda}{\theta}} + o\left(\sqrt{\lambda}\right), \ \ -\infty < \eta_1, \eta_2 < \infty; \\ &(ii)\ \ S = \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right), \ \ -\infty < \beta < \infty; \end{aligned} \tag{4.2}$$

Note, that we ignore the restrictions, that η_1, η_2 and β are positive and assume that they are of arbitrary sign. The disadvantage is that, in this case, we have three parameters η_1, η_2 and β. First we reduce the number of parameters to two.

Theorem 4.1.1 *Let λ, S and N tend to ∞ simultaneously. Then the conditions*

$$\begin{aligned} &(i)\ \ N - S = \eta_1\sqrt{\frac{\lambda p}{\mu}} + \frac{\lambda}{\theta} + \eta_2\sqrt{\frac{\lambda}{\theta}} + o\left(\sqrt{\lambda}\right), \ -\infty < \eta_1, \eta_2 < \infty; \\ &(ii)\ \ S = \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right), \ \ -\infty < \beta < \infty; \end{aligned} \tag{4.3}$$

are equivalent to the conditions

$$\begin{aligned} &(i)\ \ N - S = \eta\sqrt{\frac{\lambda}{\theta}} + \frac{\lambda}{\theta} + o\left(\sqrt{\lambda}\right), \ -\infty < \eta < \infty; \\ &(ii)\ \ S = \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right), \ \ -\infty < \beta < \infty; \end{aligned} \tag{4.4}$$

where $\eta = \eta_1\sqrt{\frac{p\theta}{\mu}} + \eta_2$.

Proof. Clearly, one can rewrite the first condition in (4.3) in the form

$$N - S = \sqrt{\frac{\lambda}{\mu}}\left(\eta_2 + \eta_1\sqrt{\frac{p\theta}{\mu}}\right) + \frac{\lambda}{\theta} + +o\left(\sqrt{\lambda}\right), \ -\infty < \eta_1, \eta_2 < \infty.$$

Setting $\eta = \eta_1\sqrt{\frac{p\theta}{\mu}} + \eta_2$ one obtains (4.4). The second condition is the same. This proves the statement. □

The conditions (4.4) have also the following equivalent form

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty; \end{aligned} \tag{4.5}$$

In this thesis we will find approximations of performance measures, where λ, S and N tend to ∞ simultaneously and under the conditions (4.5).

Note also that the asymptotic results, found in Halfin and Whitt [17] and Massey and Wallace [26], require strict positivity of β. In the case of Halfin and Whitt [17] this is clearly understandable, because they analyzed the M/M/S queue, which is unstable if $\beta \leq 0$. The analysis in Massey and Wallace [26] is based on a relationship between M/M/S and M/M/S/N. So, the reason for strict positivity of β in [26] is the same as in [17]. Note, that in Section (7.2) we will find approximations for the probability to wait and the probability to find the system busy in the M/M/S/N queue model when $-\infty \leq \beta \leq \infty$.

In the case of a call center with an IVR, we obtain a steady state regardless of the value of β. In other words, we can say that $-\infty < \beta < \infty$. But, as it turns out, in order to avoid technical problems in calculation, it is convenient to distinguish two cases:

1) $\beta \neq 0$;

2) $\beta = 0$.

4.2 Four auxiliary lemmas

In this section, we would like to formulate and prove some statements, which will be used when calculating approximations for performance measures.

Lemma 4.2.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then

$$\lim_{\lambda\to\infty} \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!} \left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\frac{p\lambda}{S\mu}} \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i = \frac{\varphi(\beta)\Phi(\eta)}{\beta}.$$

PROOF. For convenience, let us denote the expression, which we need to approximate, by ξ_1, i.e.

$$\xi_1 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\frac{p\lambda}{S\mu}} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i; \tag{4.6}$$

In view of Stirling's formula, $S! \approx \sqrt{2S\pi}S^S e^{-S}$, one obtains for ξ_1:

$$\xi_1 \approx \frac{e^{S-\lambda\frac{p}{\mu}}}{\sqrt{2S\pi}S^S}\left(\frac{p\lambda}{\mu}\right)^S \frac{\sqrt{S}}{\beta} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \tag{4.7}$$

The last sum can be rewritten as $P\left(X_\lambda \leq N-S-1\right)$ where $X_\lambda \sim Pois(\frac{\lambda}{\theta})$ is a random variable with the Poisson distribution with parameter $\frac{\lambda}{\theta}$, thus $E[X_\lambda] = \frac{\lambda}{\theta}$, $Var[X_\lambda] = \frac{\lambda}{\theta}$. If $\lambda \to \infty$, then $\frac{\lambda}{\theta} \to \infty$ (θ-fixed). Note that

$$P\left(X_\lambda \leq N-S-1\right) = P\left(\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \leq \frac{N-S-1-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}\right) \tag{4.8}$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \Rightarrow N(0,1) \tag{4.9}$$

and due to assumption *(ii)* of the lemma we get [1]

$$\sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \longrightarrow P\left(N(0,1) \leq \eta\right) = \Phi(\eta), \quad when \;\; \lambda \to \infty \tag{4.11}$$

where $N(0,1)$ is a standard normal random variable with distribution function Φ. It follows from (4.7)-(4.11) that

$$\xi_1 \approx \frac{e^{S(1-\rho)}}{\sqrt{2\pi\beta}}\rho^S\Phi(\eta) = \frac{e^{S((1-\rho)+\ln\rho)}}{\sqrt{2\pi\beta}}\Phi(\eta), \tag{4.12}$$

[1] Here we are using the following theorem (from [7])

Theorem 4.2.1 *Let $\zeta_n \Rightarrow \zeta$ and F_ζ - the distribution function of ζ is everywhere continuous. Let also $x_n \to x_\infty$ as $n \to \infty$, where $\{x_n\}$ is a sequence of scalars. Here $x_\infty \in [-\infty, \infty]$. Then*

$$F_{\zeta_n}(x_n) \longrightarrow F_\zeta(x_\infty) \tag{4.10}$$

where $\rho = \frac{p\lambda}{S\mu}$. Making use of the expansion

$$\ln \rho = \ln (1 - (1 - \rho)) = -(1 - \rho) - \frac{(1 - \rho)^2}{2} + o(1 - \rho)^2 \quad (\rho \to 1), \tag{4.13}$$

one obtains

$$\xi_1 \approx \frac{e^{S((1-\rho)-(1-\rho)-\frac{(1-\rho)^2}{2})}}{\sqrt{2\pi}\beta}\Phi(\eta) = \frac{e^{-\frac{S(1-\rho)^2}{2}}}{\sqrt{2\pi}\beta}\Phi(\eta). \tag{4.14}$$

Recall that $\sqrt{S}(1 - \rho) \to \beta$, then $S(1 - \rho)^2 \to \beta^2$, when $\lambda \to \infty$. This implies

$$\lim_{\lambda\to\infty} \xi_1 = \frac{\varphi(\beta)\Phi(\eta)}{\beta}, \tag{4.15}$$

where $\varphi(\cdot)$ is the standard normal density function, and $\Phi(\cdot)$ is the standard normal distribution function. This proves Lemma 4.2.1. □

Lemma 4.2.2 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then

$$\lim_{\lambda\to\infty} \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\frac{p\lambda}{S\mu}}\frac{p\lambda}{S\mu}^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\frac{p\lambda}{S\mu}}\right)^i = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\frac{\eta_1^2}{2}\Phi(\eta_1)}{\beta},$$

where $\eta_1 = \eta - \sqrt{\frac{\mu}{p\theta}}\beta$.

Proof. Again, for convenience, let us denote by ξ_2 the expression

$$\xi_2 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\rho^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i, \tag{4.16}$$

where $\rho = \frac{p\lambda}{S\mu}$.

Let us consider the asymptotic behavior of ξ_2 under the assumptions of Lemma 4.2.2.

$$\xi_2 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\rho^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i. \tag{4.17}$$

Again, by applying the Stirling's formula and using that $\rho = \frac{p\lambda}{S\mu} \to 1$, as $\lambda \to \infty$ and $S \to \infty$, one obtains,

$$\xi_2 \approx \frac{e^{S-\lambda\frac{p}{\mu}+\frac{\lambda(1-\rho)}{\theta\rho}}}{\sqrt{2S\pi}}\rho^N\frac{\sqrt{S}}{\beta}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i e^{-\frac{\lambda}{\theta\rho}} \tag{4.18}$$

The last sum can be rewritten as $P\left(Y_\lambda \leq N-S-1\right)$ where $Y_\lambda \sim Pois(\frac{\lambda}{\theta\rho})$, and $E[Y_\lambda] = \frac{\lambda}{\theta\rho}$, $Var[Y_\lambda] = \frac{\lambda}{\theta\rho}$. Note that

$$P\left(Y_\lambda \leq N-S-1\right) = P\left(\frac{X_\lambda - \frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}} \leq \frac{N-S-1-\frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}}\right) \tag{4.19}$$

Let us find a bound of the following fraction

$$\begin{aligned}\lim_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}} &= \lim_{\lambda\to\infty}\frac{\frac{\lambda}{\theta}-\frac{\lambda}{\theta\rho}+\eta\sqrt{\frac{\lambda}{\theta}}}{\sqrt{\frac{\lambda}{\theta\rho}}} = \eta + \lim_{\lambda\to\infty}\frac{\sqrt{\lambda}(\rho-1)}{\sqrt{\theta\rho}} \\ &= \eta - \lim_{\lambda\to\infty}\sqrt{\frac{S\mu}{p\theta}}(1-\rho) = \eta - \sqrt{\frac{\mu}{p\theta}}\beta.\end{aligned} \tag{4.20}$$

We have used the asymptotic relation

$$N - S \approx \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}, \quad as \quad \lambda \to \infty. \tag{4.21}$$

Denote

$$\eta_1 = \eta - \sqrt{\frac{\mu}{p\theta}}\beta$$

Taking into account equations (4.18) and (4.20), the Central Limit Theorem and Theorem 4.2.1 we have that

$$\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i e^{-\frac{\lambda}{\theta\rho}} \longrightarrow P\left(N(0,1) \leq \eta_1\right) = \Phi(\eta_1). \tag{4.22}$$

It follows from the assumption (ii) and (4.21) that

$$\begin{aligned}S - \frac{\lambda p}{\mu} + \frac{\lambda(1-\rho)}{\theta\rho} + N\ln\rho &\approx S(1-\rho) + \frac{\lambda}{\rho\theta}(1-\rho) - N(1-\rho) - \frac{N}{2}(1-\rho)^2 \\ &= (S-N)(1-\rho) + \frac{\lambda}{\rho\theta}(1-\rho) - \frac{N}{2}(1-\rho)^2 \\ &= -(N-S-\frac{\lambda}{\theta\rho})(1-\rho) - \frac{N}{2}(1-\rho)^2 \\ &\approx (\frac{\lambda}{\theta\rho} - \frac{N}{2})(1-\rho)^2 - \eta\sqrt{\frac{\lambda}{\theta}}(1-\rho), \qquad \rho \longrightarrow 1.\end{aligned}$$

Using (4.21) and the asymptotics

$$S \approx \frac{p\lambda}{\mu}, \quad N \approx \left(\frac{p}{\mu} + \frac{1}{\theta}\right)\lambda + \eta\sqrt{\frac{\lambda}{\theta}}, \qquad \lambda \longrightarrow \infty$$

one obtains

$$\begin{aligned}(\frac{\lambda}{\theta\rho} - \frac{N}{2})(1-\rho)^2 - \eta\sqrt{\frac{\lambda}{\theta}}(1-\rho) &\approx (\frac{1}{\theta\rho} - \frac{p}{2\mu} - \frac{1}{2\theta})\frac{\mu}{p}\beta^2 - \eta\beta\sqrt{\frac{\mu}{p\theta}} \\ &\approx \frac{1}{2}\beta^2(\frac{\mu}{p\theta} - 1) - \eta\beta\sqrt{\frac{\mu}{p\theta}} = -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}(\eta - \sqrt{\frac{\mu}{p\theta}}\beta)^2\end{aligned}$$

Therefore,

$$\lim_{\lambda\to\infty}\xi_2 = \frac{e^{-\frac{1}{2}(\eta^2+\beta^2)+\frac{1}{2}\eta_1^2}}{\sqrt{2\pi}\beta}\Phi(\eta_1) = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1), \tag{4.23}$$

and this proves Lemma 4.2.2. □

Lemma 4.2.3 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty}\sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty,$

where μ, p, θ are fixed. Then

$$\lim_{\lambda\to\infty}\sum_{i+j\le N-1, j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} = \int_{-\infty}^{\beta}\Phi\left(\eta + (\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t).$$

Proof. As in the proof of the previous lemmas, let us denote by γ the expression

$$\gamma = \sum_{i+j\le N-1, j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}; \tag{4.24}$$

and find its asymptotic behavior. For this purpose lower and upper estimates for γ in (4.24) will be found.

Let us consider a partition $\{S_j\}_{j=0}^{l}$ of the interval $[0, S]$.

$$S_j = S - j\delta, \qquad j = 0, 1, ..., l; \quad S_{l+1} = 0, \tag{4.25}$$

where $\delta = [\varepsilon\sqrt{\frac{p\lambda}{\mu}}]$, ε is an arbitrary non-negative real and l is a positive integer.

If λ and S tend to infinity and satisfy the assumption (ii), then l is less then $\frac{S}{\delta}$ for λ big enough and all the S_j belong to $[0, S]$, $j = 0, 1, ..., l$.

Emphasize that the length δ of every interval $[S_{j-1}, S_j]$ depends on λ.

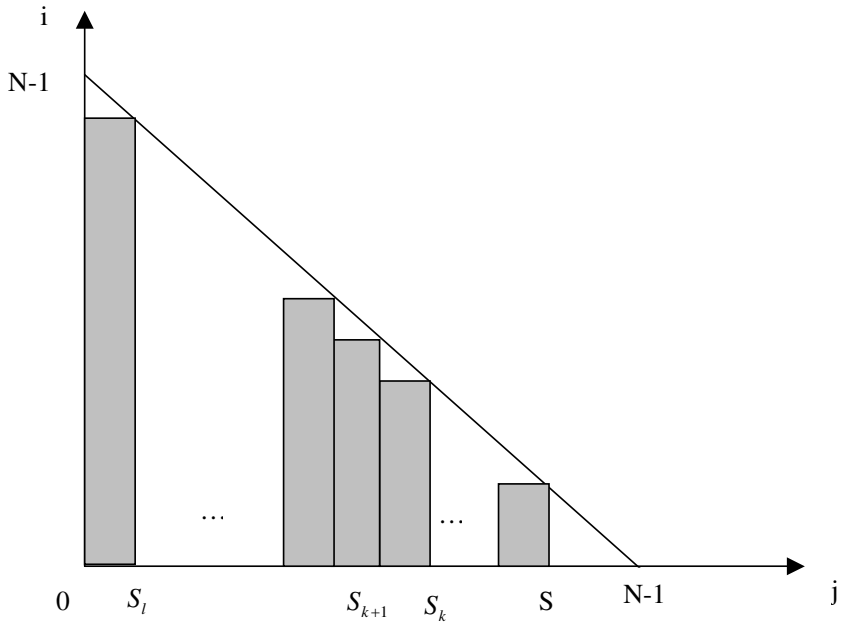

Figure 4.1: Area of the summation of the variable γ_1.

The variable γ is given by the formula (4.24), where summation is spread over the trapezoid on the Figure 4.1. Let us consider a lower estimate for γ given by the following sum with the summation over the shaded area on the Figure 4.1.

$$\begin{aligned}\gamma \geq \gamma_1 &= \sum_{k=0}^{l} \sum_{j=S_{k+1}}^{S_k-1} \frac{1}{j!} \left(\frac{p\lambda}{\mu}\right)^j e^{-\frac{p\lambda}{\mu}} \sum_{i=0}^{N-S_k} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \\ &= \sum_{k=0}^{l} P(S_{k+1} \leq Z_\lambda < S_k) P(X_\lambda \leq N - S_k),\end{aligned} \tag{4.26}$$

where

$$\begin{aligned} Z_\lambda &\sim Pois\left(\frac{p\lambda}{\mu}\right), \qquad E[Z_\lambda] = \frac{p\lambda}{\mu}, \qquad Var[Z_\lambda] = \frac{p\lambda}{\mu}; \\ X_\lambda &\sim Pois\left(\frac{\lambda}{\theta}\right), \qquad E[X_\lambda] = \frac{\lambda}{\theta}, \qquad Var[X_\lambda] = \frac{\lambda}{\theta}. \end{aligned} \tag{4.27}$$

Analogously to Lemmas 4.2.1 and 4.2.2, applying the Central Limit Theorem and

making use of the relations

$$\lim_{\lambda\to\infty}\frac{S_k-\frac{p\lambda}{\mu}}{\sqrt{\frac{p\lambda}{\mu}}}=\beta-k\varepsilon,\qquad k=0,1,...,l, \tag{4.28}$$

$$\lim_{\lambda\to\infty}\frac{N-S_k-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}},\qquad k=0,1,...,l, \tag{4.29}$$

one obtains

$$\lim_{\lambda\to\infty}P(S_{k+1}\le Z_\lambda<S_k)=\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon),\quad k=0,1,...l-1, \tag{4.30}$$

$$\lim_{\lambda\to\infty}P(0\le Z_\lambda<S_l)=\Phi(\beta-l\varepsilon), \tag{4.31}$$

$$\lim_{\lambda\to\infty}P(X_\lambda<N-S_k)=\Phi(\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}}),\quad k=0,1,...l. \tag{4.32}$$

It follows from (4.26) and (4.30), (4.31), (4.32) that

$$\begin{aligned}\lim_{\lambda\to\infty}\gamma\ge&\sum_{k=0}^{l}\Phi(\eta+k\varepsilon\sqrt{p\theta/\mu})[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]\\&+\Phi(\beta-l\varepsilon)\Phi(\eta+l\varepsilon\sqrt{p\theta/\mu}).\end{aligned} \tag{4.33}$$

It is easy to see that (4.33) is the lower Riemann-Stieltjes sum for the integral

$$-\int_0^\infty\Phi\left(\eta+s\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(\beta-s)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t), \tag{4.34}$$

corresponding to the partition $\{\beta-k\varepsilon\}_{k=0}^{l}$ of the semi axis $(-\infty,\beta)$.

Similarly let us take the upper estimate for γ as the following sum

$$\begin{aligned}\gamma\le\gamma_2&=\sum_{k=0}^{l}\sum_{j=S_{k+1}}^{S_k-1}\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j e^{-\frac{p\lambda}{\mu}}\sum_{i=0}^{N-S_{k+1}}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}}\\&=\sum_{k=0}^{l}P(S_{k+1}\le Z_\lambda<S_k)P(X_\lambda\le N-S_{k+1}),\end{aligned} \tag{4.35}$$

where the summation is widespread over the shaded area on the Figure 4.2.

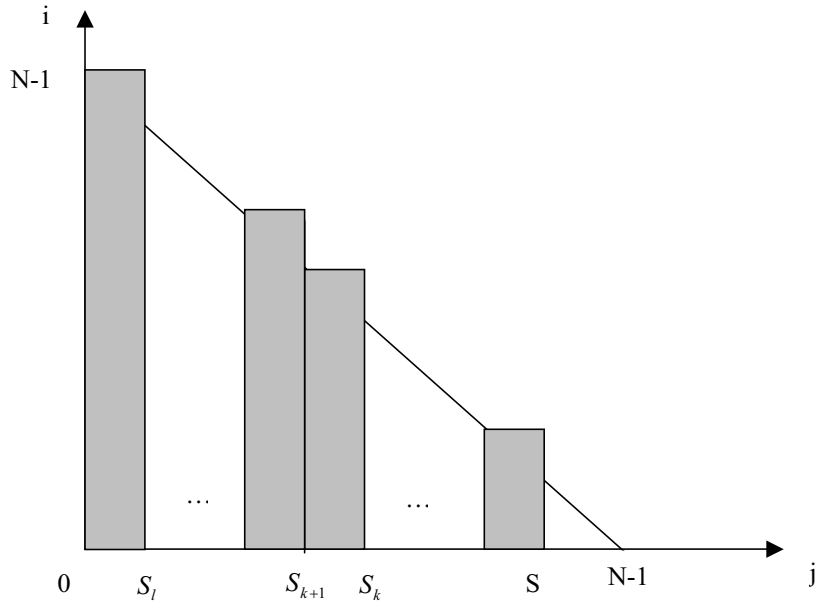

Figure 4.2: Area of the summation of the variable γ_2.

The above calculations applied to the sum (4.35) give the following asymptotic estimate for γ:

$$\lim_{\lambda\to\infty}\gamma \le \sum_{k=0}^{l}\Phi\left(\eta+(k+1)\varepsilon\sqrt{\frac{p\theta}{\mu}}\right)[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]+\Phi(\beta-l\varepsilon), \tag{4.36}$$

which is the upper Riemann-Stieltjes sum for the integral (4.34).

When $\varepsilon\to 0$ the estimates (4.33), (4.35) lead to the following equality

$$\lim_{\lambda\to\infty}\gamma = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t). \tag{4.37}$$

This proves Lemma 4.2.3. □

Lemma 4.2.4 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta, \quad -\infty<\eta<\infty;$

(ii) $\lim\limits_{\lambda\to\infty}\sqrt{S}(1-\frac{\lambda p}{\mu S})=0, \quad (\beta=0),$

where μ, p, θ are fixed, then

$$\lim_{\lambda\to\infty} e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}} \left(\frac{p\lambda}{\mu}\right)^j = \sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu}{p\theta}}\,(\eta\Phi(\eta)+\varphi(\eta)).$$

PROOF. First denote by ξ the expression

$$\xi = e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}} \left(\frac{p\lambda}{\mu}\right)^j. \tag{4.38}$$

Using Stirling's approximations and the assumptions (i) and (ii) we have

$$\begin{aligned}
\xi &= e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}} \left(\frac{p\lambda}{\mu}\right)^j \\
&= e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!} \sum_{j=0}^{N-S-i-1} \left(\frac{p\lambda}{\mu S}\right)^j \left(\frac{p\lambda}{\mu}\right)^S \\
&\approx e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \frac{e^S}{\sqrt{2\pi S}} \left(\frac{p\lambda}{\mu}\right)^S \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \sum_{j=0}^{N-S-i-1} \left(\frac{p\lambda}{\mu S}\right)^j \\
&= \frac{e^{S-\frac{\lambda p}{\mu}}}{\sqrt{2\pi S}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} \cdot \frac{1-\rho^{N-S-i}}{1-\rho},
\end{aligned} \tag{4.39}$$

where $\rho = \frac{p\lambda}{S\mu}$. Under lemma's condition (ii) $\beta = 0$ and this happens when $\rho = 1$ or $\rho \to 1$. When $\rho \to 1$, $\rho \neq 1$ we use well known approximations

$$\frac{1-\rho^k}{1-\rho} \approx k. \tag{4.40}$$

This implies that in this case

$$\xi \approx \frac{e^{S-\frac{\lambda p}{\mu}}}{\sqrt{2\pi S}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} (N-S-i),$$

When $\rho = 1$ the sum $\sum_{j=0}^{N-S-i-1} \rho^j$ in (4.39) is equal to $N-S-i$ and this leads to the same expression for ξ.

Simple calculations show that

$$\begin{aligned}
\xi &\approx \frac{1}{\sqrt{2\pi S}} \left(\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} (N-S) - \sum_{i=0}^{N-S-1} \frac{i e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} \right) \\
&= \frac{1}{\sqrt{2\pi S}} \left((N-S) \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} - \frac{\lambda}{\theta} \sum_{i=0}^{N-S-2} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} \right) \\
&\approx \frac{1}{\sqrt{2\pi S}} \left((N-S-\frac{\lambda}{\theta}) \sum_{i=0}^{N-S-1} e^{-\frac{\lambda}{\theta}} \frac{(\frac{\lambda}{\theta})^i}{i!} + e^{-\frac{\lambda}{\theta}} \frac{(\frac{\lambda}{\theta})^{N-S}}{(N-S-1)!} \right).
\end{aligned} \tag{4.41}$$

Due to the equation (4.11) the first term in (4.39) can be rewritten as follows

$$(N-S-\frac{\lambda}{\theta})\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^{i}e^{-\frac{\lambda}{\theta}}\approx\eta\sqrt{\frac{\lambda}{\theta}}\Phi(\eta),\qquad(\lambda\to\infty).\tag{4.42}$$

The Stirling's formula gives the following asymptotics for the second term in (4.41)

$$\begin{aligned}\frac{(\frac{\lambda}{\theta})^{N-S}e^{-\frac{\lambda}{\theta}}}{(N-S-1)!} &= \frac{N-S}{(N-S)!}\frac{\left(\frac{\lambda}{\theta}\right)^{N-S}e^{-\frac{\lambda}{\theta}}}{}\\ &\approx \sqrt{\frac{N-S}{2\pi}}e^{N-S-\frac{\lambda}{\theta}}\left(\frac{\lambda}{\theta(N-S)}\right)^{N-S}\\ &\approx \sqrt{\frac{N-S}{2\pi}}e^{(N-S)[1-\frac{\lambda}{\theta(N-S)}+\ln\frac{\lambda}{\theta(N-S)}]}\\ &\approx \sqrt{\frac{N-S}{2\pi}}e^{-\frac{\eta^2}{2}}\approx\sqrt{\frac{\lambda}{\theta}}\varphi(\eta).\end{aligned}\tag{4.43}$$

We have used in (9.1) the assumption (i) and the following approximations for the exponent

$$\begin{aligned}&(N-S)[1-\frac{\lambda}{\theta(N-S)}+\ln\frac{\lambda}{\theta(N-S)}]\\ &\quad\approx(N-S)[1-\frac{\lambda}{\theta(N-S)}-\left(1-\frac{\lambda}{\theta(N-S)}\right)-\frac{1}{2}\left(1-\frac{\lambda}{\theta(N-S)}\right)]\\ &\quad\approx-\frac{(N-S)}{2}\left(1-\frac{\lambda}{\theta(N-S)}\right)^{2}\approx-\frac{\lambda}{2\theta}\left(\frac{N-S-\frac{\lambda}{\theta}}{\theta(N-S)}\right)^{2}\\ &\quad\approx-\frac{\lambda}{2\theta}\left(\frac{\eta\sqrt{\frac{\lambda}{\theta}}}{\frac{\lambda}{\theta}}\right)^{2}\longrightarrow-\frac{\eta^2}{2}.\end{aligned}$$

Using the approximation

$$\frac{\lambda}{\theta S}\approx\frac{\frac{\mu}{p}(S-\beta\sqrt{S})}{\theta S}\approx\frac{\mu}{p\theta}.\tag{4.44}$$

and (4.42), (9.1) we obtain

$$\xi\approx\frac{1}{\sqrt{2\pi S}}\sqrt{\frac{\lambda}{\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right)\approx\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right).\tag{4.45}$$

This proves the Lemma 4.2.4. □

4.3 Approximation of $P(W > 0)$

First, assume that $\beta \neq 0$. In this case we can prove the following approximation of $P(W > 0)$.

Theorem 4.3.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then the probability $P(W > 0)$ that a customer will wait after IVR process has the following asymptotic behavior:

$$\alpha = \lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\beta \int\limits_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right) d\Phi(t)}{\varphi(\beta)\Phi(\eta) - \varphi(\sqrt{\eta^2 + \beta^2})\exp\frac{\eta_1^2}{2}\Phi(\eta_1)} \right)^{-1},$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}$.

Proof. We would like to find an approximation for the probability that a customer will wait after service in IVR. For this purpose, look more carefully at the definition of this probability, which appeared in (3.12):

$$\begin{aligned}
P \quad (W > 0) &= \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \chi(i,j) = \\
&= \frac{\sum\limits_{i=0}^{N-S-1} \sum\limits_{j=S}^{N-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j}{\sum\limits_{i+j\leq N-1, i\geq 1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j + \sum\limits_{i=1}^{N-S-1} \sum\limits_{j=S}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \\
&= \left(1 + \frac{\sum\limits_{i+j\leq N-1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j}{\sum\limits_{i=0}^{N-S-1} \sum\limits_{j=S}^{N-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \right)^{-1} \\
&= \left(1 + \frac{A}{B}\right)^{-1}, \qquad (4.46)
\end{aligned}$$

where

$$A = \sum_{i+j\leq N-1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j; \tag{4.47}$$

$$B = \sum_{i=0}^{N-S-1}\sum_{j=S}^{N-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j. \tag{4.48}$$

The iterated sum in (4.48) can be simplified. By changing indices $l = j - S$ we have:

$$\begin{aligned} B &= \sum_{i=0}^{N-S-1}\sum_{j=S}^{N-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j \\ &= \sum_{i=0}^{N-S-1}\sum_{l=0}^{N-S-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{l}}\left(\frac{p\lambda}{\mu}\right)^{l+S} \\ &= \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \sum_{l=0}^{N-S-i-1}\left(\frac{p\lambda}{S\mu}\right)^l \end{aligned} \tag{4.49}$$

Under the assumption of Theorem 4.3.1 one has $\beta \neq 0$. Therefore, the right hand side of (4.49) can be rewritten as

$$\begin{aligned} B &= \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1-\rho^{N-S-i}}{1-\rho} \\ &= \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\left[\sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i - \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \rho^{N-S-i}\right] \\ &= \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i - \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{\rho^{N-S}}{1-\rho} \sum_{i=0}^{N-S-1} \frac{\rho^{-i}}{i!}\left(\frac{\lambda}{\theta}\right)^i \\ &= B_1 - B_2, \end{aligned} \tag{4.50}$$

where $\rho = \frac{p\lambda}{S\mu}$.

Multiplying A, B_1 and B_2 by $e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}$ we have

$$P(W > 0) = \left(1 + \frac{\gamma}{\xi_1 - \xi_2}\right)^{-1}, \tag{4.51}$$

where

$$\xi_1 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i; \tag{4.52}$$

$$\xi_2 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\rho^{N-S} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i; \tag{4.53}$$

$$\gamma = \sum_{i+j\leq N-1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}. \tag{4.54}$$

Note, that ξ_1, ξ_2 and γ are the same as in (4.6), (4.16) and (4.24), respectively. So, it follows from Lemma 4.2.1, Lemma 4.2.2 and Lemma 4.2.3, that under the assumption of Theorem 4.3.1

$$\lim_{\lambda\to\infty} \xi_1 = \frac{\varphi(\beta)\Phi(\eta)}{\beta}; \tag{4.55}$$

$$\lim_{\lambda\to\infty} \xi_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\frac{\eta_1^2}{2}\Phi(\eta_1)}{\beta}, \tag{4.56}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}$;

$$\lim_{\lambda\to\infty} \gamma = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right) d\Phi(t). \tag{4.57}$$

Thus, equation (4.51) and equations (4.55)-(4.57) prove Theorem 4.3.1.□

In the case when $\beta = 0$ we can find an approximation for $P(W > 0)$ using the following theorem.

Theorem 4.3.2 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = 0, \quad (\beta = 0),$

where μ, p, θ are fixed. Then the probability that a customer will wait after the IVR process has the following asymptotic behavior:

$$\alpha = \lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{p\theta}{\mu}}\right) d\Phi(t)}{\sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta) + \varphi(\eta)\right)}\right)^{-1}.$$

Proof. In the case when $\beta = 0$, the probability that a customer will wait after the IVR process has the following form

$$P(Wait > 0) = \left(1 + \frac{A}{B}\right)^{-1},$$

where A and B are defined in (4.47) and (4.48), respectively. Multiplying A and B by $e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}$ we have

$$P(Wait > 0) = \left(1 + \frac{\gamma}{\xi}\right)^{-1}, \tag{4.58}$$

where $\gamma = Ae^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}$ and $\xi = Be^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}$. By Lemma 4.2.3, when $\beta = 0$

$$\lim_{\lambda\to\infty} \gamma = \int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{p\theta}{\mu}}\right) d\Phi(t). \tag{4.59}$$

Now, due to Lemma 4.2.4 we can say, that

$$\lim_{\lambda\to\infty} \xi = \sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta) + \varphi(\eta)\right). \tag{4.60}$$

Combining (4.59) and (4.60), we proved Theorem 4.3.2. □

4.4 Upper and lower bounds for the approximation of $P(W > 0)$

Let us find a cruder approximation for $P(W > 0)$ under the assumptions of Theorem 4.3.1. As we showed in the proof of Theorem 4.3.1, we can represent $P(W > 0)$ in the following way:

$$P(W > 0) = \left(1 + \frac{\gamma}{\xi_1 - \xi_2}\right)^{-1}.$$

Now, let us find upper and lower bounds for the approximation of γ. The picture below shows domains that yield these bounds.

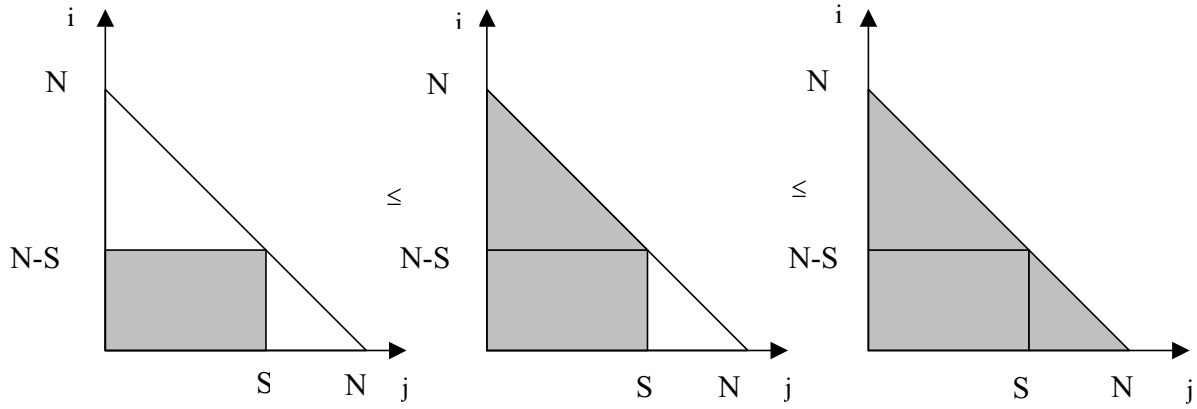

Figure 4.3: Graphical comparison of γ_1, γ and γ_2 areas simultaneously.

Instead of a summation over the trapezoid we will examine the summation over the rectangle (the upper bound of approximation of γ) and over the parallelogram (the lower bound of approximation of γ).

It is easy to see that:

$$\gamma_1 \leq \gamma \leq \gamma_2; \tag{4.61}$$

where

$$\gamma_1 = \sum_{i=0}^{N-S-1} \sum_{j=0}^{S} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})},$$

and

$$\gamma_2 = \sum_{k=0}^{S} \sum_{j=0}^{k} \frac{1}{(k-j)!} \left(\frac{\lambda}{\theta}\right)^{k-j} \frac{1}{j!} \left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}.$$

Separation of variables i and j in γ_1 leads to the equality

$$\gamma_1 = \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \sum_{j=0}^{S} \frac{1}{j!} \left(\frac{p\lambda}{\mu}\right)^j e^{-\frac{p\lambda}{\mu}} = P(X_\lambda \leq S) P(Y_\lambda \leq N-S-1),$$

where

$$X_\lambda \sim Pois\left(\frac{p\lambda}{\mu}\right) \qquad E[X_\lambda] = \frac{p\lambda}{\mu} \qquad Var[X_\lambda] = \frac{p\lambda}{\mu},$$
$$Y_\lambda \sim Pois\left(\frac{\lambda}{\theta}\right) \qquad E[Y_\lambda] = \frac{\lambda}{\theta} \qquad Var[Y_\lambda] = \frac{\lambda}{\theta}.$$

Applying the Central Limit theorem, Theorem 4.2.1 and using assumptions (i) and (ii) of Theorem 4.3.1, as in the case of ξ_1, we can show that

$$\lim_{\lambda\to\infty} P\left(Y_\lambda \le N-S-1\right) = \Phi(\eta) \quad \textit{and} \quad \lim_{\lambda\to\infty} P\left(X_\lambda \le S\right) = \Phi(\beta). \tag{4.62}$$

Thus we obtain that

$$\lim_{\lambda\to\infty} \gamma_1 = \Phi(\eta)\Phi(\beta). \tag{4.63}$$

The estimate γ_2 can be rewritten as

$$\gamma_2 = \sum_{i=0}^{N}\sum_{j=0}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} = \sum_{k=0}^{N} \frac{1}{k!}\left(\frac{\lambda}{\theta}+\frac{p\lambda}{\mu}\right)^k e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} = P(Z_\lambda \le N), \tag{4.64}$$

where

$$Z_\lambda \sim Pois\left(\frac{\lambda}{\theta}+\frac{p\lambda}{\mu}\right) \qquad EZ_\lambda = \frac{\lambda}{\theta}+\frac{p\lambda}{\mu} \qquad VarZ_\lambda = \frac{\lambda}{\theta}+\frac{p\lambda}{\mu}.$$

Again, by the Central Limit Theorem, the assumptions (i), (ii) and Theorem 4.2.1 one obtains

$$\begin{aligned} P(Z_\lambda \le N) &= \Phi\left(\frac{N-\frac{\lambda}{\theta}-\frac{p\lambda}{\mu}}{\sqrt{\frac{\lambda}{\theta}+\frac{p\lambda}{\mu}}}\right) \approx \Phi\left(\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}+\frac{p\lambda}{\mu}}} + \frac{S-\frac{p\lambda}{\mu}}{\sqrt{\frac{\lambda}{\theta}+\frac{p\lambda}{\mu}}}\right) = \\ &= \Phi\left(\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}\cdot\frac{1}{\sqrt{1+\frac{p\theta}{\mu}}} + \frac{S-\frac{p\lambda}{\mu}}{\sqrt{\frac{p\lambda}{\mu}}}\cdot\frac{1}{\sqrt{1+\frac{p\theta}{\mu}}}\right) \approx \\ &\approx \Phi\left(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}} + \frac{\beta}{\sqrt{1+\frac{p\theta}{\mu}}}\right). \end{aligned} \tag{4.65}$$

Consequently, we can formulate the following Remark.

Remark 4.4.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} = \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then the probability that a customer will wait after IVR process has the following estimates for its asymptotic behavior:

$$\left(1+\frac{A_1}{B}\right)^{-1} \leq \lim_{N\to\infty} P_N(W>0) \leq \left(1+\frac{A_2}{B}\right)^{-1},$$

where

$$A_1 = \Phi\left(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}} + \frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\beta, \quad A_2 = \Phi(\beta)\Phi(\eta)\beta,$$

$$B = \varphi(\beta)\Phi(\eta) - \varphi(\sqrt{\eta^2+\beta^2})\exp(\frac{\eta_1^2}{2})\Phi(\eta_1), \quad \eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}.$$

4.5 The boundary cases for $P(W>0)$

Now let us consider the boundary cases of our approximations for $P(W>0)$. For this purpose we distinguish the following cases:

a) $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta}) \longrightarrow \eta$, $(\eta \in (-\infty,+\infty))$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu}) \longrightarrow +\infty$;

b) $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta}) \longrightarrow \eta$, $(\eta \in (-\infty,+\infty))$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu}) \longrightarrow -\infty$;

c) $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta}) \longrightarrow +\infty$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu}) \longrightarrow \beta$, $(\beta \in (-\infty,+\infty))$;

d) $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta}) \longrightarrow -\infty$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu}) \longrightarrow \beta$, $(\beta \in (-\infty,+\infty))$.

In each case we will find a limit of approximation of the probability $P(W>0)$ that the customer will wait after IVR-service and before the agents' service.

Case a. This is the case when the number of agents is bigger then the rate of the arrival of customers to the agents' pool. So it is reasonable to suppose that $P(W>0)$ will tend to zero.

To prove this supposition consider the behavior of γ, ξ_1 and ξ_2 from (4.15), (4.23) and (4.37), respectively, under the assumptions (a).

It is easy to see that

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\xi_1 = \lim_{\beta\to\infty}\frac{\varphi(\beta)}{\beta}\Phi(\eta) = 0,$$

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\xi_2 = \lim_{\beta\to\infty}\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{\eta_1}{2}}\Phi(\eta_1) = 0.$$

In this case γ tends to 1, since

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\gamma=\lim_{\beta\to\infty}\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t)=1.$$

So,

$$\lim_{\beta\to\infty}\alpha=\lim_{\beta\to\infty}\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1}=0,$$

as we supposed.

Case b. In this case the number of customers is noticeably bigger than the number of agents. Thus, we can guess that almost all of the customers have to wait before receiving the agents' service. Therefore, the bound of $P(W>0)$ will tend to 1.

Indeed,

$$\lim_{\beta\to-\infty}(\xi_1-\xi_2)=\lim_{\beta\to-\infty}\left(\frac{\varphi(\beta)}{\beta}\Phi(\eta)-\frac{e^{-\frac{\beta^2}{2}(1-\frac{\mu}{p\theta})-\beta\eta\sqrt{\frac{\mu}{p\theta}}}}{\sqrt{2\pi}\beta}\Phi(\eta_1)\right)$$
$$=\begin{cases}0, & \text{if } \frac{\mu}{p\theta}<1;\\ \infty, & \text{if } \frac{\mu}{p\theta}\geq 1.\end{cases}$$

The variable γ is infinitesimal when β tends to $-\infty$

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\gamma=\lim_{\beta\to-\infty}\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t)=0.$$

Let us consider two cases. First, when

$$\frac{\mu}{p\theta}\geq 1,$$

then, clearly,

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1}=1.$$

In the second case, when

$$\frac{\mu}{p\theta}<1,$$

γ, under the previous section, has the following lower bound

$$\lim_{\lambda\to\infty}\gamma\leq\Phi(\eta)\Phi(\beta).$$

Using the well - known approximation for $\Phi(x)$, when $x \to -\infty$, we have

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\gamma \le \lim_{\beta\to-\infty} -\frac{\varphi(\beta)\Phi(\eta)}{\beta},$$

In that way

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\frac{\gamma}{\xi_1-\xi_2} \le \lim_{\beta\to-\infty}\frac{\Phi(\eta)\varphi(\beta)}{\beta\left(\frac{\varphi(\beta)}{\beta}\Phi(\eta)-\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}(\eta-\beta\sqrt{\frac{\mu}{p\theta}})^2}\Phi(\eta-\beta\sqrt{\frac{\mu}{p\theta}})\right)}$$

$$= \lim_{\beta\to-\infty}\frac{1}{1-e^{\frac{1}{2}(\eta-\beta\sqrt{\frac{\mu}{p\theta}})^2}/\Phi(\eta)} = 0$$

This implies

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1} = \begin{cases}1, & \frac{\mu}{p\theta}\ge 1,\\ 1, & \frac{\mu}{p\theta}<1.\end{cases}$$

So, the probability to wait for the customers which finished their IVR service and want to continue in the agents' service goes to 1, as we supposed.

Case c. In this case $\eta \to \infty$. This means there are infinite number of trunk lines in a call center. Because of this, we can guess that the number of customers in the IVR is not influenced by the state of the agents' pool. So, let us consider only the agents' pool. As we said, there is an infinite number of trunk lines, meaning that almost all the customers which seek to be served by an agent can enter the system. The customers come with the rate $p\lambda$. The agents' service time is exponential with rate μ. Our intuition is thus as follows:

• when $\beta \le 0$ the model is not in a steady state and almost all the customers are waiting, hence we can guess that $P(W > 0)$ tends to 1.

• when $\beta > 0$ we cannot give such a simple answer, but we guess that the customer's arrival process to the agents' pool can be modelled by the Poisson process, and then the system can be modelled by the M/M/S queue model, which was analyzed by Halfin and Whitt in [17]. So in this case we guess that our approximation will tend to Halfin and Whitt's approximation from [17].

Now, let us look mathematically at what happens with $P(W > 0)$ when $\eta \to \infty$. Since

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\xi_1 = \frac{\varphi(\beta)}{\beta},$$

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\xi_2 = \lim_{\eta\to\infty}\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{\eta_1^2}{2}}\Phi(\eta_1) = \begin{cases}0, & \beta>0,\\ -\infty, & \beta<0,\end{cases}$$

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\gamma = \lim_{\eta\to\infty}\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t) = \Phi(\beta).$$

it follows that

$$\lim_{\eta\to\infty}\alpha=\lim_{\eta\to\infty}\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1}=\begin{cases}\left(1+\frac{\Phi(\beta)\beta}{\varphi(\beta)}\right)^{-1}, & \beta>0,\\ 1, & \beta<0,\end{cases}\tag{4.66}$$

Note, that the approximation for probability to wait when $\beta > 0$ coincides with the approximation for probability to wait in the M/M/S system, which was founded by Halfin and Whitt in [17].

Case d. This is practically an impossible case for the call center with IVR, because in this case the number of agents is close to the number of the trunk lines. It is obvious that almost nobody waits in the queue, so the probability to wait tends to zero.

Actually, we have the following approximation for $\xi_1 - \xi_2$ when $\eta \to -\infty$

$$\begin{aligned}\xi_1-\xi_2&\approx-\frac{\varphi(\beta)\varphi(\eta)}{\beta\eta}+\frac{e^{-\frac{\eta^2+\beta^2}{2}}}{\sqrt{2\pi}\beta}\cdot\frac{\varphi(\eta_1)}{\eta_1}\cdot e^{\eta_1^2/2}\\&\approx-\frac{\varphi(\beta)\varphi(\eta)}{\beta\eta}+\frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\beta(\eta-\beta\sqrt{\mu/p\theta})}\\&\approx\frac{\varphi(\beta)\varphi(\eta)}{\eta^2}\sqrt{\frac{\mu}{p\theta}}.\end{aligned}\tag{4.67}$$

Using the lower estimate for γ from (4.63) and the asymptotic formulas for $\Phi(\eta)$ when $\eta \to -\infty$ we obtain

$$\begin{aligned}\lim_{\eta\to-\infty}\lim_{\lambda\to\infty}\frac{\gamma}{\xi_1-\xi_2}&\geq\lim_{\eta\to-\infty}\lim_{\lambda\to\infty}\frac{\gamma_1}{\xi_1-\xi_2}\\&=\lim_{\eta\to-\infty}\frac{\Phi(\eta)\Phi(\beta)}{\frac{\varphi(\beta)\varphi(\eta)}{\eta^2}\sqrt{\frac{\mu}{p\theta}}}\\&=\lim_{\eta\to-\infty}\frac{\Phi(\beta)}{\varphi(\beta)\sqrt{\frac{\mu}{p\theta}}}(-\eta)=\infty.\end{aligned}\tag{4.68}$$

This implies

$$\lim_{\lambda\to\infty}\lim_{\eta\to-\infty}\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1}=0.\tag{4.69}$$

So, just as expected, the probability to wait in this case tends to zero.

Analyzing the approximation of the probability to wait in the case when $\beta = 0$, we can say that

$$P(W>0)\longrightarrow\begin{cases}1, & \text{when } \eta\to+\infty\\ 0, & \text{when } \eta\to-\infty.\end{cases}$$

4.6 Approximation of $\sqrt{S}P(block)$

Let us find an approximation for $\sqrt{S}P(block)$. As in the previous section, we distinguish two cases

1) $\beta \neq 0$;

2) $\beta = 0$.

In the first case this approximation is given by the following theorem.

Theorem 4.6.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then the probability of blocking has the following asymptotic behavior:

$$\lim_{S\to\infty} \sqrt{S}P(block) = \frac{\nu\varphi(\nu_1)\,\Phi(\nu_2) + \varphi(\sqrt{\eta^2+\beta^2})\exp^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int\limits_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right) d\Phi(t) + \frac{\varphi(\beta)\Phi(\eta)}{\beta} - \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}\exp^{\frac{\eta_1^2}{2}}\Phi(\eta_1)},$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}, \quad \nu_1 = \frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}, \quad \nu_2 = \frac{\beta\sqrt{\frac{\mu}{p\theta}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}, \quad \nu = \frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}.$

Proof. The probability that an arriving call finds all trunk lines busy is given by the following formula:

$$\begin{aligned} P(block) &= \sum_{j=0}^{N} \pi(N-j, j) \\ &= \pi_0 \left(\sum_{j=0}^{S} \frac{1}{(N-j)!}\frac{1}{j!}\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{p\lambda}{\mu}\right)^{j} \right. \\ &+ \left. \sum_{j=S+1}^{N} \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^{j}\frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j} \right) \\ &= \frac{\widetilde{C}_1 + \widetilde{C}_2}{\widetilde{A} + \widetilde{B}_1 - \widetilde{B}_2} \cdot \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}} = \frac{\delta_1 + \delta_2}{\widetilde{\gamma} + \widetilde{\xi}_1 - \widetilde{\xi}_2}, \end{aligned} \tag{4.70}$$

where

$$\widetilde{A} = \sum_{i+j\leq N, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j; \tag{4.71}$$

$$\widetilde{B}_1 = \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i; \tag{4.72}$$

$$\widetilde{B}_2 = \frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S \frac{1}{1-\rho}\rho^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i; \tag{4.73}$$

$$\widetilde{C}_1 = \sum_{j=0}^{S}\frac{1}{(N-j)!}\frac{1}{j!}\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{p\lambda}{\mu}\right)^j; \tag{4.74}$$

$$\widetilde{C}_2 = \sum_{j=S+1}^{N}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j \frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j}. \tag{4.75}$$

So,

$$\widetilde{\gamma} = e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{i+j\leq N, j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j; \tag{4.76}$$

$$\widetilde{\xi}_1 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S\frac{1}{1-\rho}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i; \tag{4.77}$$

$$\widetilde{\xi}_2 = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{p\lambda}{\mu}\right)^S\frac{1}{1-\rho}\rho^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i; \tag{4.78}$$

$$\delta_1 = e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{j=0}^{S}\frac{1}{(N-j)!}\frac{1}{j!}\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{p\lambda}{\mu}\right)^j; \tag{4.79}$$

$$\delta_2 = e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{j=S+1}^{N}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j\frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j}. \tag{4.80}$$

Using (4.15), (4.23) and (4.37), notice that

$$\lim_{\lambda\to\infty}\widetilde{\gamma} = \lim_{\lambda\to\infty}\gamma = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t); \tag{4.81}$$

$$\lim_{\lambda\to\infty}\widetilde{\xi}_1 = \lim_{\lambda\to\infty}\xi_1 = \frac{\varphi(\beta)}{\beta}\Phi(\eta); \tag{4.82}$$

$$\lim_{\lambda\to\infty}\widetilde{\xi}_2 = \lim_{\lambda\to\infty}\rho\xi_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}exp^{\frac{1}{2}{\eta_1}^2}\Phi(\eta_1); \tag{4.83}$$

where β, η and η_1 are the same as in the Theorem 4.6.1.

Now let us consider

$$\begin{aligned}\delta_1 &= \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{N!}\sum_{j=0}^{S}\frac{N!}{(N-j)!j!}\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{p\lambda}{\mu}\right)^{j}\\ &= \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{N!}\left(\lambda\left(\frac{1}{\theta}+\frac{p}{\mu}\right)\right)^{N}P(X_\lambda\le S)\\ &= P(Y_\lambda = N)P(X_\lambda\le S),\end{aligned}$$

where

$$X_\lambda \sim Bin\left(N, \frac{\frac{p}{\mu}}{\frac{1}{\theta}+\frac{p}{\mu}}\right);\quad EX_\lambda = \frac{Np}{\mu(\frac{1}{\theta}+\frac{p}{\theta})};\quad VarX_\lambda = \frac{\frac{N\mu}{p\theta}}{(1+\frac{\mu}{p\theta})^2};$$

$$Y_\lambda \sim Pois\left(\lambda\left(\frac{1}{\theta}+\frac{p}{\mu}\right)\right);\quad EY_\lambda = \lambda\left(\frac{1}{\theta}+\frac{p}{\mu}\right);\quad VarY_\lambda = \lambda\left(\frac{1}{\theta}+\frac{p}{\mu}\right).$$

By the Central Limit theorem and Theorem 4.2.1 one obtains

$$\begin{aligned}P(X_\lambda\le S) &= \Phi\left(\frac{S-\frac{Np}{\mu(\frac{1}{\theta}+\frac{p}{\theta})}}{\sqrt{\frac{\frac{N\mu}{p\theta}}{(1+\frac{\mu}{p\theta})^2}}}\right) = \Phi\left(\frac{(S(\frac{\mu}{p\theta}+1)-N)}{\sqrt{\frac{N\mu}{p\theta}}}\right) =\\ &= \Phi\left(\frac{-\sqrt{\frac{p\theta}{\mu}}\left(S(1+\frac{\mu}{p\theta})-\sqrt{S}(\frac{\beta\mu}{p\theta}-\eta\sqrt{\frac{\mu}{p\theta}})-S(1-\frac{\mu}{p\theta})\right)}{\sqrt{S(1+\frac{\mu}{p\theta})}}\right) =\\ &= \Phi\left(\frac{\beta\sqrt{\frac{\mu}{p\theta}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}\right);\end{aligned}\tag{4.84}$$

$$\begin{aligned}P(Y_\lambda = N) &= P\left(\frac{N-1-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}} < \frac{Y_\lambda-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}} \le \frac{N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}\right) =\\ &= P\left(\frac{N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}} - \frac{1}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}} < Z \le \frac{N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}\right)\\ &\approx \frac{1}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}\varphi\left(\frac{N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}\right)\\ &\approx \frac{1}{\sqrt{S}\sqrt{1+\frac{\mu}{p\theta}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right).\end{aligned}\tag{4.85}$$

In the last case we have used the equivalences (as λ tends ∞)

$$N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})\approx S+\frac{\lambda}{\theta}+\eta\sqrt{\frac{\lambda}{\theta}}-\frac{\lambda}{\theta}-\frac{p\lambda}{\mu}\approx S+\eta\sqrt{\frac{\lambda}{\theta}}-S+\beta\sqrt{S}\approx\eta\sqrt{\frac{S\mu}{p\theta}}+\beta\sqrt{S};$$

$$\lambda(\frac{1}{\theta}+\frac{p}{\mu})\approx\frac{S\mu}{p\theta}+S;$$

$$\frac{N-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}{\sqrt{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}\approx\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}=\frac{\eta+\beta\sqrt{\frac{p\theta}{\mu}}}{\sqrt{1+\frac{p\theta}{\mu}}}.$$

It follows from (4.84), (4.85) that

$$\lim_{\lambda\to\infty}\delta_1=\frac{1}{\sqrt{S}\sqrt{1+\frac{\mu}{p\theta}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{p\theta}{\mu}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\tag{4.86}$$

Let us find an approximation for δ_2.

$$\begin{aligned}\delta_2 &= e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{j=S+1}^{N}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j\frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j}=\\ &= \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!S^{-S}}\sum_{j=S+1}^{N}\frac{1}{(N-j)!}\left(\frac{p\lambda}{\mu S}\right)^j\left(\frac{\lambda}{\theta}\right)^{N-j}=\\ &= \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!S^{-S}}\left(\frac{p\lambda}{\mu S}\right)^N\sum_{i=0}^{N-S-1}\left(\frac{\lambda}{\theta\rho}\right)^i\frac{1}{i!}.\end{aligned}\tag{4.87}$$

Comparing (4.17) and (4.87) one obtains

$$\delta_2=\sqrt{S}\beta\xi_2,\quad\lambda\to\infty.$$

In the previous section (cf. (4.18) and (4.23) an approximation of ξ_2 was found, therefore

$$\lim_{\lambda\to\infty}\widetilde{\delta}_2=\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\sqrt{S}}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1),\tag{4.88}$$

where $\eta_1=\eta-\beta\sqrt{\frac{\mu}{p\theta}}$. This proves Theorem 4.6.1. □

Now let us consider what happens in the case when $\beta=0$. Approximation for $\widetilde{\gamma}$, $\widetilde{\xi}$ are the same as the approximation of γ and ξ (4.59), (4.45), in the proof of Theorem 2.2 and approximation for δ_1 are the same as in the proof of Theorem

3.1. So, we just need to find an approximation for δ_2. For this purpose one can use the formulas (4.86) and (4.88) from the proof of Theorem 4.6.1.

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}\delta_2 = \frac{1}{\sqrt{2\pi}}\Phi(\eta) \tag{4.89}$$

and

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}\delta_1 = \frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}\varphi(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}})\Phi(\frac{\eta}{\sqrt{1+\frac{\mu}{p\theta}}}) \tag{4.90}$$

Now we can formulate the following theorem.

Theorem 4.6.2 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty}\sqrt{S}(1-\frac{\lambda p}{\mu S}) = 0, \quad (\beta = 0),$

where μ, p, θ are fixed. Then the probability of blocking has the following asymptotic behavior:

$$\lim_{S\to\infty}\sqrt{S}P(block) = \frac{\nu\varphi(\nu_1)\,\Phi(\nu_2) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int\limits_{-\infty}^{0}\Phi\left(\eta - t\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t) + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu}{p\theta}}\,(\eta\Phi(\eta)+\varphi(\eta))},$$

where $\nu_1 = \frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}}, \quad \nu_2 = \frac{\eta}{\sqrt{1+\frac{\mu}{p\theta}}}, \quad \nu = \frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}.$

4.7 Upper and lower bounds for the approximation of $\sqrt{S}P(block)$

Calculations similar to those in Remark 4.2.1 yield upper and lower bounds for approximation of $\sqrt{S}P(block)$.

Remark 4.7.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} = \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \ \beta \neq 0,$

μ, p, θ are fixed. Then the probability of blocking has the the following estimates for its asymptotic behavior:

$$\frac{\nu\varphi\left(\nu_1\right)\Phi\left(\nu_2\right)+\beta B_2}{A_1+B_1-B_2} \leq \lim_{S\to\infty}\sqrt{S}P(block) \leq \frac{\nu\varphi\left(\nu_1\right)\Phi\left(\nu_2\right)+\beta B_2}{A_2+B_1-B_2},$$

where

$$A_1=\Phi(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}}+\frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}}), \quad A_2=\Phi(\beta)\Phi(\eta),$$

$$B_1=\frac{1}{\beta}\varphi(\beta)\Phi(\eta), \quad B_2=\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}\exp(\frac{\eta_1^2}{2})\Phi(\eta_1),$$

$$\eta_1=\eta-\beta\sqrt{\frac{\mu}{p\theta}}, \quad \nu_1=\frac{\eta+\beta\sqrt{\frac{p\theta}{\mu}}}{1+\frac{p\theta}{\mu}}, \quad \nu_2=\frac{\beta\sqrt{\frac{p\theta}{\mu}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}, \quad \nu=\frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}.$$

4.8 The boundary cases for $\sqrt{S}P(block)$

Now let us see what happens with $\sqrt{S}P(block)$ when either β or η is going to ∞. The cases are the same as in the analyzing of behavior of the approximation of $P(W>0)$.

Case a. $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta}) \to \eta$, $(\eta \in (-\infty,+\infty))$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu}) \to +\infty$.

In this case the number of agents is bigger than the number of customers which want to receive agents' service. So, it is possible to suppose that almost nobody waits before the agents' service and therefore there are many vacancies in the system, then $\sqrt{S}P(block)$ tends to zero.

As we saw when proving the Theorem 4.6.1 the probability of blocking can be written in the following form

$$P(block)=\frac{\delta_1+\delta_2}{\widetilde{\gamma}+\widetilde{\xi_1}-\widetilde{\xi_2}} \tag{4.91}$$

Approximations of $\widetilde{\xi_1}$, $\widetilde{\xi_2}$, γ, δ_1 and δ_2 are given in (4.15), (4.23), (4.37), (4.86) (4.88). It is easy to see that

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\widetilde{\xi_1}=\lim_{\beta\to\infty}\frac{\varphi(\beta)}{\beta}\Phi(\eta)=0,$$

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\widetilde{\xi_2}=\lim_{\beta\to\infty}\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)=0,$$

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sqrt{S}\delta_1=\lim_{\beta\to\infty}\frac{1}{\sqrt{S}\sqrt{1+\frac{\mu}{p\theta}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{p\theta}{\mu}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)=0,$$

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sqrt{S}\delta_2=\lim_{\beta\to\infty}\varphi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)=0,$$

and

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\widetilde{\gamma}=\lim_{\beta\to\infty}\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t)=\lim_{\beta\to\infty}E[\Phi(\eta+(\beta-t)\sqrt{p\theta/\mu})]=1,$$

So,

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sqrt{S}P(block)=0.$$

Case b. $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta})\to\eta,\ \ (\eta\in(-\infty,+\infty))$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu})\to-\infty$;

In this case the rate of customers is bigger than the number of agents and this implies that there are many people in the queue and all the system is busy. Thus, the probability to find the system busy multiplied by the square root from the number of agents tends to infinity. This heuristic analysis is approved by the following mathematical calculations.

When $\beta\longrightarrow-\infty$ then

$$\eta_1\longrightarrow\infty,\qquad\nu_1\longrightarrow-\infty,\qquad\nu_2\longrightarrow-\infty$$

As we saw in (4.61), (4.64) and (4.65)

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\widetilde{\gamma}=\lim_{\beta\to-\infty}\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t)\le\Phi(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}}+\frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}})$$

$$=\lim_{\beta\to-\infty}-\frac{\varphi(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}}+\frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}})}{\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}}+\frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}}}=\lim_{\beta\to-\infty}-\frac{\varphi(\nu_1)}{\nu_1}=0$$

We can see also that

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\delta_1=0,\qquad\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\delta_2=0,$$

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\widetilde{\xi}_1=0,\qquad\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\widetilde{\xi}_2=\infty.$$

So,

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\sqrt{S}P(block)=\frac{\delta_1+\beta\xi_2}{\widetilde{\gamma}+\widetilde{\xi}_1-\widetilde{\xi}_2}=\infty,$$

as we supposed.

Case c. $\sqrt{\frac{\theta}{\lambda}}(N - S - \frac{\lambda}{\theta}) \to +\infty$, and $\sqrt{S}(1 - \frac{p\lambda}{S\mu}) \to \beta$, $(\beta \in (-\infty, +\infty))$.

In this case we have an enormous number of places in the system. As we saw in Section 4.3 there are two different situations in the agents' pool. When $\beta > 0$ it is not a problem to enter into the system and therefore the probability to block tends to zero. In the case $\beta \leq 0$ it is not clearly understandable what happens with the number of customers in the system, and we can guess that the probability to find the system busy differs from 0. Indeed, in this case

$$\eta_1 \to \infty, \qquad \nu_2 \to -\infty, \qquad \lim_{\beta\to\infty}\lim_{\lambda\to\infty} \Phi(\nu_2) = \frac{\varphi(\nu_2)}{-\nu_2},$$

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty} \Phi(\eta_1) = 1, \qquad \lim_{\beta\to\infty}\lim_{\lambda\to\infty} \varphi(\nu_1) = 0,$$

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty} \sqrt{S}P(block) = \lim_{\eta\to\infty}\lim_{\lambda\to\infty} \frac{\nu\varphi(\nu_1)\Phi(\nu_2) + \beta\xi_2}{\gamma + \xi_1 - \xi_2} = \begin{cases} 0, & \beta > 0; \\ -\beta, & \beta < 0. \end{cases}$$

Case d. $\sqrt{\frac{\theta}{\lambda}}(N - S - \frac{\lambda}{\theta}) \to -\infty$, and $\sqrt{S}(1 - \frac{p\lambda}{S\mu}) \to \beta$, $(\beta \in (-\infty, +\infty))$.

In this case the number of places in the queue tends to zero, so it is difficult to come into the system and the probability to find the system busy, multiplied by $\sqrt{S}$ tends to infinity.

When $\eta \longrightarrow -\infty$ then

$$\eta_1 \longrightarrow -\infty, \qquad \nu_1 \longrightarrow -\infty, \qquad \nu_2 \longrightarrow \infty,$$

and

$$\lim_{\eta\to-\infty}\lim_{\lambda\to-\infty} \Phi(\eta_1) = \lim_{\eta\to\infty} -\frac{\varphi(\eta_1)}{\eta_1}, \quad \lim_{\eta\to\infty} \varphi(\nu_1) = 0, \quad \lim_{\eta\to\infty} \Phi(\nu_2) = 1.$$

Also we can see that

$$\lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \widetilde{\xi}_1\beta = \lim_{\eta\to-\infty} \frac{C}{\eta^2} e^{-\frac{\eta^2+\beta^2}{2}}, \qquad \lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \widetilde{\xi}_2\beta = \lim_{\eta\to-\infty} \frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\eta},$$

$$\lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \varphi(\nu_1)\Phi(\nu_2) = C_1 e^{-\frac{1}{2}\frac{\eta^2}{1+\frac{\mu}{p\theta}}},$$

$$\lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \widetilde{\gamma} \leq \lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \gamma_1 = \lim_{\eta\to-\infty} -\frac{\varphi(\nu_1)}{\nu_1},$$

where C and C_1 does not depend on η.

Consequently,

$$\lim_{\eta\to-\infty}\lim_{\lambda\to\infty} \sqrt{S}P(block) = \lim_{\eta\to-\infty} -\frac{\nu\varphi(\nu_1)\varphi(\nu_2)}{\frac{\varphi(\nu_1)}{\nu_1}} = \infty,$$

as we supposed.

Analyzing the approximation when $\beta = 0$ we can say that

$$\sqrt{S}P(block) \longrightarrow \begin{cases} 0, & \text{when } \eta \to +\infty \\ +\infty, & \text{when } \eta \longrightarrow -\infty. \end{cases}$$

4.9 Approximation of $\sqrt{S}E[W]$

Now let us find an approximation for the expected waiting time $E[W]$ before service, for customers that finished their IVR process and continue to receive service from the agents.

Theorem 4.9.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \ \beta \neq 0,$

where μ, p, θ are fixed. Then the expectation of the waiting time has the following asymptotic behavior:

$$\lim_{S\to\infty} \sqrt{S}E[W] = \frac{\frac{1}{\mu}\left(\frac{1}{\beta}\varphi(\beta)\Phi(\eta) + (\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}})\varphi(\sqrt{\eta^2+\beta^2})\exp(\frac{\eta_1^2}{2})\Phi(\eta_1)\right)}{\beta\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t) + \varphi(\beta)\Phi(\eta) - \varphi(\sqrt{\eta^2+\beta^2})\exp(\frac{\eta_1^2}{2})\Phi(\eta_1)},$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}$.

Proof. It follows from (3.11) that the expectation of the waiting time is given by

$$\begin{aligned} E[W] &= \frac{1}{\mu S}\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)(j-S+1) \\ &= \frac{1}{\mu S}\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)(j-S) + \frac{1}{\mu S}P(W>0) \\ &= C + D, \end{aligned}$$

where

$$
\begin{aligned}
C &= \frac{\frac{1}{\mu S}\sum\limits_{i=1}^{N-S}\sum\limits_{j=S}^{N-i}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j(j-S)}{\sum\limits_{i+j\le N,i\ge 1,j<S}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j+\sum\limits_{i=1}^{N-S}\sum\limits_{j=S}^{N-i}\frac{i}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \\
&= \frac{\frac{1}{\mu S}\sum\limits_{m=0}^{N-S-1}\sum\limits_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j(j-S)}{\sum\limits_{m+j\le N-1,j<S}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j+\sum\limits_{m=0}^{N-S-1}\sum\limits_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j} \\
&= \frac{1}{\mu S}\frac{G}{A+B}; \qquad (4.92)
\end{aligned}
$$

and

$$
D = \frac{1}{\mu S}\frac{B}{A+B}. \qquad (4.93)
$$

The numerator G in (4.92), can be rewritten as follow:

$$
\begin{aligned}
G &= \sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\sum_{j=S}^{N-m-1}\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j(j-S) = \\
&= \sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!S^{-S}}\sum_{l=0}^{N-S-m-1}\left(\frac{p\lambda}{\mu S}\right)^{l+S}l = \\
&= \sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S\sum_{l=0}^{N-S-m-1}\rho^l l = \\
&= \frac{1}{S!}\sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\left(\frac{p\lambda}{\mu}\right)^S\sum_{l=0}^{N-S-m-1}\rho^l l. \qquad (4.94)
\end{aligned}
$$

Using the formula

$$
\sum_{n=0}^{M}n\rho^n = \rho\left(\sum_{n=0}^{M}\rho^n\right)' = \rho\left(\frac{1-\rho^{M+1}}{1-\rho}\right)' = M\frac{\rho^{M+1}}{\rho-1}+\frac{1-\rho^M}{(1-\rho)^2}\rho, \qquad (4.95)
$$

one can rewrite G as a sum:

$$
G = G_1 + G_2,
$$

where

$$
G_1 = \sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S\frac{\rho^{N-S-m}}{\rho-1}(N-S-m-1); \qquad (4.96)
$$

and

$$G_2 = \sum_{m=0}^{N-S-1} \frac{1}{m!} \left(\frac{\lambda}{\theta}\right)^m \frac{1}{S!} \left(\frac{p\lambda}{\mu}\right)^S \frac{1-\rho^{N-S-m-1}}{(1-\rho)^2}\rho. \tag{4.97}$$

It follows from (4.96) that

$$\begin{aligned} G_1 = & \frac{(N-S-1)\rho^{N-S}}{S!\rho - 1} \left(\frac{p\lambda}{\mu}\right)^S \sum_{m=0}^{N-S-1} \frac{1}{m!} \left(\frac{\lambda}{\theta\rho}\right)^m \\ & - \frac{1}{S!}\frac{\rho^{N-S}}{\rho-1} \left(\frac{p\lambda}{\mu}\right)^S \sum_{m=0}^{N-S-1} \frac{m}{m!} \left(\frac{\lambda}{\theta\rho}\right)^m \\ & = -(N-S)e^{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\xi_2 + \frac{\lambda}{\theta\rho}e^{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\xi_2, \end{aligned} \tag{4.98}$$

where ξ_2 was defined in (4.15). Next, (4.6), (4.16) and (4.97) yield

$$\begin{aligned} G_2 = & \frac{1}{(1-\rho)^2}\frac{1}{S!} \left(\frac{p\lambda}{\mu}\right)^S \sum_{m=0}^{N-S-1} \frac{1}{m!} \left(\frac{\lambda}{\theta}\right)^m \\ & - \frac{1}{1-\rho}\frac{1}{S!}\frac{\rho^{N-S-1}}{1-\rho} \left(\frac{p\lambda}{\mu}\right)^S \sum_{m=0}^{N-S-1} \frac{1}{m!} \left(\frac{\lambda}{\theta\rho}\right)^m \\ & = \frac{e^{\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{1-\rho}\left(\xi_1 - \xi_2\right). \end{aligned} \tag{4.99}$$

Multiplying G by $e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}$ one obtains from (4.98), (4.99)

$$\begin{aligned} Ge^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} &= \frac{\lambda}{\theta\rho}\xi_2 - (N-S)\xi_2 + \frac{1}{1-\rho}(\xi_1-\xi_2) \approx \\ &\approx \xi_1\frac{\sqrt{S}}{\beta} + \xi_2\left(\frac{\mu}{p\theta}S - \frac{\mu}{p\theta}S + \beta\frac{\mu}{p\theta}\sqrt{S} - \eta\sqrt{\frac{\mu}{p\theta}}\sqrt{S} - \frac{1}{\beta}\sqrt{S}\right) = \\ &= \sqrt{S}\left(\xi_1\frac{1}{\beta} + \xi_2(\beta\frac{\mu}{p\theta} - \eta\sqrt{\frac{\mu}{p\theta}} - \frac{1}{\beta})\right). \end{aligned} \tag{4.100}$$

Since

$$E[W] = \frac{1}{\mu S}\frac{G}{A+B} + \frac{1}{\mu S}\frac{B}{A+B},$$

and using previous calculations for A and B, we may say that

$$\sqrt{S}E[W] \approx \frac{1}{\mu\sqrt{S}}\frac{G}{A+B} \approx \frac{\xi_1\frac{1}{\beta} + \xi_2(\beta\frac{\mu}{p\theta} - \eta\sqrt{\frac{\mu}{p\theta}} - \frac{1}{\beta})}{\mu(\gamma+\xi_1-\xi_2)}. \tag{4.101}$$

The approximations for γ, ξ_1 and ξ_2 and the formula (4.101) prove the statement of Theorem 4.1. □

In the case of $\beta = 0$ we can formulate the following

Theorem 4.9.2 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = 0, \quad (\beta = 0),$

where μ, p, θ are fixed. Then the expectation of the waiting time has the following asymptotic behavior:

$$\lim_{S\to\infty} \sqrt{S}E[W] = \frac{1}{2\mu} \cdot \frac{\eta^2 \frac{\mu}{p\theta}\Phi(\eta) + \eta\sqrt{\frac{\mu}{p\theta}}\left(1 + \sqrt{\frac{\mu}{p\theta}}\right)\varphi(\eta)}{\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta) + \varphi(\eta)\right) + \sqrt{2\pi}\int_{-\infty}^{0} \Phi(\eta - t\sqrt{\frac{p\theta}{\mu}})d\Phi(t)}.$$

Proof. First, let us calculate the last sum in (4.94). The condition $\beta = 0$ means that $\rho = 1$ or $\rho \to 1$. If $\rho = 1$ it is easy to see that

$$\sum_{l=0}^{N-S-m-1} l\rho^l = \frac{(N-S-m-1)(N-S-m)}{2}. \tag{4.102}$$

If $\rho \to 1$ by using the Teylor's formula and (4.13) we can say that

$$\begin{aligned} \rho^M = e^{M\ln\rho} &\approx 1 + M\ln\rho + \frac{M^2\ln^2\rho}{2} \\ &\approx 1 + M\left((\rho-1) - \frac{(\rho-1)^2}{2}\right) + \frac{M^2(\rho-1)^2}{2}. \end{aligned} \tag{4.103}$$

Then by using the relation (4.95) the sum in (4.94) we can rewrite in the form

$$\sum_{l=0}^{M} l\rho^l = M\frac{\rho^{M+1}-1}{\rho-1} + \frac{M}{\rho-1} - \frac{\rho^M - 1}{(\rho-1)^2}, \tag{4.104}$$

where $M = N - S - m - 1$. Taking into account (4.40) and (4.103), one obtains

$$\begin{aligned} \sum_{l=0}^{M} l\rho^l k &\approx M(M+1) + \frac{M}{\rho-1} - \frac{1 + M(\rho-1) - \frac{M(\rho-1)^2}{2} + \frac{M^2(\rho-1)^2}{2} - 1}{(\rho-1)^2} - \frac{\rho^M - 1}{\rho-1} \\ &= M^2 - \frac{M(M-1)}{2} = \frac{M(M+1)}{2}. \end{aligned} \tag{4.105}$$

Thus, (4.94) has the following form:

$$\begin{aligned} Ge^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} &= e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\frac{1}{S!S^{-S}}\sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{N-S-m-1}{2}(N-S-m) \\ &\approx \frac{e^{S(1-\rho)}}{\sqrt{2\pi S}}\frac{1}{2}\sum_{m=0}^{N-S-1}\frac{e^{-\frac{\lambda}{\theta}}}{m!}\left(\frac{\lambda}{\theta}\right)^m(N-S-m-1)(N-S-m) \\ &\approx \frac{N-S-1}{2\sqrt{2\pi S}}\sum_{m=0}^{N-S-1}\frac{e^{-\frac{\lambda}{\theta}}}{m!}\left(\frac{\lambda}{\theta}\right)^m(N-S-m) \\ &- \frac{1}{2\sqrt{2\pi S}}\sum_{m=0}^{N-S-1}\frac{me^{-\frac{\lambda}{\theta}}}{m!}\left(\frac{\lambda}{\theta}\right)^m(N-S-m) \qquad (4.106) \end{aligned}$$

In the notations used in the proof of Theorem 4.9.1 one can rewrite the last sum in the form

$$\frac{N-S-1}{2}\xi - \frac{\lambda}{2\theta}\left(\xi - \frac{1}{\sqrt{2\pi}}\frac{(\frac{\lambda}{\theta})^{N-S-1}e^{-\frac{\lambda}{\theta}}}{(N-S-1)!}\right)$$

It follows from (9.1), (4.45) and the assumption (ii) that

$$\begin{aligned} Ge^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} &\approx (N-S-\frac{\lambda}{\theta}-1)\frac{\xi}{2} + \frac{\left(\frac{\lambda}{\theta}\right)^{N-S}e^{-\frac{\lambda}{\theta}}}{2(N-S-1)!\sqrt{2\pi}} \qquad (4.107) \\ &\approx \eta\sqrt{\frac{\lambda}{\theta}}\frac{\xi}{2} + \frac{1}{2\sqrt{2\pi}}\sqrt{\frac{\lambda}{\theta}}\varphi(\eta) \qquad (4.108) \\ &\approx \frac{\sqrt{S}}{2\sqrt{2\pi}}\left[\eta^2\frac{\mu}{p\theta}\Phi(\eta) + \eta\sqrt{\frac{\mu}{p\theta}}\left(1+\sqrt{\frac{\mu}{p\theta}}\right)\varphi(\eta)\right] \qquad (4.109) \end{aligned}$$

Making use of (4.107), (4.59), (4.45) one obtains

$$\begin{aligned} \sqrt{S}E[W] &\approx \frac{1}{\mu S}\frac{G}{A+B} \\ &\approx \frac{\frac{1}{2\mu\sqrt{2\pi}}\left[\eta^2\frac{\mu}{p\theta}\Phi(\eta) + \eta\sqrt{\frac{\mu}{p\theta}}\left(1+\sqrt{\frac{\mu}{p\theta}}\right)\varphi(\eta)\right]}{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right) + \int_{-\infty}^{0}\Phi(\eta - t\sqrt{\frac{p\theta}{\mu}})d\Phi(t)} \\ &\approx \frac{1}{2\mu}\frac{\eta^2\frac{\mu}{p\theta}\Phi(\eta) + \eta\sqrt{\frac{\mu}{p\theta}}\left(1+\sqrt{\frac{\mu}{p\theta}}\right)\varphi(\eta)}{\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right) + \sqrt{2\pi}\int_{-\infty}^{0}\Phi(\eta - t\sqrt{\frac{p\theta}{\mu}})d\Phi(t)}. \end{aligned}$$

This proves the Theorem 4.9.1. □

4.10 Upper and lower bounds for the approximation of $E[W]$

For the approximation of $E[W]$, one can also identify upper and lower bounds that are relatively simple for calculations and analysis.

Remark 4.10.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} = \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty, \ \beta \neq 0,$

where μ, p, θ are fixed. Then the expectation of the waiting time has the following estimates for its asymptotic behavior:

$$\frac{\frac{1}{\beta^2}B_1 + (\frac{\mu}{p\theta} - \frac{1}{\beta^2} - \frac{\eta}{\beta}\sqrt{\frac{\mu}{p\theta}})B_2}{A_1 + B_1 - B_2} \lim_{S\to\infty} \leq \sqrt{S}E[W] \leq \frac{\frac{1}{\beta^2}B_1 + (\frac{\mu}{p\theta} - \frac{1}{\beta^2} - \frac{\eta}{\beta}\sqrt{\frac{\mu}{p\theta}})B_2}{A_2 + B_1 - B_2},$$

where $A_1 = \Phi(\frac{\eta}{\sqrt{1+\frac{p\theta}{\mu}}} + \frac{\beta}{\sqrt{1+\frac{\mu}{p\theta}}}), \quad A_2 = \Phi(\beta)\Phi(\eta), \quad B_1 = \frac{\varphi(\beta)\Phi(\eta)}{\beta}, \quad B_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}\exp^{\frac{\eta_1^2}{2}}\Phi(\eta_1), \quad \eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}.$

Proof. The proof is identical to the proof of Remark 4.4.1. Only, instead of the approximation of γ from (4.24) there is a need to take the approximation of the upper and lower bounds of γ from (4.61). □

4.11 The boundary cases for $\sqrt{S}E[W]$

In this section let us see what happens with $\sqrt{S}E[W]$ when either β or η goes to ∞. The cases are the same as in the analyzing of behavior of the approximation of $P(W > 0)$ and $\sqrt{S}P(block)$.

As we saw when proving the Theorem 4.9.1, the expectation of the waiting time can be written down in the following form

$$\sqrt{S}E[W] \approx \frac{1}{\mu}\frac{\sigma_1 + \sigma_2}{\gamma + \xi_1 - \xi_2}, \tag{4.110}$$

where

$$\sigma_1 \approx \frac{1}{\beta}\xi_1 = \frac{\varphi(\beta)}{\beta^2}\Phi(\eta), \tag{4.111}$$

$$\sigma_2 \approx \frac{1}{\beta}\xi_1 = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{\eta_1}{2}} \Phi(\eta)(\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}}). \tag{4.112}$$

Case a. $\sqrt{\frac{\theta}{\lambda}}(N - S - \frac{\lambda}{\theta}) \to \eta$, $(\eta \in (-\infty, +\infty))$, and $\sqrt{S}(1 - \frac{p\lambda}{S\mu}) \to +\infty$.

As was told in the previous analysis, in this case the number of agents is bigger than the number of customers which want to receive an agents' service. We see that in this case $P(W > 0)$. So, it is possible to suppose that expectation of the waiting time before the agents' service will tend to zero.

It is easy to see that

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sigma_1 = \lim_{\beta\to\infty}\varphi(\beta)\Phi(\eta) = 0,$$

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sigma_2 = \lim_{\beta\to\infty}\rho\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)(\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}}) = 0,$$

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\gamma = \lim_{\beta\to\infty}\int_{-\infty}^{\beta}\Phi(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t) = \lim_{\beta\to\infty}E[\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})] = 1.$$

So,

$$\lim_{\beta\to\infty}\lim_{S\to\infty}\sqrt{S}E[W] = \frac{1}{\mu}\frac{\sigma_1+\sigma_2}{\gamma+\xi_1-\xi_2} = 0.$$

Case b. $\sqrt{\frac{\theta}{\lambda}}(N - S - \frac{\lambda}{\theta}) \to \eta$, $(\eta \in (-\infty, +\infty))$, and $\sqrt{S}(1 - \frac{p\lambda}{S\mu}) \to -\infty$;

In this case the rate of customers is bigger than the number of agents and this implies that there are many people in the queue and all the system is busy. Thus, it is possible to suppose, that the expectation time multiplied by square root from the number of agents tends to infinity. This heuristic analysis is approved by the following mathematical calculations:

When $\beta \longrightarrow -\infty$ then

$$\eta_1 \longrightarrow \infty, \qquad \nu_1 \longrightarrow -\infty, \qquad \nu_2 \longrightarrow -\infty$$

First look at

$$\lim_{\beta\to\infty}\lim_{\lambda\to\infty}\gamma = \lim_{\beta\to\infty}\int_{-\infty}^{\beta}\Phi(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t) = 0.$$

We can see also that

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\sigma_1 = 0, \qquad \lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\sigma_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)(\beta\frac{\mu}{p\theta}),$$

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\widetilde{\xi}_1=0,\qquad \lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\widetilde{\xi}_2=\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1).$$

So,

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W]=\frac{\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)(\beta\frac{\mu}{p\theta})}{\frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}=\lim_{\beta\to-\infty}-\beta=\infty,$$

as we supposed.

Case c. $\sqrt{\frac{\theta}{\lambda}}(N-S-\frac{\lambda}{\theta})\to+\infty$, and $\sqrt{S}(1-\frac{p\lambda}{S\mu})\to\beta,\ \ (\beta\in(-\infty,+\infty))$.

In this case we have an infinite number of places in the queue. As we saw in Sections 4.3, in this case the behaviour of the agents' pool system looks like M/M/S queue system. When $\beta>0$ it is not a problem to come into the system and therefore the probability to block tends to zero. So, we can suppose that the expectation of the waiting time in queue before agent's service tends to expectation of variate which is distributed exponential with rate $\mu\beta$. When $\beta\leq 0$ it is an "explosion" of the number of customers in the system, so it is not so easy to come and almost every one wait in the queue.

There were our assumptions and now let us see mathematically what happens with the expectation of the waiting time in queue before the agent's service. Easy to see that

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\sigma_2=0,\qquad \lim_{\beta\to\infty}\lim_{\lambda\to\infty}\xi_2=0,$$

and when $\beta>0$

$$\begin{aligned}\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W]&=\lim_{\eta\to\infty}\frac{\frac{\varphi(\beta)}{\beta^2}\Phi(\eta)}{\mu(\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t)+\frac{\varphi(\beta)}{\beta}\Phi(\eta))}\\&=\lim_{\eta\to\infty}\frac{\frac{\varphi(\beta)}{\beta^2}}{\mu\left(\int_{-\infty}^{\beta}d\Phi(t)+\frac{\varphi(\beta)}{\beta}\right)}\\&=\frac{1}{\mu\beta}\left(1+\frac{\beta\Phi(\beta)}{\varphi(\beta)}\right)^{-1}.\end{aligned}$$

So,

$$\sqrt{S}E[W]\longrightarrow\begin{cases}\frac{1}{\mu\beta}\left(1+\frac{\beta\Phi(\beta)}{\varphi(\beta)}\right)^{-1}, & \text{when } \beta>0\\ +\infty, & \text{when } \beta<0.\end{cases}$$

From this formula one can say that the conditional expectation of the waiting time multiplied by the square root of the number of agents is following:

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W|W>0]=\frac{1}{\mu\beta}. \tag{4.113}$$

This fact is corresponded with the well-known asymptotic for conditional waiting time in the M/M/S queue system.

Case d. $\sqrt{\frac{\theta}{\lambda}}(N - S - \frac{\lambda}{\theta}) \to -\infty$, and $\sqrt{S}(1 - \frac{p\lambda}{S\mu}) \to \beta$, $(\beta \in (-\infty, +\infty))$.

In this case the number of places in the queue tends to zero, so it is difficult to come in the system and the expectation of the waiting time multiplied by $\sqrt{S}$ tends to infinity.

Using the approximation for $\Phi(x)$, when $x \to -\infty$, one obtains

$$\lim_{\eta\to-\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W] =$$

$$= \lim_{\eta\to-\infty} \frac{-\frac{\varphi(\beta)\varphi(\eta)}{\eta\beta^2} - \frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\beta(\eta-\beta\sqrt{\frac{\mu}{p\theta}})}\left(\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}}\right)}{\mu\left(\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t) - \frac{\varphi(\beta)\varphi(\eta)}{\eta\beta} + \frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\beta(\eta-\beta\sqrt{\frac{\mu}{p\theta}})}\right)}$$

$$= \lim_{\eta\to-\infty} \frac{\frac{\varphi(\beta)e^{-\frac{\eta^2}{2}}}{\beta^2\sqrt{2\pi}} + \frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\beta}\left(\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}}\right)}{\mu\left(\frac{\varphi(\beta)e^{-\frac{\eta^2}{2}}}{\beta\sqrt{2\pi}} + \frac{e^{-\frac{\eta^2+\beta^2}{2}}}{2\pi\beta}\right)}$$

$$= \lim_{\eta\to-\infty} \frac{\frac{\varphi(\beta)}{\beta^2\sqrt{2\pi}} + \frac{e^{-\frac{\beta^2}{2}}}{2\pi\beta}\left(\beta\frac{\mu}{p\theta} - \frac{1}{\beta} - \eta\sqrt{\frac{\mu}{p\theta}}\right)}{\mu\left(\frac{\varphi(\beta)}{\beta\sqrt{2\pi}} + \frac{e^{-\frac{\beta^2}{2}}}{2\pi\beta}\right)} = \infty,$$

as we supposed.

Analyzing the approximation when $\beta = 0$ we can say that

$$\lim_{\beta\to-\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W] = \begin{cases} \infty, & \text{when } \eta \to +\infty \\ 0, & \text{when } \eta \longrightarrow -\infty. \end{cases}$$

Chapter 5

Graphical analysis

5.1 Illustration of the approximations

The examples in this chapter illustrate the results found in the previous chapter. For this purpose we will present some graphs, which include both the real and approximation values. First the real values were calculated by a program written on Visual Basic (see Appendix), and the approximations' values were calculated on Excel. Next, all this data was processed in Excel and the graphs were built.

For the examination of approximations we compare the performance measures of small-, and two mid-sized call centers. The model of each call center is the same as it was in the previous chapter: The arrival process is a Poisson process with rate λ. There are N trunk lines and S agents in the system ($S \leq N$). First the customer is served by the IVR processor. We assume that IVR processing times are independent and identically distributed exponential random variables with the rate θ. After finishing the IVR process, a call may leave the system with the probability $1 - p$ or request service from an agent with the probability p. In each of these cases the average of call handling time in IVR is $\theta = 1$, the average of call-handling time by an agent is $\mu = 1$, and the probability for the call to be served by an agent is $p = 1$.

In the small call center the arrival rate λ is 10 customers per minute and the number of trunk lines is 50. The first mid-sized call center has an arrival rate of $\lambda = 50$ customers per minute and the number of trunk lines is 150. The second mid-sized call center has 200 trunk lines, and the arrival rate λ is 80 customers per minute. In each case the number of agents S is in the domain where the traffic intensity $\rho = \frac{\lambda p}{\mu S}$ is about 1. Namely, for the small call center the number of agents is between 1 and 30, in the first mid-sized call center it is between 30 and 70, and in the second mid-sized call center the number of agents is between

60 and 100.

5.1.1 The probability of delay $P(W > 0)$

First, look at the approximation for probability to wait in the small-sized call center.

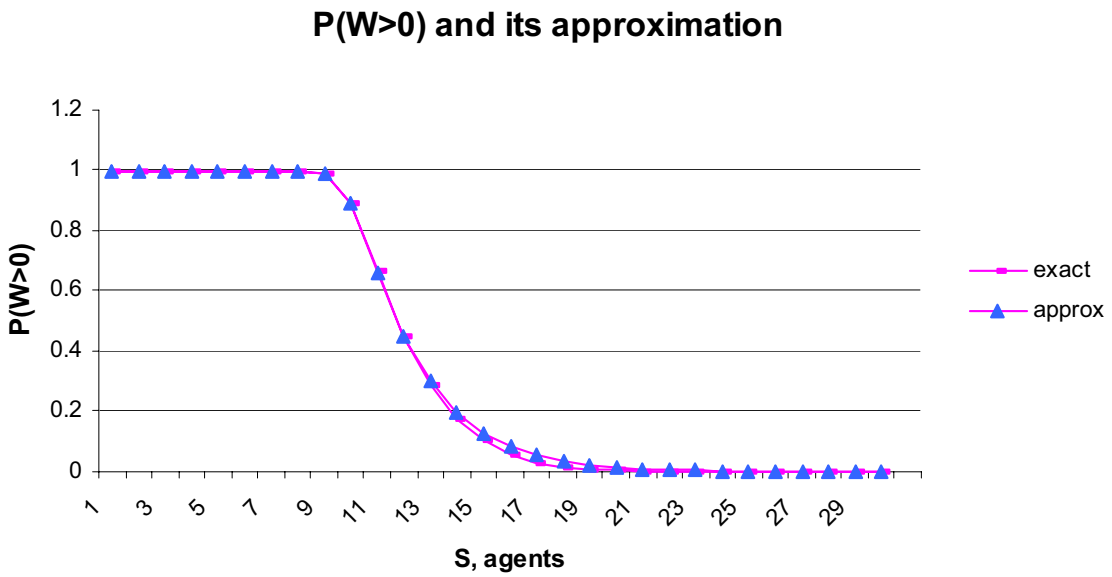

Figure 5.1: Comparison of the exact calculated probability to wait and it's approximation for a small-sized call center.

Figure 5.1 depicts the comparison of exact calculated probability to wait and its approximation. Mention that, in spite of small values of S and λ, the approximation is good enough.

In order to show the working of approximation in all graphs in this Chapter, we take a domain defined by the theorems about approximation.

Thus, in the case of the small-sized call center, the differences becomes noticeable when the number of agents is more than 20. Actually, it is possible to say that this is already not the domain, which was defined in the theorem. But nevertheless, the approximation still works well.

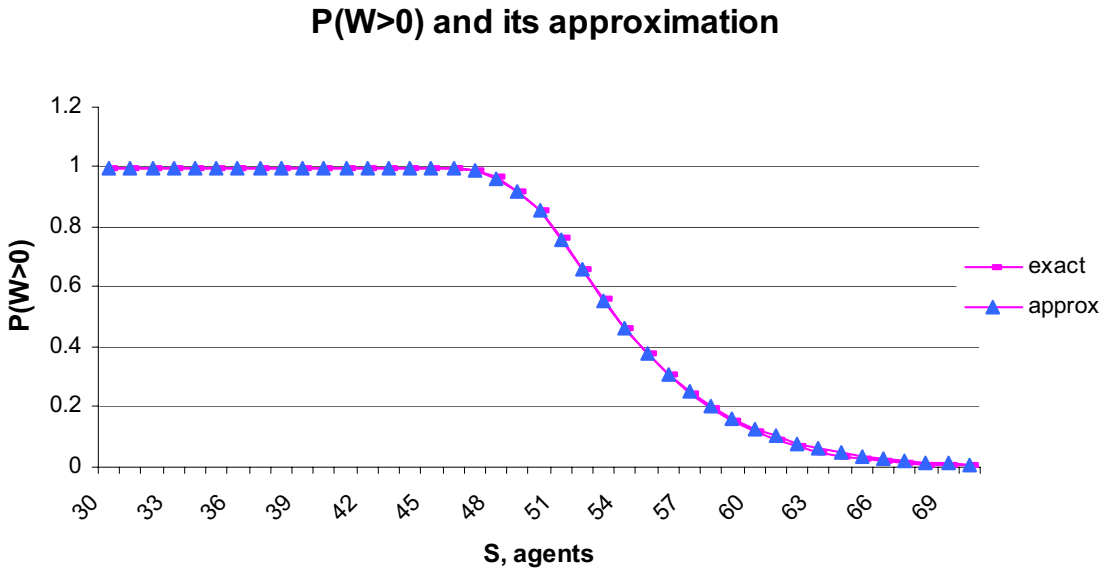

Figure 5.2: Comparison of the exact calculated probability to wait and it's approximation for a mid-sized call center with the arrival rate 50 and the number of trunk lines 150.

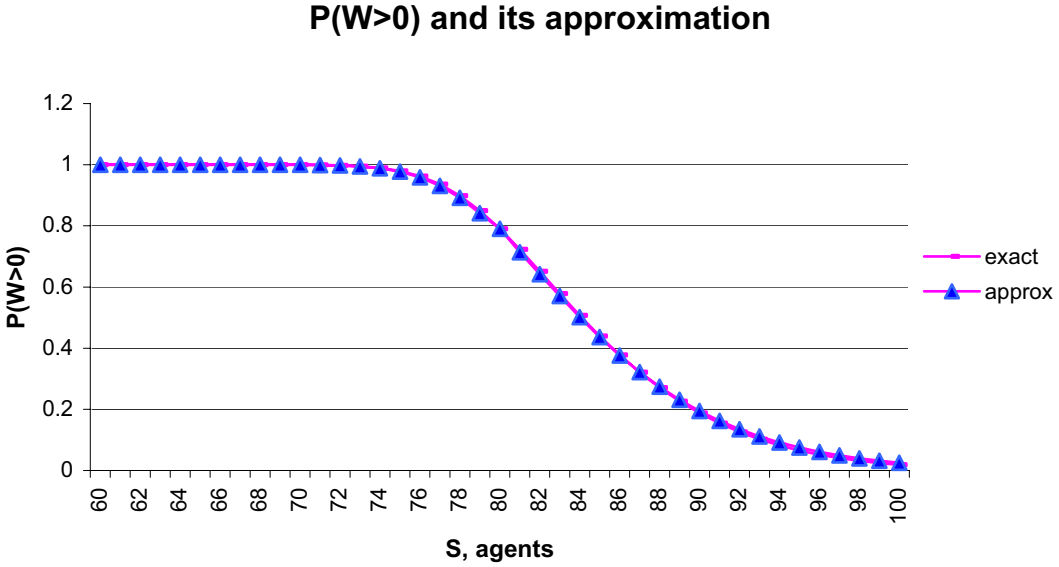

Figure 5.3: Comparison of the exact calculated probability to wait and its approximation for a mid-sized call center with the arrival rate 80 and the number of trunk lines 200.

In the case of mid-sized call centers, it is easy to see that the picture is much better. One of the conclusions which can be derived is the fact that the approximation which was founded is close to the exact value although in the small-sized call center. Certainly, when parameters of the system are increasing, the value of approximation goes to the exact value.

Let us note, that the calculation of the exact value is very difficult practically in the case of a bigger call center, for example, when the arrival rate λ is 500, the number of trunk lines N is 1500, and the number of agents S is between 450 and 550 agents. But, as we see from the following picture, the approximation is very close to the exact value. So, it can be used for the calculations of probability to wait in such call centers. In this case one obtains the following picture.

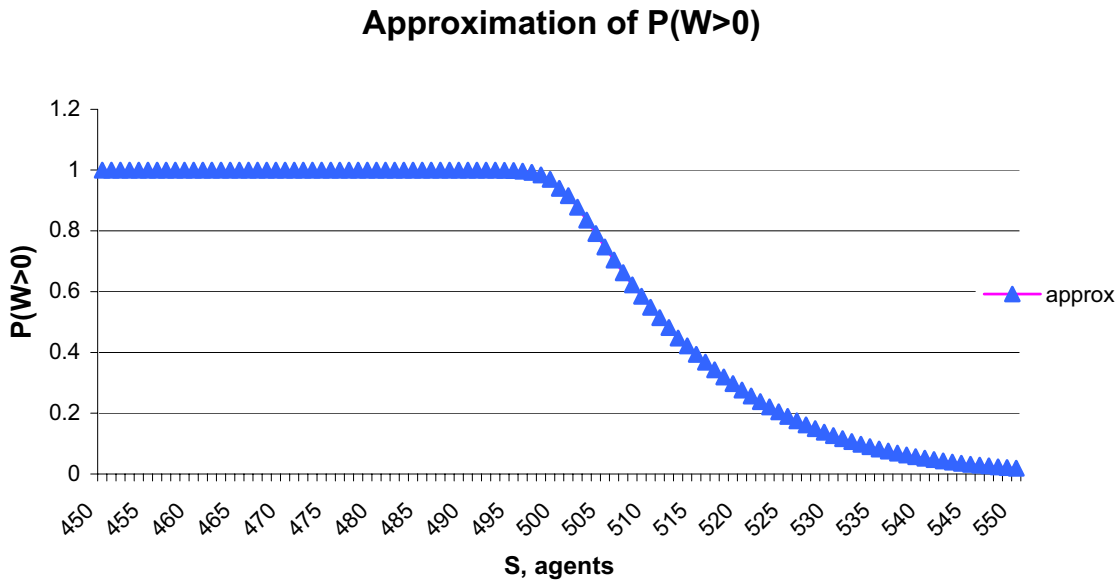

Figure 5.4: Comparison of the exact calculated probability to wait and it's approximation for a big call center with the arrival rate 500 and the number of trunk lines 1500.

It is worth mentioning that the calculation of the approximation is easy to do in Excel, Matlab, Maple and many others well-known programs.

5.1.2 The probability to find the system busy $P(block)$

To analyze the accuracy of approximation for $P(block)$ - probability to hear the sound that the all trunk lines are busy, we compare the exact calculated value of $P(block)$, and its approximation.

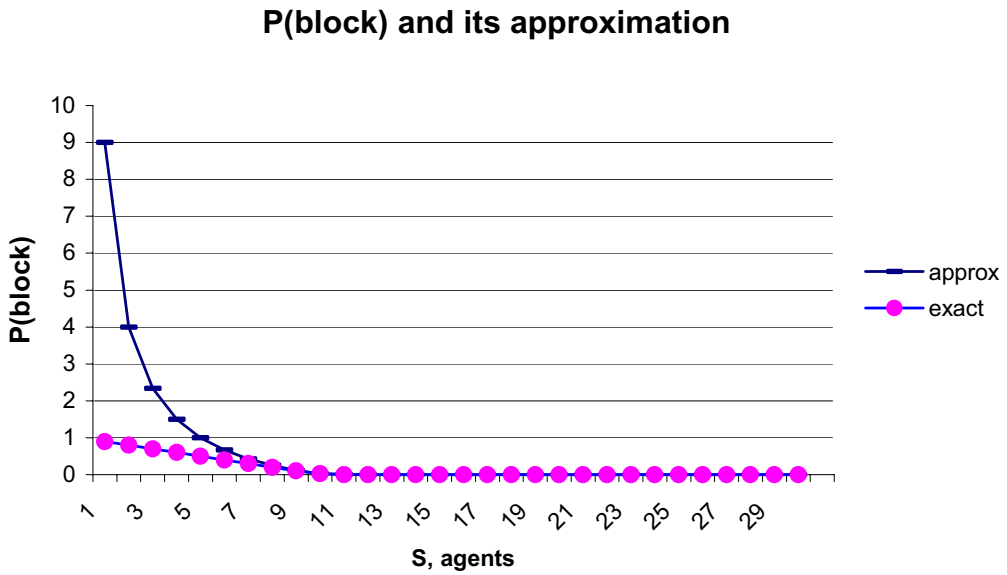

Figure 5.5: Comparison of the exact calculated probability of blocking and it's approximation for a small-sized call center.

Figure 3.16 depicts the probability to be blocked in a small-sized call center. As we can see again, in spite of small values, the approximation is good enough. Actually, it is even closer than in the case of the probability to wait.

From this graph we can see that at the beginning, when the number of agents is too small, the probability to wait is close to 0.9. It is easy to explain in the following way: in the case when there is only one agent in the system, it is full. So, the customers come to the system only when someone leaves the system. In this case, customers leave the system after servicing by an agent, i.e. with average rate 1 in unit of time. Thus, from 10 customers, which arrive to the system only one comes in. Therefore, the probability to be blocked is 0.9.

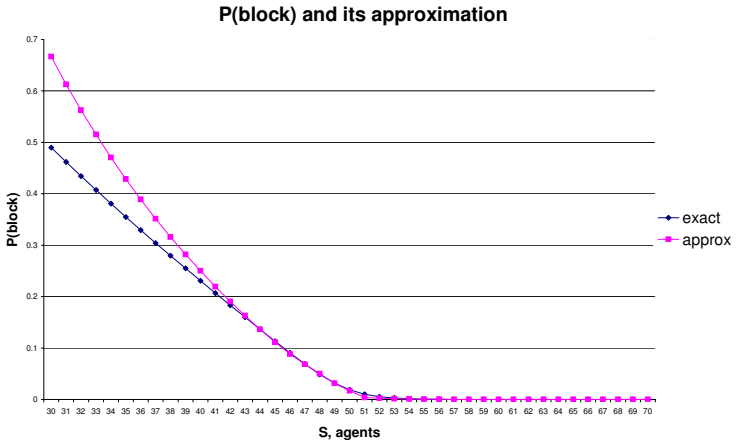

Figure 5.6: Comparison of the exact calculated probability of blocking and it's approximation for a mid-sized call center with the arrival rate 50 and the number of trunk lines 150.

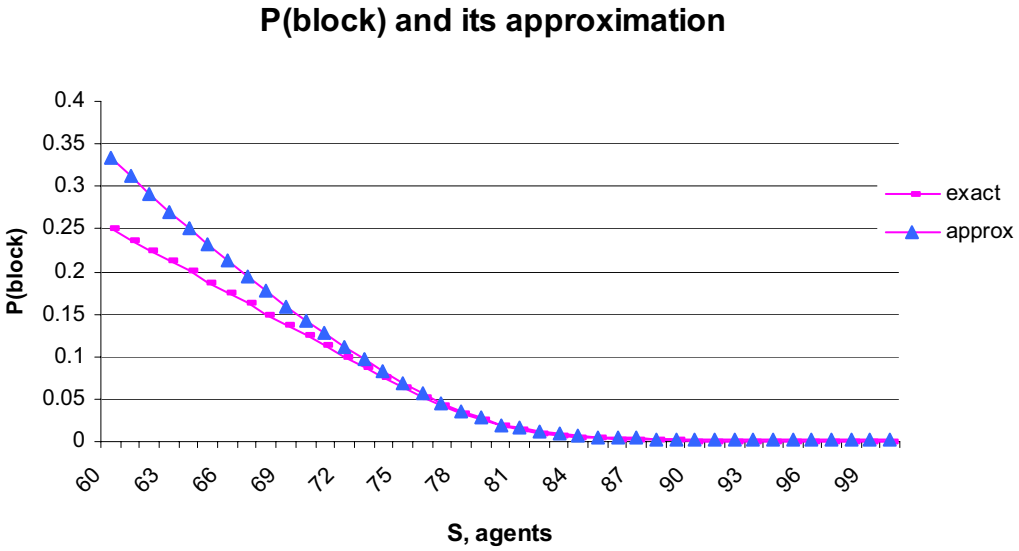

Figure 5.7: Comparison of the exact calculated probability of blocking and it's approximation for a mid-sized call center with the arrival rate 80 and the number of trunk lines 200.

In the case of mid-sized call centers we have the analogous pictures.

As in the previous section, one can conclude that the approximation is close to the exact value although in the small-sized call center. Certainly, when parameters of the system are increasing, the value of approximation goes closer to the exact value. So, in the case of a call center with the arrival rate $\lambda = 500$, the number of trunk lines $N = 1500$, and the number of agents $450 \leq S \leq 550$, one obtains the following picture.

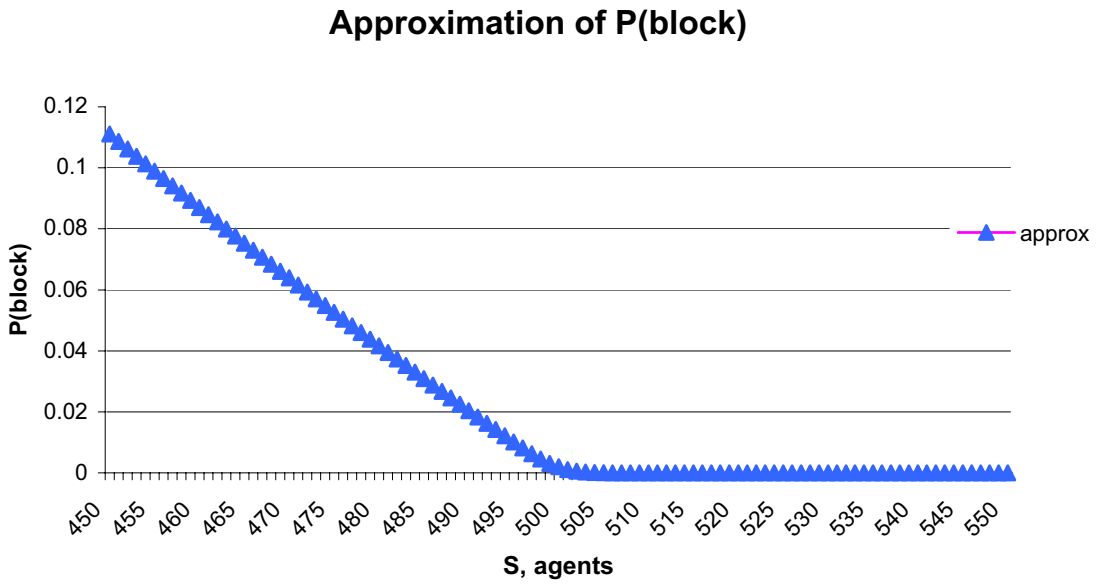

Figure 5.8: Comparison of the exact calculated probability of blocking and it's approximation for a big call center with the arrival rate 500 and the number of trunk lines 1500.

5.1.3 The expected waiting time $E[W]$

Finally, look at the the approximation for the $E[W]$ in the small-sized call center.

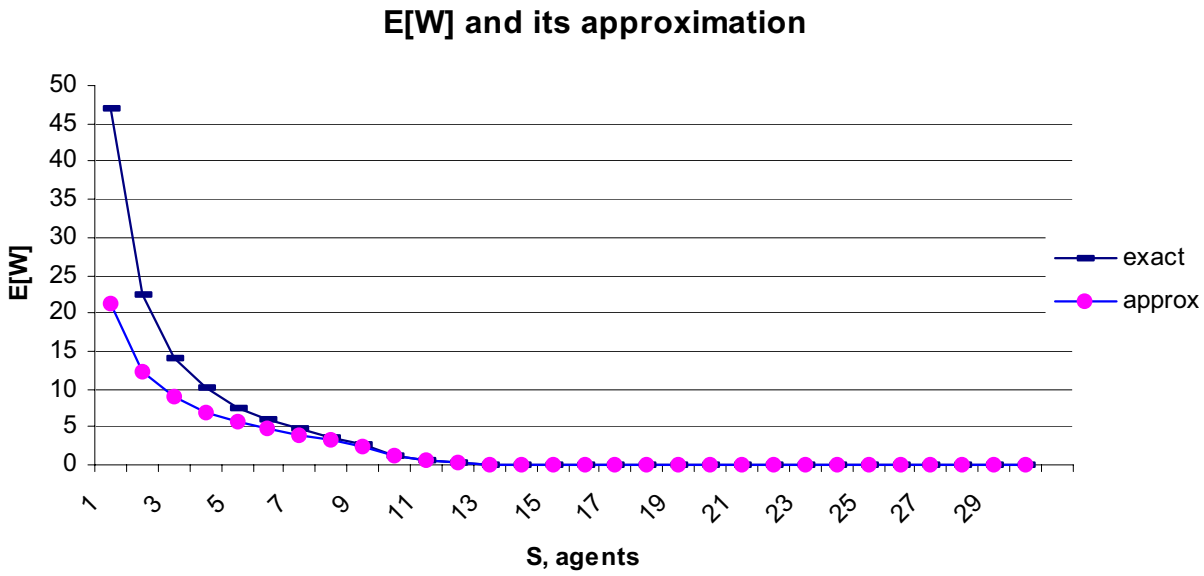

Figure 5.9: Comparison of the exact calculated expectation of the waiting time and it's approximation for a small-sized call center.

And again, in spite of small values, the approximation is good enough, when the number of agents is equal to 8 and more. Before this the offered load $\rho = \frac{\lambda p}{\mu S}$ is bigger than 1 and there are many places to wait. So, one can say that this is an "explosion" in the system. On the other side, the expectation of the waiting time is bounded from above by 49. This happens in the case, when the number of agents is 1, and the number of trunk lines is 49. From our model it is possible to conclude that in this case, the customers will wait on average 49 times the average of service time. So, their waiting time has Erlang(49,1) distribution, and the expectation is equal to 49, as we have.

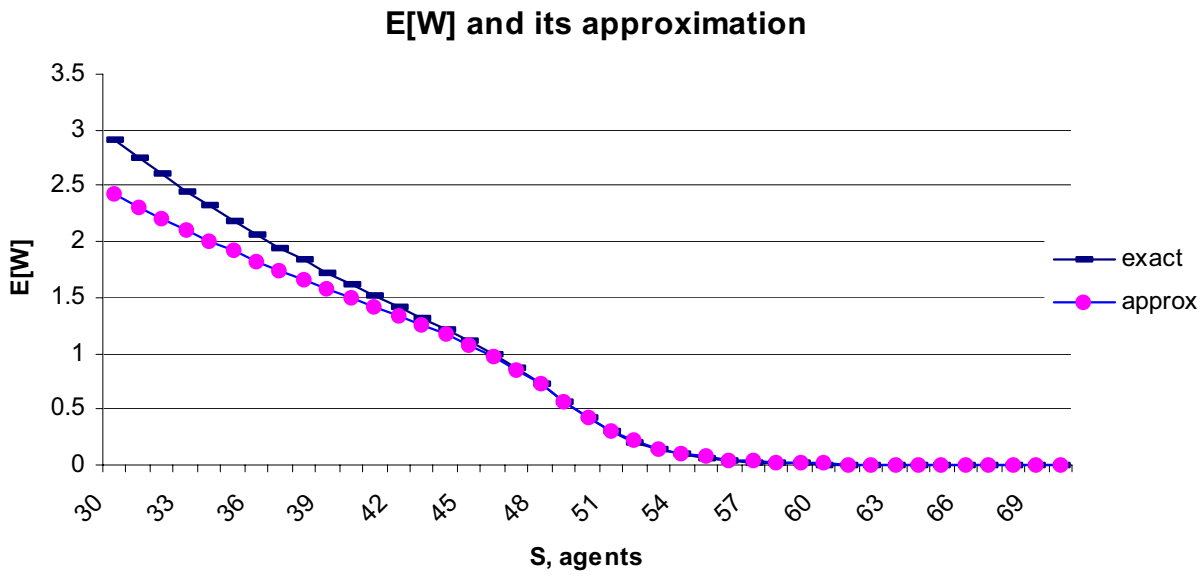

Figure 5.10: Comparison of the exact calculated expectation of the waiting time and it's approximation for a mid-sized call center with the arrival rate 50 and the number of trunk lines 150.

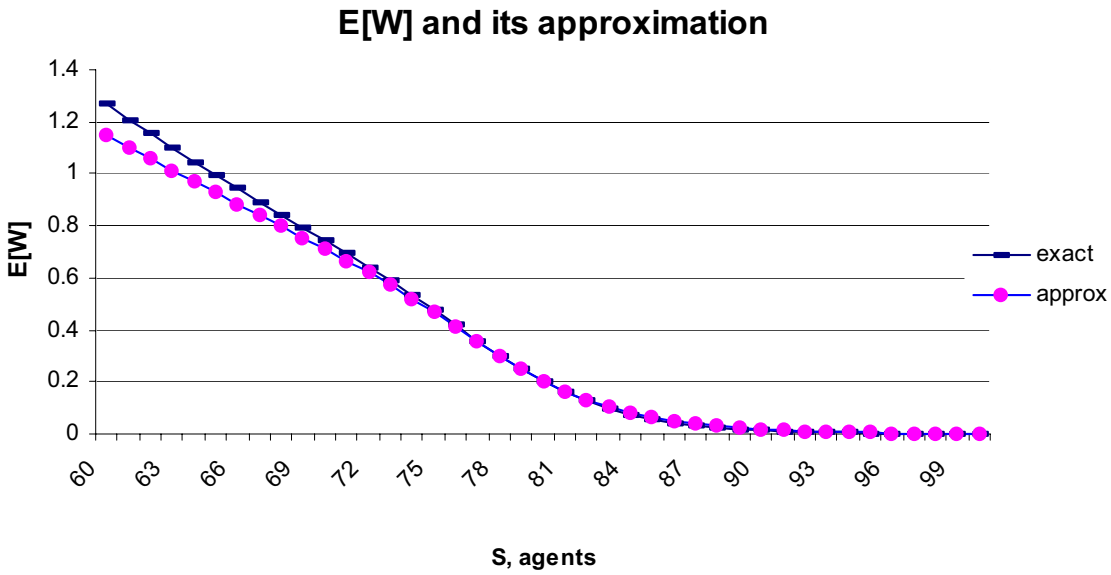

Figure 5.11: Comparison of the exact calculated expectation of the waiting time and it's approximation for a mid-sized call center with the arrival rate 80 and the number of trunk lines 200.

In the case of a mid-sized call center with arrival rate $\lambda = 50$ and the number of trunk lines $N = 150$, one can see that the approximation of expectation of the waiting time is close to the exact calculated value. In the call center with arrival rate $\lambda = 80$ and the number of trunk lines $N = 200$ the picture is more accurate. So, it is natural to suppose, that when $\lambda \to \infty$ the approximation is very close to the exact value.

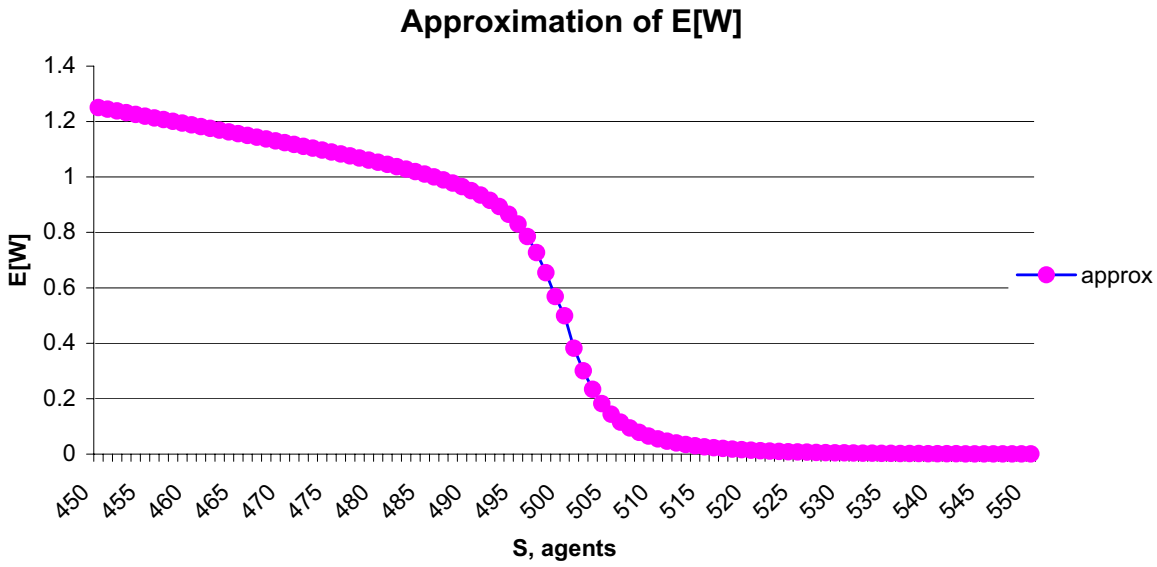

Figure 5.12: Comparison of the exact calculated expectation of the waiting time and it's approximation for a big call center with the arrival rate 500 and the number of trunk lines 1500.

Thus, without long and complicated calculations in the case of the great values, one can use the approximation formula for the expectation of the waiting time.

5.2 Effect of number of trunk lines on performance measures

In the previous section we analyzed the approximation for the probability to find a system busy $P(block)$, the probability to wait $P(W > 0)$ and the average waiting time $E[W]$ for different sized call centers with all parameters being constant exept for the agents number. The qualitative picture is the following: all the measures are decreasing with the growth of the number of agents. Now, let us analyze the case of a call center with all parameters constant except for the number of trunk lines. For this purpose, we use a mid-sized call center with the arrival rate 50, when the number of agents S in the range $30 \leq S \leq 70$ and the number of trunk lines is 95, 110, 150.

First, consider the behaviour of the probability to wait $P(W > 0)$ in such a call center.

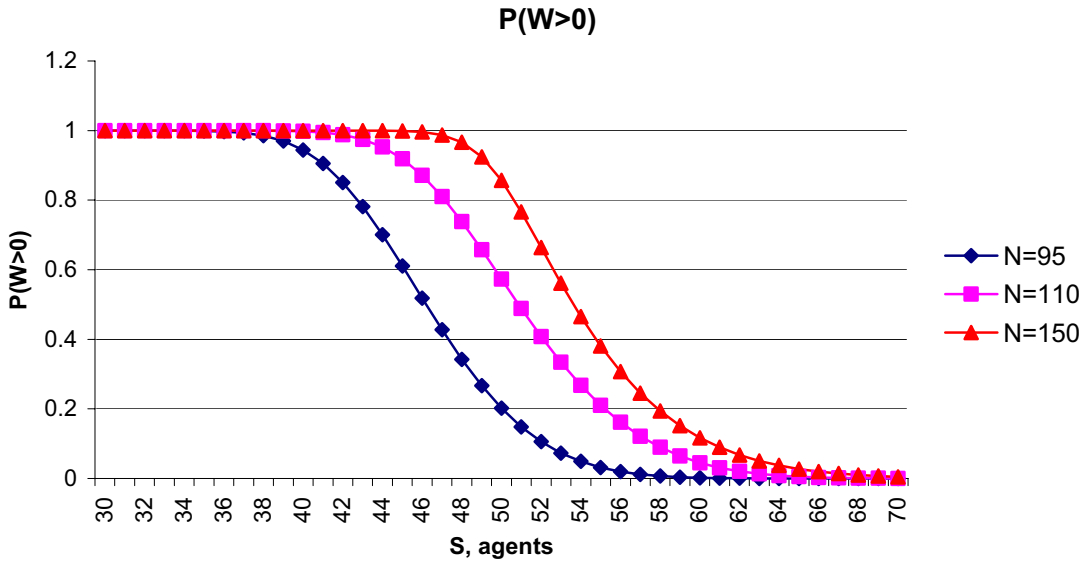

Figure 5.13: An illustration of the exact calculated probability to wait for a call center with arrival rate 50 and number of trunk lines 95, 110 and 150.

The figure above demonstrates that as the number of trunk lines increases so does the probability to wait $P(W > 0)$. Consequently, in case of increasing the number of trunk lines, if we want to keep the probability to wait constant, we need to increase the number of agents. For example, the probability to wait in a call center with 95 trunk lines and 50 agents is 0.2. Adding 15 trunk lines will

increase the probability to wait up to 0.6. In order to reduce it back to 0.2 the call center must add 5 extra agents.

A similar behaviour we identify in the expectation of the waiting time $E[W]$.

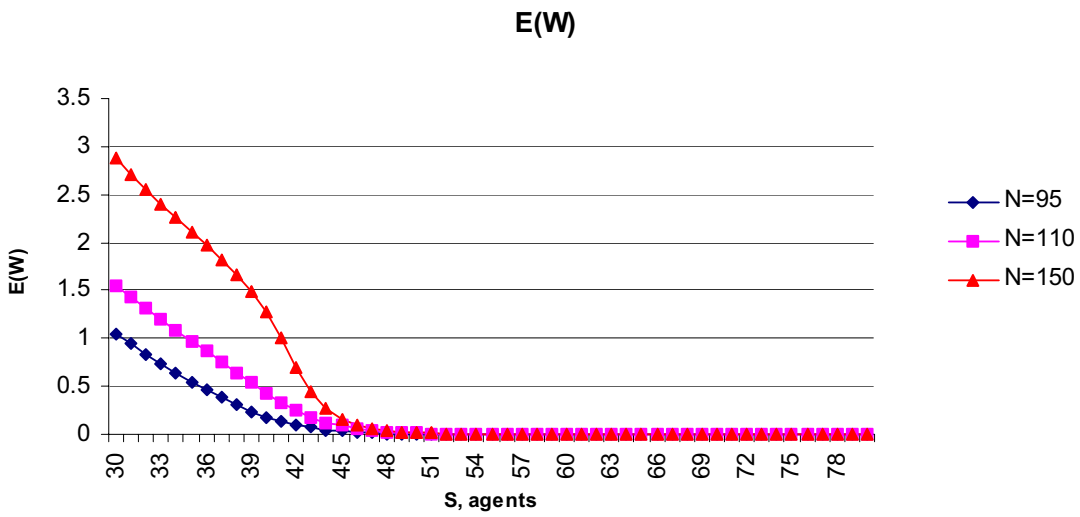

Figure 5.14: An illustration of the exact calculated expectation of the waiting time for a call center with arrival rate 50 and number of trunk lines 95, 110 and 150.

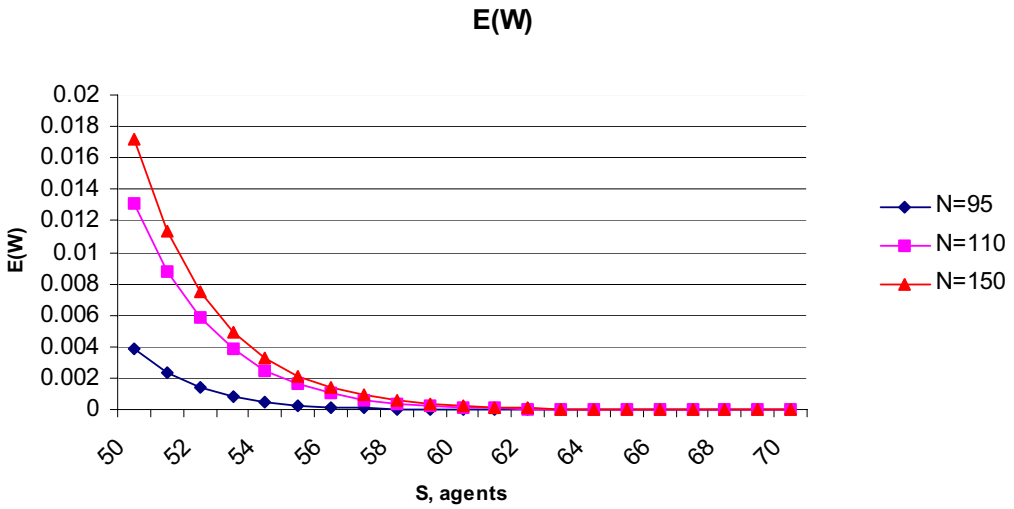

Figure 5.15: An illustration of the exact calculated expectation of the waiting time for a call center with arrival rate 50 and number of trunk lines 95, 110 and 150.

In Figure 5.14 we see that as the number of trunk lines increasing, so does the average waiting time $E[W]$. When the number of agents is 45 or more those changes seems not significant. Figure 5.15 provides a closer look at those values of agents number S and as can be seen the changes are very small. As in the previous case, the value of expectation of the waiting time is decreasing with decreasing of the number of trunk lines. Thus, in order to reduce the probability to wait and the average waiting time we should reduce the number of trunk lines. Having that, we should now examine how it will affect the probability to find the system busy.

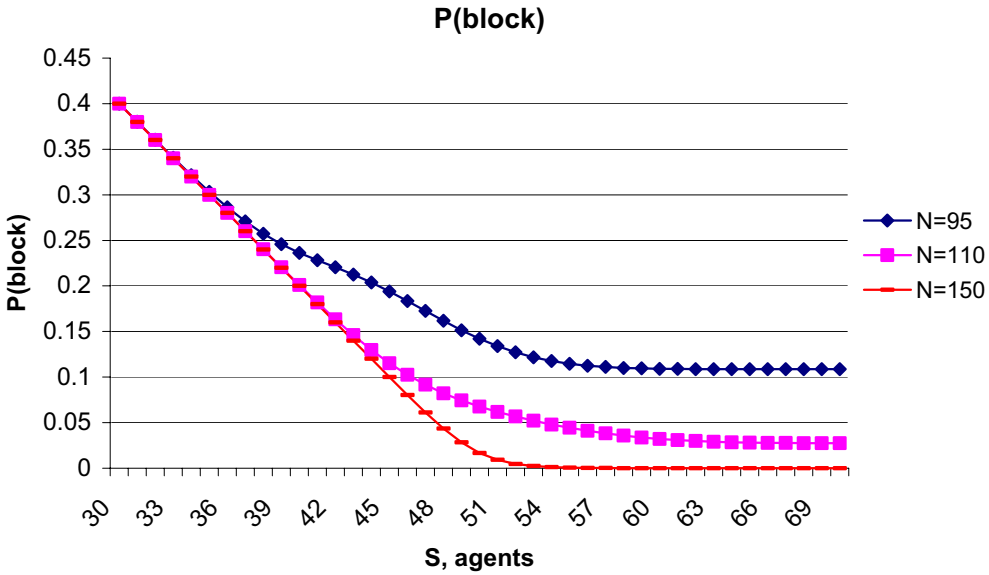

Figure 5.16: An illustration of the exact calculated probability to hear a busy sound for a call center with arrival rate 50 and number of trunk lines 95, 110 and 150.

Figure 5.16 shows that the reduction of the trunk lines number causes this probability to be larger. Moreover, small changes in the trunk lines number may cause a significant changes in the probability to find the system busy. For example, taking a call center with 110 trunk lines and 60 agents, are reducing its trunk lines number by 15, will enlarge the probability to find the system busy from 0.03 to 0.11. Changes to this probability caused by additional agents will be negligible.

We have considered the behaviour of performance measures just in one special case. There is no doubt that in the case with higher value of λ we will get even better approximations, but what is important to us is the general tendency of changing these values.

Chapter 6

The waiting time distribution

6.1 Density of the waiting time

The distribution function of the waiting time was given in Section 3 with the awkward expression (3.10), which is not so easy to calculate. Let us try to simplify this formula. For this purpose, we find first the density function of the waiting time. Start with

$$\begin{aligned} W(t) &= 1 - \sum_{i=1}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!} \\ &= 1 - \sum_{i=1}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) e^{-\mu S t} \left(1 + \frac{\mu S t}{1!} + ... + \frac{(\mu S t)^{j-S}}{(j-S)!} \right). \end{aligned}$$

To find the density function, let us take the derivative of the distribution function

$$\begin{aligned} f_W(t) = &- \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) \left[-\mu S e^{-\mu S t} \left(1 + \frac{\mu S t}{1!} + ... + \frac{(\mu S t)^{j-S}}{(j-S)!} \right) \right. \\ &\left. + e^{-\mu S t} \left(\mu S + \frac{(\mu S)^2 t}{1!} + ... + \frac{(\mu S t)^{j-S} t^{j-S-1}}{(j-S-1)!} \right) \right] \\ &= \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) \frac{e^{-\mu S t} (\mu S)^{j-S+1} t^{j-S}}{(j-S)!} \end{aligned}$$

Now let us find the Laplace transform of this density function:

$$\begin{aligned} L_W(x) = E(e^{-xt}) &= \sum_{i=1}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) \int_0^\infty \frac{e^{-\mu S t} (\mu S)^{j-S+1} t^{j-S}}{(j-S)!} e^{-xt} dt \\ &= \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j) \left(\frac{\mu S}{\mu S + x} \right)^{j-S+1} \int_0^\infty \frac{e^{-t(\mu S + x)} (\mu S + x)^{j-S+1} t^{j-S}}{(j-S)!} dt. \end{aligned}$$

Notice that

$$\int_0^\infty \frac{e^{-t(\mu S+x)}(\mu S+x)^{j-S+1}t^{j-S}}{(j-S)!}dt = 1, \qquad \forall S \le j \le N \text{ and } \quad x > 0, \tag{6.1}$$

because the expression, which is in the integral, is a density function of the Gamma distribution.

It follows from (3.8) that, for all $j > S$, $\chi(k, j)$ is equal to

$$\chi(i+j, j) = \frac{\frac{1}{S!S^{j-S}}(\frac{p\lambda}{\mu})^j \frac{i}{i!}(\frac{\lambda}{\theta})^i}{\frac{1}{\theta}(A+B)}, \tag{6.2}$$

where

$$A = \sum_{i+j \le N-1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j; \tag{6.3}$$

$$B = \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j. \tag{6.4}$$

Substituting (6.1) and (6.2) into (6.1) one obtains

$$\begin{aligned} L_W(x) &= \frac{\frac{S^S}{S!}\left(\frac{\mu S+x}{\mu S}\right)^{S-1}}{A+B}\left(\frac{p\lambda}{\mu S+x}\right)^S \sum_{i=0}^{N-S-1}\frac{1}{i!}(\frac{\lambda}{\theta})^i \sum_{j=0}^{N-S-i-1}\left(\frac{p\lambda}{\mu S+x}\right)^j \\ &= \frac{\frac{S^S}{S!}\left(\frac{\mu S+x}{\mu S}\right)^{S}}{A+B}\cdot\frac{\mu S}{\mu S+x}\left(\frac{p\lambda}{\mu S+x}\right)^S \sum_{i=0}^{N-S-1}\frac{1}{i!}(\frac{\lambda}{\theta})^i \frac{1-\left(\frac{p\lambda}{\mu S+x}\right)^{N-S-i}}{\frac{\mu S+x-p\lambda}{\mu S+x}} \\ &= \frac{\frac{S^S}{S!}}{A+B}\cdot\frac{\mu S}{\mu S(1-\frac{p\lambda}{\mu S})+x}\left(\frac{p\lambda}{\mu S}\right)^S \sum_{i=0}^{N-S-1}\frac{1}{i!}(\frac{\lambda}{\theta})^i\left[1-\left(\frac{p\lambda}{\mu S+x}\right)^{N-S-i}\right] \end{aligned}$$

By defining

$$a = \mu S, \quad b = \mu S - p\lambda, \quad \rho = \frac{p\lambda}{\mu S}, \tag{6.5}$$

and

$$C = \frac{1}{A+B}\cdot\frac{\mu S}{S!}\cdot\left(\frac{p\lambda}{\mu}\right)^S, \tag{6.6}$$

the Laplace transform can be rewritten as follows

$$L_W(x) = C\cdot\frac{1}{b+x}\sum_{i=0}^{N-S-1}\frac{1}{i!}(\frac{\lambda}{\theta})^i\left[1-\left(\frac{a-b}{a+x}\right)^{N-S-i}\right]. \tag{6.7}$$

To find the inverse Laplace transform we need to use the following lemma.

Lemma 6.1.1 *For all $n > 0$ the function*

$$f_n(x) = \frac{1}{b+x}[1 - \left(\frac{a-b}{a+x}\right)^n] \tag{6.8}$$

has an inverse Laplace transform, which is equal to

$$L^{-1}_{f_n}(t) = e^{-at}\left(1 + (a-b)t + ... + \frac{(a-b)^{n-1}t^{n-1}}{(n-1)!}\right). \tag{6.9}$$

Using this lemma one can find the inverse Laplace transform for the function $L_W(x)$ in (6.7):

$$\begin{aligned} L^{-1}_W(t) &= C\sum_{i=0}^{N-S-1}\sum_{j=0}^{N-S-i-1}\frac{1}{i!j!}(\frac{\lambda}{\theta})^i\left[(a-b)t\right]^j e^{-at} \\ &= C\sum_{k=0}^{N-S-1}\frac{e^{-at}}{k!}\sum_{i+j=k}\frac{k!}{i!j!}(\frac{\lambda}{\theta})^i\left[(a-b)t\right]^j \\ &= C\sum_{k=0}^{N-S-1}\frac{1}{k!}\left(\frac{\lambda}{\theta}+(a-b)t\right)^k e^{-at} \end{aligned}$$

In view of (6.5) and (6.6), the density function of the waiting time for the customers that finished their IVR service and continue to get agents' service is equal to

$$f_W(t) = \frac{1}{A+B}\cdot\frac{1}{S!}\cdot\left(\frac{p\lambda}{\mu}\right)^S\cdot\mu S e^{-\mu St}\sum_{k=0}^{N-S-1}\frac{\left[\lambda(\frac{1}{\theta}+pt)\right]^k}{k!}. \tag{6.10}$$

When multiplying and dividing (6.10) by the expression

$$\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\sum_{j=0}^{N-S-i-1}\left(\frac{p\lambda}{\mu S}\right)^j, \tag{6.11}$$

when $\beta \neq 0$ one obtains (see (4.46) and (4.50))

$$f_W(t) = \frac{P(W>0)(1-\rho)\mu S e^{-(1-\rho)\mu St}}{\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i(1-\rho^{N-S-i})}\cdot\sum_{k=0}^{N-S-1}\frac{\left[\lambda(\frac{1}{\theta}+pt)\right]^k}{k!}e^{-pt\lambda},$$

where $\rho = \frac{p\lambda}{\mu S}$. Dividing $f_W(t)$ by $P(W>0)$ one gets the conditional probability to wait in the case when $\beta \neq 0$:

$$f_{W|W>0}(t) = \frac{(1-\rho)\mu S e^{-(1-\rho)\mu St}}{\sum_{i=0}^{N-S-1}\frac{e^{-\frac{\lambda}{\theta}}}{i!}\left(\frac{\lambda}{\theta}\right)^i(1-\rho^{N-S-i})}\cdot\sum_{k=0}^{N-S-1}\frac{\left[\lambda(\frac{1}{\theta}+pt)\right]^k}{k!}e^{-\lambda(\frac{1}{\theta}+pt)}. \tag{6.12}$$

When $\rho < 1$, this function can be written as follows:

$$f_{W|W>0}(t) = f_X(x)\frac{P(Y < N-S)}{P(Z_1 < N-S) - \rho^{N-S}e^{-\frac{\lambda(1-\rho)}{\theta\rho}}P(Z_2 < N-S)} \tag{6.13}$$

where

$$X \sim exp(\mu(1-\rho)), \qquad Y \sim Pois(\lambda(\frac{1}{\theta}+pt)),$$
$$Z_1 \sim Pois(\frac{\lambda}{\theta}), \qquad Z_2 \sim Pois(\frac{\lambda}{\theta\rho}).$$

When $\beta = 0$, a density function of the waiting time as below shows:

$$f_W(t) = \frac{P(W>0)\mu S e^{-\mu S t}}{\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \sum_{j=0}^{N-S-i-1}\left(\frac{p\lambda}{\mu S}\right)^j}\sum_{k=0}^{N-S-1}\frac{\left(\lambda(\frac{1}{\theta}+pt)\right)^k}{k!}. \tag{6.14}$$

This is a conditional density function of the waiting time when $\beta = 0$ can also be rewritten as the following:

$$f_{W|W>0}(t) = \frac{\mu S}{\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \sum_{j=0}^{N-S-i-1}\left(\frac{p\lambda}{\mu S}\right)^j}\sum_{k=0}^{N-S-1}\frac{\left(\lambda(\frac{1}{\theta}+pt)\right)^k}{k!}e^{-p\lambda t}. \tag{6.15}$$

This function can be represented in the following way:

$$f_{W|W>0}(t) = \frac{\mu S \cdot P(Y < N-S)}{(N-S)P(Z_1 < N-S) - \frac{\lambda}{\theta}P(Z_1 < N-S-1)} \tag{6.16}$$

where

$$Y \sim Pois(\lambda(\frac{1}{\theta}+pt)), \qquad Z_1 \sim Pois(\frac{\lambda}{\theta}).$$

We have not been able, however, to find a probabilistic explanation or derivation of (6.16).

6.2 An approximation of $\frac{1}{\sqrt{S}}f_{W|W>0}(\frac{t}{\sqrt{S}})$

Let us assume that the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions

(i) $\lim\limits_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} = \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty \ \ \beta \neq 0,$

where μ, p, θ are fixed. Under this assumptions let us consider the asymptotic behavior of the conditional density function of the waiting time.

$$\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \frac{\sqrt{S}(1-\rho)\mu e^{-\sqrt{S}(1-\rho)\mu t}}{\sum\limits_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!}\left(\frac{\lambda}{\theta}\right)^i (1-\rho^{N-S-i})} \cdot \sum_{k=0}^{N-S-1} \frac{\left[\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})\right]^k}{k!} e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})}. \tag{6.17}$$

First, look at the last sum

$$\sum_{k=0}^{N-S-1} \frac{\left[\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})\right]^k}{k!} e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})} = P(X_\lambda < N-S),$$

where

$$X_\lambda \sim Pois(\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})), \qquad E[X_\lambda] = \lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}}), \qquad Var[X_\lambda] = \lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}}).$$

We can say that

$$\lim_{\lambda\to\infty} P(X_\lambda < N-S) = \lim_{\lambda\to\infty} P\left(\frac{X_\lambda - \frac{\lambda}{\theta} - \frac{\lambda pt}{\sqrt{S}}}{\sqrt{\frac{\lambda}{\theta}+\frac{\lambda pt}{\sqrt{S}}}} < \frac{N-S-\frac{\lambda}{\theta}-\frac{\lambda pt}{\sqrt{S}}}{\sqrt{\frac{\lambda}{\theta}+\frac{\lambda pt}{\sqrt{S}}}}\right).$$

The right side in the parentheses can be approximated in the following way

$$\begin{aligned} \frac{N-S-\frac{\lambda}{\theta}-\frac{\lambda pt}{\sqrt{S}}}{\sqrt{\frac{\lambda}{\theta}+\frac{\lambda pt}{\sqrt{S}}}} &\approx \frac{\eta\sqrt{\frac{\lambda}{\theta}}-\sqrt{\lambda p\mu t}}{\sqrt{\frac{\lambda}{\theta}+\sqrt{\lambda p\mu t}}} \\ &\approx \eta - \sqrt{p\mu\theta}t. \end{aligned} \tag{6.18}$$

Thus, by the Central Limits Theorem one obtains

$$\frac{X_\lambda - \frac{\lambda}{\theta} - \frac{\lambda pt}{\sqrt{S}}}{\sqrt{\frac{\lambda}{\theta}+\frac{\lambda pt}{\sqrt{S}}}} \Rightarrow N(0,1),$$

when $N(0,1)$ is a normal distributed random value. Then by using (6.18) and Theorem 4.2.1, one obtains

$$\lim_{\lambda\to\infty} P(X_\lambda < N-S) = \Phi(\eta - \sqrt{p\mu\theta}t). \tag{6.19}$$

Under our assumptions (i), (ii) in beginning of the section

$$\lim_{\lambda\to\infty} \sqrt{S}(1-\rho)\mu e^{-\sqrt{S}(1-\rho)\mu t} = \beta\mu e^{-\beta\mu t}. \tag{6.20}$$

Now, let us consider the denominator in (6.17)

$$\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i \left(1-\rho^{N-S-i}\right) = \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i - \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i \rho^{N-S-i} \tag{6.21}$$

As we have seen in (4.7)-(4.11) the first sum tends to $\Phi(\eta)$, so it remains only to estimate the second sum

$$\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i \rho^{N-S-i} = \rho^{N-S} e^{\frac{\lambda(1-\rho)}{\theta\rho}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta\rho}}}{i!} \left(\frac{\lambda}{\theta\rho}\right)^i. \tag{6.22}$$

It follows from (4.18), (4.22) that

$$\lim_{\lambda\to\infty} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta\rho}}}{i!} \left(\frac{\lambda}{\theta\rho}\right)^i = \Phi(\eta_1), \tag{6.23}$$

where

$$\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}.$$

Now let us find the limit of $\rho^{N-S} e^{\frac{\lambda(1-\rho)}{\theta\rho}}$ as $\lambda \to \infty$.

$$\rho^{N-S} e^{\frac{\lambda(1-\rho)}{\theta\rho}} = e^{\frac{\lambda(1-\rho)}{\theta\rho} + (N-S)\ln\rho}. \tag{6.24}$$

By using (4.13) to the degree of exponent one obtains as $\lambda \to \infty$

$$\begin{aligned}
\frac{\lambda(1-\rho)}{\theta\rho} + (N-S)\ln\rho &\approx \frac{\lambda(1-\rho)}{\theta\rho} - (N-S)(1-\rho) - \frac{1}{2}(N-S)(1-\rho)^2 \\
&\approx -(N-S-\frac{\lambda}{\theta\rho})(1-\rho) - \frac{1}{2}(N-S)(1-\rho)^2 \\
&\approx -\left(\eta\sqrt{\frac{\lambda}{\theta}} + \frac{\lambda}{\theta} - \frac{\lambda}{\theta\rho}\right)(1-\rho) - \frac{1}{2}\eta\sqrt{\frac{\lambda}{\theta}}(1-\rho)^2 \\
&\approx -\eta\sqrt{\frac{\lambda}{\theta}}(1-\rho) + \left(\frac{\lambda}{\theta\rho} - \frac{1}{2}\frac{\lambda}{\theta} -\right)(1-\rho)^2 \\
&\approx \left(\frac{1}{\theta\rho} - \frac{1}{2\theta}\right)\lambda(1-\rho)^2 - \eta\sqrt{\frac{\lambda}{\theta}}(1-\rho) \\
&\approx \left(\frac{1}{\theta\rho} - \frac{1}{2\theta}\right)\frac{\mu S}{p}(1-\rho)^2 - \eta\sqrt{\frac{\mu}{p\theta}}\sqrt{S}(1-\rho) \\
&\approx \frac{1}{2}\frac{\mu}{p\theta}\beta^2 - \eta\beta\sqrt{\frac{\mu}{p\theta}}.
\end{aligned}$$

So,

$$\lim_{\lambda\to\infty} \rho^{N-S} e^{\frac{\lambda(1-\rho)}{\theta\rho}} = e^{\frac{1}{2}\frac{\mu}{p\theta}\beta^2 - \eta\beta\sqrt{\frac{\mu}{p\theta}}}. \tag{6.25}$$

Combining (6.19), (6.20), (6.23) and (6.25) one obtains the following corollary.

Corollary 6.2.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty \ \ (\beta \neq 0);$

where μ, p, θ are fixed. Then the conditional density function of the waiting time, evaluated at $t/\sqrt{S}$ and divided by $\sqrt{S}$, has the following asymptotic behavior:

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \beta\mu e^{-\beta\mu t} \frac{\Phi(\eta - t\sqrt{p\mu\theta})}{\Phi(\eta) - e^{\eta_2}\Phi(\eta_1)}, \tag{6.26}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{p\theta}{\mu}}$, *and* $\eta_2 = \frac{1}{2}\frac{\mu}{p\theta}\beta^2 - \eta\beta\sqrt{\frac{\mu}{p\theta}}$.

Now consider the case when $\beta = 0$. As it was shown in the proof of Lemma 4.2.4, when $\beta = 0$, i.e. $\rho \to 1$ or $\rho = 1$,

$$\lim_{\rho\to 1} \sum_{j=0}^{N-S-i-1} \left(\frac{p\lambda}{\mu S}\right)^j = N - S - i.$$

So, we can say that

$$f_{W|W>0}(t) \approx \frac{\mu S}{\sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i (N-S-i)} \sum_{k=0}^{N-S-1} \frac{\left(\lambda(\frac{1}{\theta}+pt)\right)^k}{k!} e^{-p\lambda t}. \tag{6.27}$$

Now, let us assume that the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} = \sqrt{S}(1-\frac{\lambda p}{\mu S}) = 0,$

where μ, p, θ are fixed. Under this assumptions let us consider the asymptotic behavior of the $\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)$:

$$\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \frac{\mu S}{\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!}\left(\frac{\lambda}{\theta}\right)^i (N-S-i)} \sum_{k=0}^{N-S-1} \frac{\left(\lambda(\frac{1}{\theta}+pt)\right)^k}{k!} e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})}. \tag{6.28}$$

Using (6.19) one obtains

$$\lim_{\lambda\to\infty} \sum_{k=0}^{N-S-1} \frac{\left(\lambda(\frac{1}{\theta}+pt)\right)^k}{k!} e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})} = \Phi(\eta - t\sqrt{p\mu\theta}). \tag{6.29}$$

Now let us consider the denominator. According to (4.42), (4.41) and (9.1) it can be approximated as follows

$$\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i (N-S-i) \approx \sqrt{S}\left(\eta\sqrt{\frac{\mu}{p\theta}}\Phi(\eta) + \sqrt{\frac{\mu}{p\theta}}\varphi(\eta)\right). \tag{6.30}$$

So,

$$\lim_{\lambda\to\infty} \frac{\mu S}{\sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i (N-S-i)} = \frac{\mu}{\eta\sqrt{\frac{\mu}{p\theta}}\Phi(\eta) + \sqrt{\frac{\mu}{p\theta}}\varphi(\eta)}, \tag{6.31}$$

and we can formulate the following corollary.

Corollary 6.2.2 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions*

(i) $\lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty;$

(ii) $\lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = 0;$

where μ, p, θ are fixed. Then the conditional density function of the waiting time evaluated at $t/\sqrt{S}$ and divided by $\sqrt{S}$ has the following asymptotic behavior

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \frac{\mu\Phi(\eta - t\sqrt{p\mu\theta})}{\eta\sqrt{\frac{\mu}{p\theta}}\Phi(\eta) + \sqrt{\frac{\mu}{p\theta}}\varphi(\eta)}. \tag{6.32}$$

6.3 Graphical analysis of the approximation for $\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)$

This section is devoted to investigating the behavior of the approximation for $\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}, \beta, \eta\right)$. For this purpose, we plotted the graphs in the cases as in the previous section.

In order to simplify our analysis let us define

$$g(t, \beta, \eta) = \lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}, \beta, \eta\right).$$

First, look at the case, when η is negative, for instance $\eta = -2$. Let us take three value of β, specifically -1, 0 and 1. In the figure below, we can see how the function $g(t, \beta, \eta)$ behaves in each case.

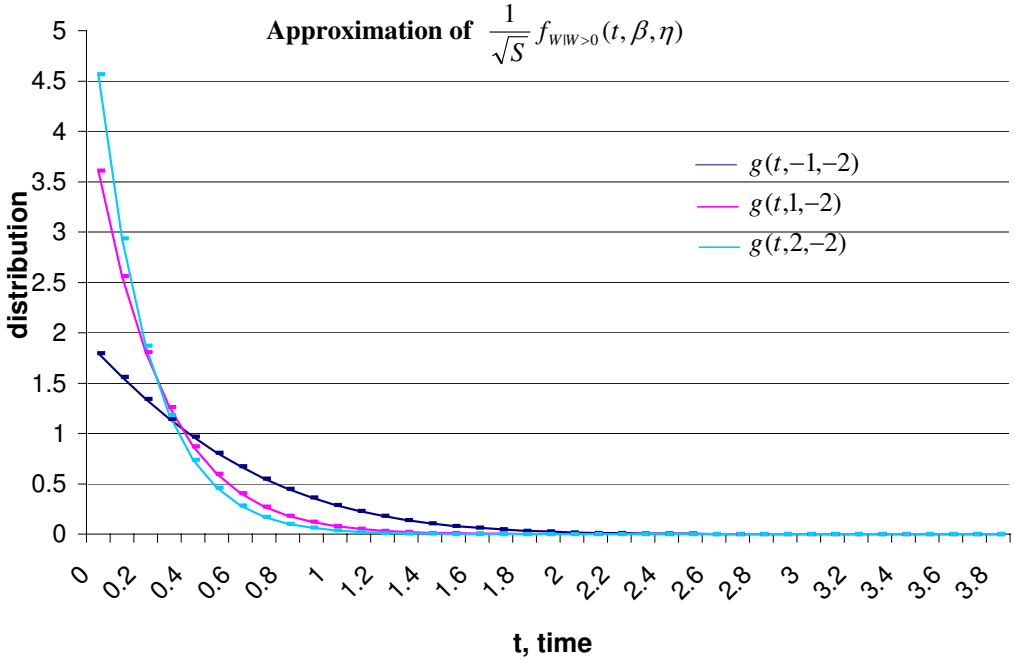

Figure 6.1: An illustration of the function $g\,(t,\beta,\eta)$ in the cases when $\eta = -2$ and β is equal to -1, 0 and 1.

We see that all the graphs have a similar form, which looks like the exponential density function. In each case we have different functions value when $t = 0$, and it is easy to see that this value is increasing with an increase of β. Let us try to approximate these lines by the density of an exponential distribution with parameter that is equal to the value of the function at the point $t = 0$.

The function at the origin is equal to $g(0,-1,-2) = 1.798$, and we have the following figure:

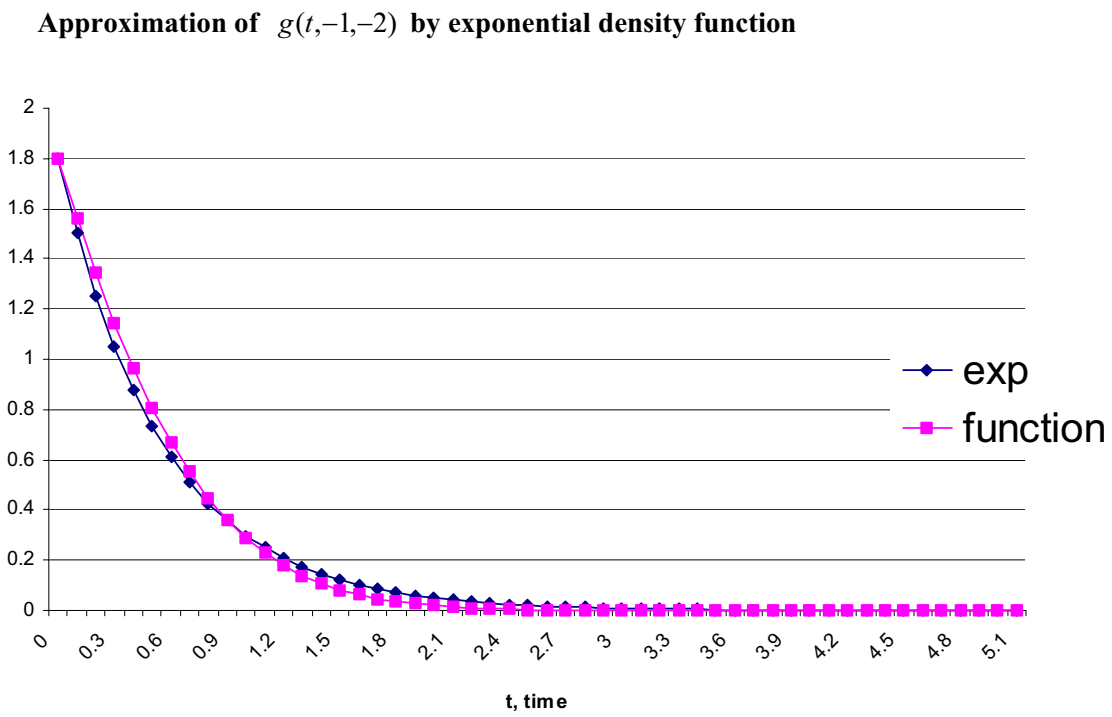

Figure 6.2: An illustration of the $g(t, \beta, \eta)$ approximation by an exponential density function with rate 1.798.

When $\beta = 0$, the rate of the exponential density function is equal to $g(0, -1, -2) = 2.675$.

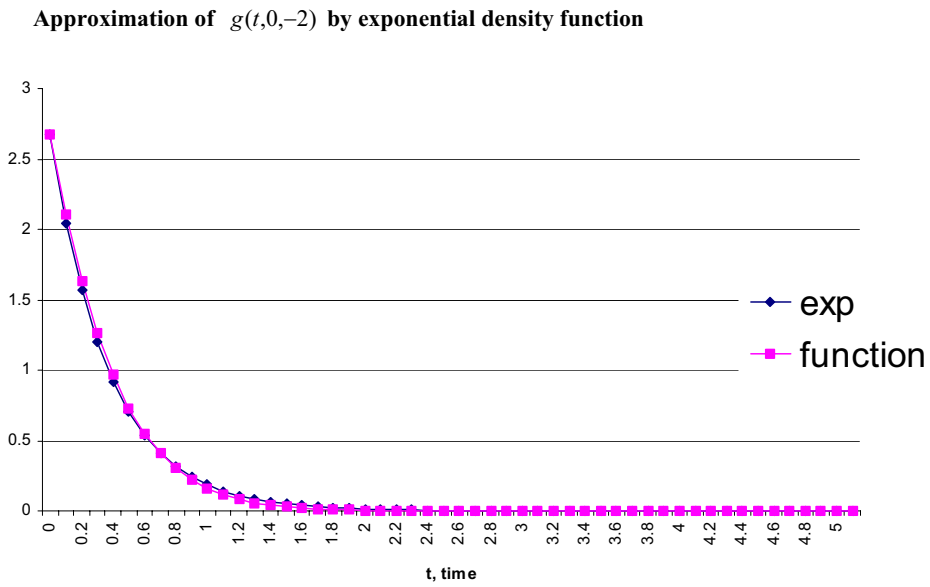

Figure 6.3: An illustration of the $g(t, \beta, \eta)$ approximation by exponential density function with rate 2.675.

In the last case, when $\beta = 1$, the rate of an exponential density function is equal to $g(0, -1, -2) = 3.61$.

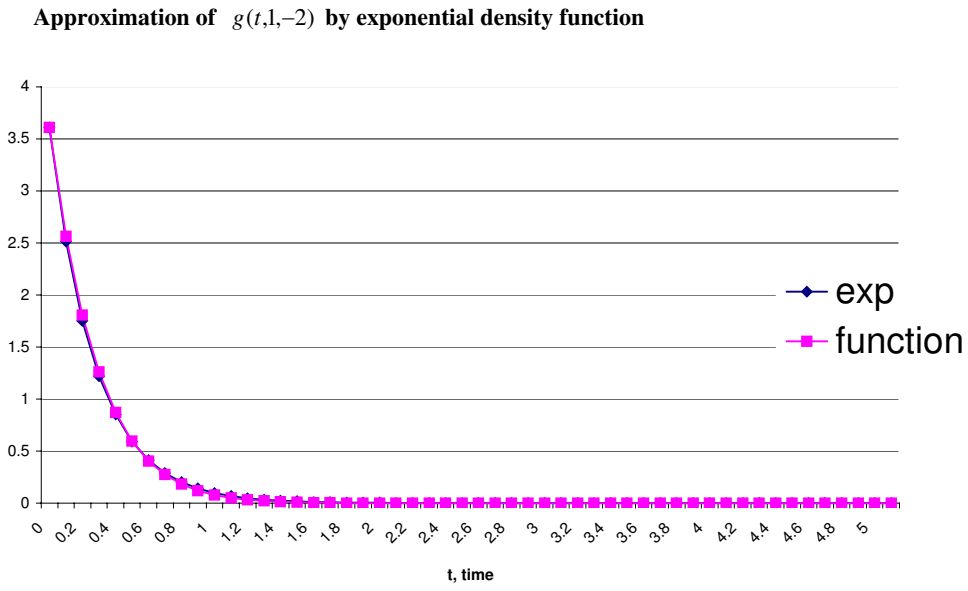

Figure 6.4: An illustration of the $g(t, \beta, \eta)$ approximation by an exponential density function with rate 3.61.

Figures 6.2-6.4 show that the exponential approximation is working satisfactory, especially in the last case. Under the condition (i) from Corollaries 6.2.1 and 6.2.2 $(N-S)-\frac{\lambda}{\theta} \approx \eta\sqrt{\frac{\lambda}{\theta}}$. Note, that $N-S$ is a number of customers in a queue. In the case where $\eta < 0$ (current case), the number of customers in a queue is smaller than the number of arrival customers. Also, under condition (ii) from Corollaries 6.2.1 and 6.2.2 $S \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}}$ and in our cases $-1 \leq \beta \leq 1$. Thus, the number of customers in a queue is smaller than the number of agents. Therefore, we can say that most of customers will have to wait a little time.

Figure 6.5 shows the case when $\eta = 0$ and the values of β equal: -1, 0 and 1. We see that the value of the function when $t = 0$ is increasing with the increase of β. But the forms of the graphs are different. Consider also Figure 6.6. It describes the case when $\eta = 2$. In this case, the graphs forms are different. When $\beta = -1$ most of the customers will wait a time between $\frac{1}{\sqrt{S}}$ and $\frac{3.6}{\sqrt{S}}$, and when $\beta = 0$ or $\beta = 1$, the larger percentage of customers will wait less than this interval of time.

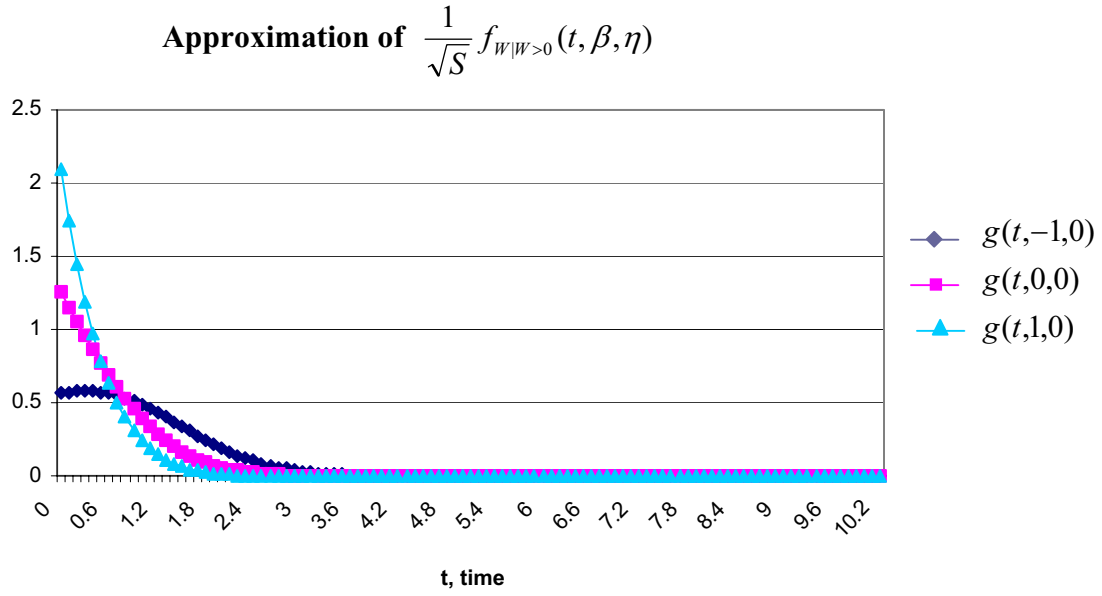

Figure 6.5: An illustration of the function $g(t,\beta,\eta)$ in the cases when $\eta = 0$ and β is equal to -1, 0 and 1.

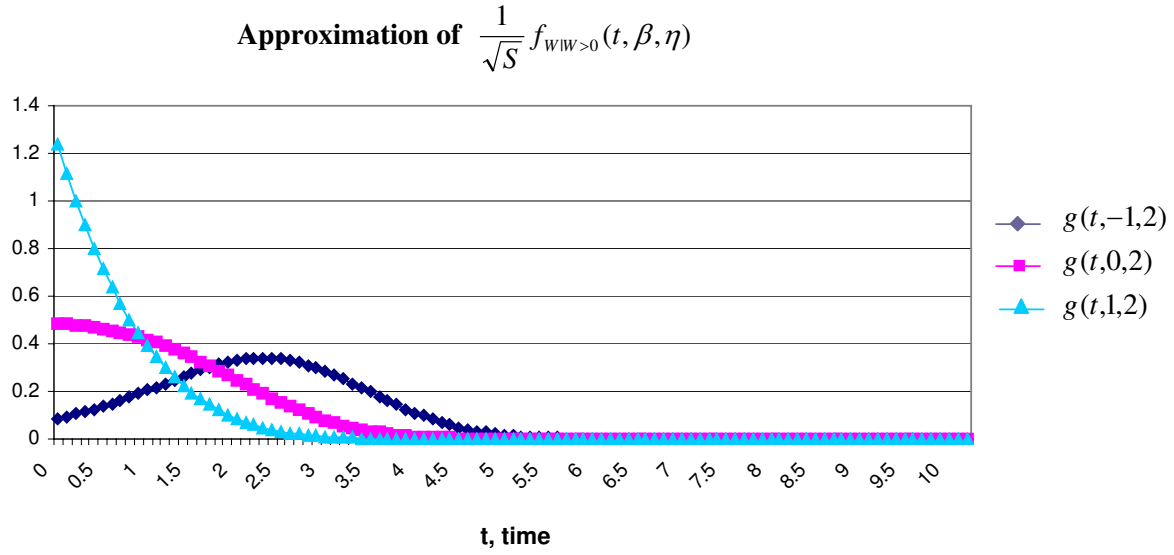

Figure 6.6: An illustration of the function $g(t,\beta,\eta)$ in the cases when $\eta = 2$ and β is equal to -1, 0 and 1.

At last, let us see what happen in the case when η is not a small value, for instance $\beta = 10$. In this case we have the following picture.

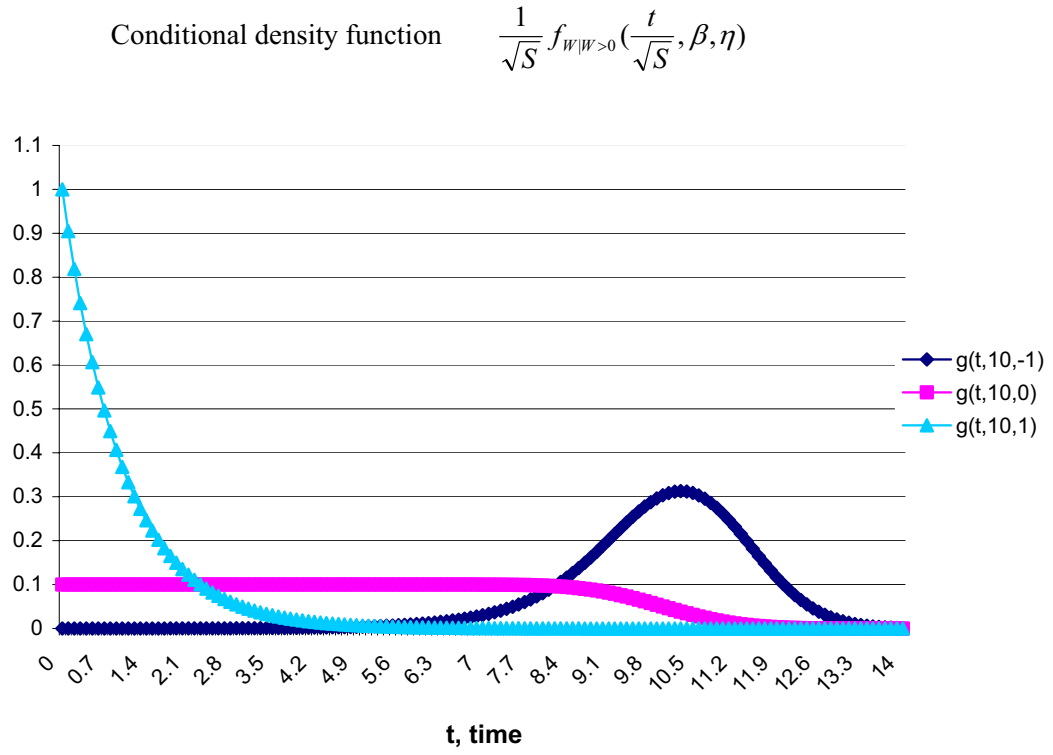

Figure 6.7: An illustration of the function $g(t, \beta, \eta)$ in the cases when $\eta = 10$ and β is equal to -1, 0 and 1.

Figure 6.7 shows that when η is high enough the density function multiplied by $\sqrt{S}$ changes in three different ways. Note that when η is a big value there are a big number of trunk lines in the call centers and the system of agents' pool can be described as the M/M/S queue model. Really, in the case when β is strictly positive the distribution of the waiting time seems like distribution of the exponential random variable. This supports the assumption about M/M/S queue model as a model for agents' pool. When β equals 0 or negative, we do not know much about the waiting time distribution. In these cases there are too many customers in the system, which waiting time is enormous and we call such a situation an "explosion" of the system. Our example give us an approximate picture about behaviour of the waiting time. Thus, when $\beta = 0$ we see that this distribution is basically (till $t = 9$) uniform with density function $\frac{1}{\eta}$. When $\beta < 0$ it is hard to match some specific known distribution without additional analysis. We can just say that this function looks like χ or bound Normal density distribution function.

Chapter 7

Special cases

In this chapter some special cases of our model are presented. In all such cases our model, under certain assumptions, becomes one of rather well-analyzed models, such as Erlang-B, Erlang-C and others. The goal is to show that our model and the results obtained for it coincide in these special cases with the well-known results for the corresponding models.

To facilitate the reading, we reproduce here our general model (Figure 3.2):

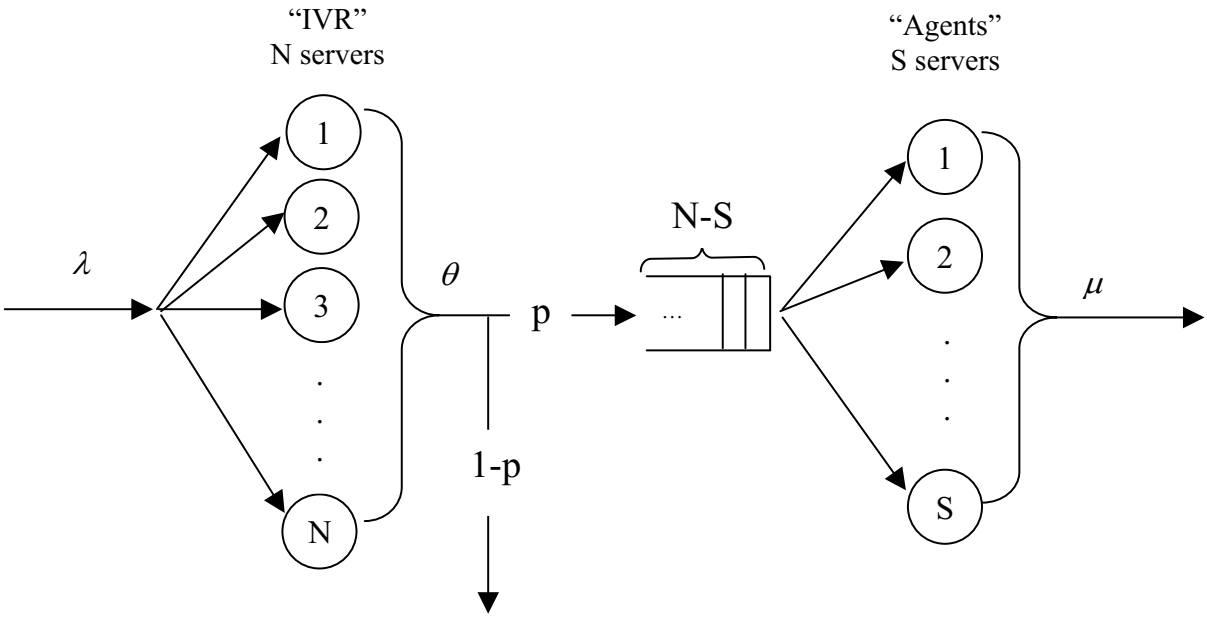

Figure 7.1: Schematic model of a queueing system with an interactive voice response, S agents and N trunk lines.

7.1 The M/G/S/S loss system

When the number of agents is equal to the number of trunk lines ($N = S$), the call center model can be presented as an M/G/S/S loss system. There is no waiting in this model. The service time G has the Phase Type distribution which is described on the following figure:

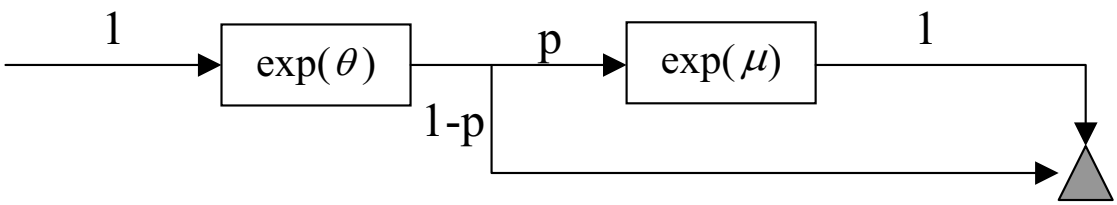

Figure 7.2: Schematic model of the Phase Type distribution, which corresponds to the service time in a call center with an IVR, when the number of agents is equal to the number of trunk lines ($N = S$).

The loss probability is then:

$$P_N(block) = \pi_N = \sum_{j=0}^{N} \pi(N-j, j) = \frac{\frac{\lambda}{N!}\left(\frac{1}{\theta} + \frac{p}{\mu}\right)}{\sum_{j=0}^{N} \frac{\lambda^j}{j!}\left(\frac{1}{\theta} + \frac{p}{\mu}\right)^j} \tag{7.1}$$

and analogously to Halfin and Whitt [17] (see also Jagerman [19]), it is possible to show that if

(i) $\lambda \to \infty$;

(ii) θ, p, μ are fixed;

(iii) $\sqrt{N}\left(1 - \frac{1}{N}\left(\frac{\lambda}{\theta} + \frac{p\lambda}{\mu}\right)\right) \to \bar{\beta}$, $-\infty < \bar{\beta} < \infty$

then

$$\sqrt{N} P_N(block) \to \frac{\phi(\bar{\beta})}{\Phi(\bar{\beta})} \tag{7.2}$$

where

$\phi(\cdot)$ is the standard normal density function;

$\Phi(\cdot)$ is the standard normal distribution function.

We thus recover a known result for the M/G/S/S queue.

It is easy to see that when $p = 0$, i.e. no one wishes to be served by an agent, our system is the well-known M/M/N/N queue (Erlang-B model).

Similarly, when the service time of the agents goes to 0 (μ goes to infinity), the system is equivalent to the M/M/S/S loss system with exponential service time with the rate θ. Only the IVR phase is taken into account. We have precisely the same picture if the service time in the IVR goes to 0 (θ goes to infinity). We still have the M/M/S/S loss system, but now the service time is exponential with the rate μ. In each case, by letting $\mu \to \infty$, or $\theta \to \infty$, approximation for the loss probability agrees with the well-known asymptotic for the Erlang-B formula (Jagerman [19]).

7.2 The M/M/S/N system

Masey and Wallace in [26] found approximations for the following operational characteristics for the M/M/S/N queue:

- the probability to find the system busy $P(block)$;
- the probability to wait more than t units of time $P(W > t)$;

when λ, S and N tend to ∞ simultaneously and:

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{\frac{\lambda}{\mu}}} = \eta, \quad 0 < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad 0 < \beta < \infty; \end{aligned} \tag{7.3}$$

The condition $\eta > 0$ in assumption (i) is completely natural, because $N - S$ is the maximal number of places in the queue, but the condition $\beta > 0$ is not required. The reason of strict positivity of β in [26] is using the M/M/S queue for finding the operational characteristics for M/M/S/N. Thus, in this section we find approximations for the probability to wait and the probability to find the system busy for M/M/S/N when $-\infty < \beta < \infty$. We also find approximations for the expected waiting time and the density of the waiting time, and show that with some specific parameters the system can be represented as the M/M/S/N queue.

7.2.1 Operational characteristics for M/M/S/N

Recall that the M/M/S/N queue has the following stationary distribution:

$$\pi_i = \begin{cases} \pi_0 \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i, & 0 \le i < S; \\ \pi_0 \frac{1}{S! S^{i-S}} \left(\frac{\lambda}{\mu}\right)^i, & S \le i \le N; \\ 0, & \text{otherwise.} \end{cases} \tag{7.4}$$

where

$$\pi_0 = \left(\sum_{i=0}^{S-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \sum_{i=S}^{N} \frac{1}{S! S^{i-S}} \left(\frac{\lambda}{\mu}\right)^i \right)^{-1}. \tag{7.5}$$

As in the previous analysis we define the waiting time by W. By using PASTA we can find the probability to wait:

$$P(W > 0) = \sum_{i=S}^{N} \pi_i = \frac{\sum_{i=S}^{N-1} \frac{1}{S! S^{i-S}} \left(\frac{\lambda}{\mu}\right)^i}{\sum_{i=0}^{S-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \sum_{i=S}^{N} \frac{1}{S! S^{i-S}} \left(\frac{\lambda}{\mu}\right)^i} \tag{7.6}$$

and the probability to find the system busy:

$$P(block) = \pi_N. \tag{7.7}$$

The expectation of the waiting time we can find by using Little's formula:

$$E[W] = \frac{L_{queue}}{\lambda_{eff}} = \frac{\sum_{i=S+1}^{N} (i-S)\pi_i}{\lambda(1 - P(block))}. \tag{7.8}$$

The conditional density function of the waiting time for M/M/S/N queue has the following form:

$$f_{W|W>0}(t) = \begin{cases} \frac{\mu S (1 - \frac{\lambda}{\mu S}) e^{-\mu S (1 - \frac{\lambda}{\mu S}) t}}{1 - \left(\frac{\lambda}{\mu S}\right)^{N-S}} \sum_{k=0}^{N-S-1} \frac{e^{-\lambda t} (\lambda t)^k}{k!}, & \rho \neq 1; \\ \frac{\mu S}{N-S} \sum_{k=0}^{N-S-1} \frac{e^{-\lambda t} (\lambda t)^k}{k!}, & \rho = 1. \end{cases} \tag{7.9}$$

In the case $\beta > 0$ this formula was found in [26] by Massey and Wallace. When $\beta \le 0$ it can be obtained with the help of Laplace transform in a way similar to that in Section 6.1.

Theorem 7.2.1 *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions* [1]*:*

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{\frac{\lambda}{\mu}}} = \eta, \quad 0 < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \frac{S - \frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad -\infty < \beta < \infty; \end{aligned} \tag{7.10}$$

where μ is fixed. Then

- *the probability to wait in the system $P(W > 0)$ has the following asymptotic behavior:*

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1 + \dfrac{\beta\Phi(\beta)}{\varphi(\beta)(1-e^{-\eta\beta})}\right)^{-1}, & \beta \neq 0; \\ \left(1 + \dfrac{\sqrt{\pi}}{\eta\sqrt{2}}\right)^{-1}, & \beta = 0; \end{cases} \tag{7.11}$$

- *the probability to find the system busy $P(block)$ has the following asymptotic behavior:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \begin{cases} \dfrac{\beta\varphi(\beta)e^{-\eta\beta}}{\beta\Phi(\beta) + \varphi(\beta)(1-e^{-\eta\beta})}, & \beta \neq 0; \\ \dfrac{1}{\sqrt{\dfrac{\pi}{2}} + \eta}, & \beta = 0; \end{cases} \tag{7.12}$$

- *the expectation of the waiting time has the following asymptotic behavior:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W] = \begin{cases} \dfrac{\dfrac{\varphi(\beta)}{\beta}\left[\dfrac{1-e^{-\eta\beta}}{\beta\mu} - \eta e^{-\eta\beta}\right]}{\beta\Phi(\beta) + \varphi(\beta)(1-e^{-\eta\beta})}, & \beta \neq 0; \\ \dfrac{\eta^2}{2\mu(\eta + \sqrt{\dfrac{\pi}{2}})}, & \beta = 0. \end{cases} \tag{7.13}$$

[1]Note, that these conditions can be also rewritten in the following form:

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{S}} = \eta, \quad 0 < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \frac{S - \frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad -\infty < \beta < \infty; \end{aligned}$$

- *the density function of the waiting time has the following asymptotic behavior:*

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{S}}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)=\begin{cases}\frac{\mu\beta e^{-\mu\beta t}}{(1-e^{-\eta\beta})}, & \mu t<\eta, \qquad \beta\neq 0;\\ \frac{\mu}{\eta}, & \mu t<\eta, \qquad \beta=0;\\ 0, & \mu t\geq\eta.\end{cases} \tag{7.14}$$

PROOF. Define

$$\gamma=\sum_{i=0}^{S-1}\frac{e^{-\frac{\lambda}{\mu}}}{i!}\left(\frac{\lambda}{\mu}\right)^i; \tag{7.15}$$

$$\xi=\sum_{i=S}^{N-1}\frac{e^{-\frac{\lambda}{\mu}}}{S!S^{i-S}}\left(\frac{\lambda}{\mu}\right)^i; \tag{7.16}$$

$$\delta=\frac{e^{-\frac{\lambda}{\mu}}}{S!S^{N-S}}\left(\frac{\lambda}{\mu}\right)^N; \tag{7.17}$$

$$\zeta=\frac{1}{\lambda}\sum_{i=S+1}^{N}\frac{e^{-\frac{\lambda}{\mu}}}{S!S^{i-S}}(i-S)\left(\frac{\lambda}{\mu}\right)^i. \tag{7.18}$$

Thus, we can rewrite operational characteristics of the M/M/S/N queue as follows:

$$P(W>0)=\frac{\xi}{\gamma+\xi}; \tag{7.19}$$

$$P(block)=\frac{\delta}{\gamma+\xi}; \tag{7.20}$$

$$E[W]=\frac{\zeta}{\gamma+\xi}. \tag{7.21}$$

Note that γ can be rewritten as $P(X_\lambda<S)$ where $X_\lambda\sim Pois(\frac{\lambda}{\mu})$, and $E[X_\lambda]=\frac{\lambda}{\mu}$, $Var[X_\lambda]=\frac{\lambda}{\mu}$. Then by the Central Limit Theorem

$$\frac{X_\lambda-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}\Rightarrow N(0,1).$$

Using condition (ii) of the Theorem and Theorem 4.2.1 one obtains

$$\gamma=P\left(\frac{X_\lambda-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}<\frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}\right)\to\Phi(\beta). \tag{7.22}$$

For convenience, let us denote $\frac{\lambda}{\mu S} = \rho$.

Let us find the approximation for ξ. In view of the Stirling's formula, $S! \approx \sqrt{2\pi S} S^S e^{-S}$, one obtains for ξ:

$$\xi \approx \frac{e^{S-\frac{\lambda}{\mu}}}{\sqrt{2\pi S}} \rho^S \sum_{k=0}^{N-S-1} \rho^k.$$

Making use of the expansion

$$\ln \rho = \ln(1-(1-\rho)) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2, \qquad (\rho \to 1), \quad (7.23)$$

one obtains

$$\xi \approx \frac{e^{S((1-\rho)-(1-\rho)-\frac{(1-\rho)^2}{2})}}{\sqrt{2\pi S}} \sum_{k=0}^{N-S-1} \rho^k = \frac{e^{\frac{S(1-\rho)^2}{2}}}{\sqrt{2\pi S}} \sum_{k=0}^{N-S-1} \rho^k \approx \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi S}} \sum_{k=0}^{N-S-1} \rho^k. \quad (7.24)$$

Due to the conditions (i) and (ii) of theorem and (7.23) the expression ρ^{N-S} can be rewritten in the equivalent form

$$\rho^{N-S} = e^{(N-S)\ln\rho} \approx e^{(N-S)}(\rho - 1) \approx e^{-\eta\beta}. \quad (7.25)$$

When $\beta \neq 0$, then $\rho \neq 1$ and the last sum in (7.24) has the following form:

$$\sum_{k=0}^{N-S-1} \rho^k = \frac{1-\rho^{N-S}}{1-\rho} \approx \frac{1-e^{-\eta\beta}}{1-\rho}.$$

Therefore, when $\beta \neq 0$

$$\lim_{\lambda\to\infty} \xi = \frac{\varphi(\beta)}{\beta}(1-e^{-\eta\beta}). \quad (7.26)$$

Now, let $\beta = 0$, i.e. $\rho = 1$ or $\rho \to 1$. When $\rho = 1$ it is easy to see that

$$\sum_{i=S}^{N-1} \rho^i = N - S.$$

When $\rho \to 1$ the last sum in (7.24) can be approximate as

$$\frac{\rho^{N-S}-1}{\rho-1} = \frac{e^{(N-S)\ln\rho}}{\rho-1} \approx \frac{(N-S)\ln\rho}{\rho-1} \approx N - S.$$

Here we have used the relation

$$(N-S)\ln\rho = o(1), \qquad (\rho \to 1).$$

The Stirling's formula and (7.23) imply

$$\xi \approx \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi S}}(N-S).$$

Using the conditions (i) and (ii) of the Theorem, one obtains

$$\xi \approx \frac{1}{\sqrt{2\pi S}}\eta\sqrt{\frac{\lambda}{\mu}} \approx \frac{\eta}{\sqrt{2\pi}}.$$

Thus, when $\beta = 0$

$$\lim_{\lambda\to\infty} \xi = \frac{\eta}{\sqrt{2\pi}}. \tag{7.27}$$

Now, consider δ. By the Stirling's formula it can be rewritten as

$$\delta \approx \frac{e^{S(1-\rho)}}{\sqrt{2\pi S}}\rho^S\rho^{N-S}.$$

Using relations (7.23) and (7.25), one obtains

$$\lim_{\lambda\to\infty} \sqrt{S}\delta = \varphi(\beta)e^{-\eta\beta}. \tag{7.28}$$

In order to find an approximation for ζ let us use the Stirling's formula. Thus, we can rewrite ζ by the following way:

$$\zeta \approx \frac{e^{S(1-\rho)}}{\lambda\sqrt{2\pi S}}\sum_{k=1}^{N-S} k\rho^{k+S} = \frac{e^{S(1-\rho)}}{\lambda\sqrt{2\pi S}}\rho^S\sum_{k=1}^{N-S} k\rho^k. \tag{7.29}$$

When $\beta \neq 0$ we use the formula (4.95), which was found in Section 4.9. Recall this formula

$$\sum_{k=1}^{M} k\rho^k = \frac{\rho^{M+1}}{\rho-1}\cdot M + \frac{1-\rho^M}{(1-\rho)^2}\cdot\rho.$$

So, using formula (4.95), conditions (i) and (ii) of the theorem and relation (7.25), one obtains

$$\frac{\rho^{N-S+1}}{\rho-1}(N-S) + \frac{1-\rho^{N-S}}{(1-\rho)^2}\rho \approx -\eta\sqrt{\frac{\lambda}{\mu}}\frac{e^{-\eta\beta}\sqrt{S}}{\beta} + \frac{\lambda(1-e^{-\eta\beta})}{\beta^2\mu}. \tag{7.30}$$

Taking into account equations (7.23) and (7.30) we have

$$\lim_{\lambda\to\infty} \sqrt{S}\zeta = \frac{\varphi(\beta)}{\mu\beta}\left[\frac{1-e^{-\eta\beta}}{\beta} - \eta e^{-\eta\beta}\right], \tag{7.31}$$

when $\beta \neq 0$. The case $\beta = 0$ means that $\rho = 1$ or $\rho \to 1$. If $\rho = 1$ it is easy to see that

$$\frac{1}{\lambda}\sum_{k=1}^{N-S} k\rho^k = \frac{1}{\lambda}\frac{(N-S)(N-S+1)}{2} \approx \frac{\eta^2}{2\mu}. \tag{7.32}$$

If $\rho \to 1$ then (7.32) is implied from (4.105). Thus, when $\beta = 0$ the approximation for $\sqrt{S}\zeta$ has the form

$$\lim_{\lambda\to\infty} \sqrt{S}\zeta = \frac{\eta^2}{2\mu\sqrt{2\pi}}. \tag{7.33}$$

Now, consider the conditional density function of the waiting time for M/M/S/N queue. When $\beta \neq 0$ using condition (i) of the theorem and equation (7.25), one obtains

$$\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) \approx \frac{\mu S(1-\rho)e^{-\mu S(1-\rho)t}}{1-\rho^{N-S}} \sum_{k=0}^{N-S-1} \frac{e^{-\sqrt{\lambda\mu}t}\left(\sqrt{\lambda\mu}t\right)^k}{k!}.$$

The last sum can be rewritten as $P(X_\lambda < N-S)$, where $X_\lambda \sim Pois(\sqrt{\lambda\mu}t)$. From the strong low of large numbers for the Poisson process we get

$$\lim_{\lambda\to\infty} \frac{X_\lambda}{\sqrt{\frac{\lambda}{\mu}}} = \mu t. \tag{7.34}$$

Thus,

$$\lim_{\lambda\to\infty} P(X_\lambda < N-S) = \lim_{\lambda\to\infty} P\left(\frac{X_\lambda}{\sqrt{\frac{\lambda}{\mu}}} < \eta\right) = \begin{cases} 1, & \mu t < \eta, \\ 0, & \mu t \geq \eta; \end{cases} \tag{7.35}$$

and approximation of the density function when $\beta \neq 0$ has the following form:

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \dfrac{\mu\beta e^{-\mu\beta t}}{(1-e^{-\eta\beta})}, & \mu t < \eta, \quad \beta \neq 0; \\ 0, & \mu t \geq \eta, \quad \beta \neq 0. \end{cases} \tag{7.36}$$

When $\beta = 0$ using conditions (i) and (ii) of the theorem and equation (7.35), one obtains

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \dfrac{\mu}{\eta}, & \mu t < \eta, \quad \beta = 0; \\ 0, & \mu t \geq \eta, \quad \beta = 0. \end{cases} \tag{7.37}$$

Combining (7.22), (7.26), (7.27), (7.28), (7.31), (7.33), (7.36) and (7.37), we proved Theorem 7.2.1. □

7.2.2 M/M/S/N queue as a particular case of a call center with an IVR

Suppose that the IVR processing time is negligible when compared to the time between arrivals. We capture this by letting $\theta \to \infty$. We also need to suppose that all of the customers wish to be served by an agent, i.e. $p = 1$. Show, that in this case the system with an IVR can be presented as M/M/S/N queue. For this purpose consider the M/M/S/N queue with parameters:

- the arrival rate equals λ;
- the service rate equals μ;
- the number of agents equals S;
- the number of trunk lines equals N.

Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{\frac{\lambda}{\mu}}} = \eta, \quad 0 < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad -\infty < \beta < \infty; \end{aligned} \tag{7.38}$$

and μ is fixed.

Consider also a call center with an IVR where

- the arrival rate equals λ;
- the IVR's service rate equals θ;
- the agent's service rate equals $\mu^* = \frac{\mu\theta}{\theta-\mu}$;
- the number of agents equals $S^* = S - \frac{\lambda}{\theta}$;
- the number of trunk lines equals N.

Thus, relations between the parameters of the system with an IVR are the following

$$\lim_{\lambda\to\infty} \frac{N - S^* - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{\frac{\lambda}{\mu}}\sqrt{\frac{\mu}{\theta}}} = \eta\sqrt{\tfrac{\theta}{\mu}}, \quad 0 < \eta < \infty;$$

$$\lim_{\lambda\to\infty}\frac{S^*-\dfrac{\lambda}{\mu^*}}{\sqrt{\dfrac{\lambda}{\mu^*}}}=\lim_{\lambda\to\infty}\frac{S-\dfrac{\lambda}{\theta}-\dfrac{\lambda}{\mu}+\dfrac{\lambda}{\theta}}{\sqrt{\dfrac{\lambda(\theta-\mu)}{\mu\theta}}}=\beta\sqrt{\frac{(\theta-\mu)}{\theta}},\quad -\infty<\beta<\infty;$$

where μ, p, θ are fixed. Denote

$$\eta^*=\eta\sqrt{\frac{\theta}{\mu}},\qquad \beta^*=\beta\sqrt{\frac{(\theta-\mu)}{\theta}}.$$

Let us find the approximations of the operational characteristics of the call center with an IVR with the above mentioned parameters:

• the probability $P(W>0)$ that a customer will wait after the IVR has the following asymptotic behavior (Theorems 4.3.1 and 4.3.2)

$$\lim_{\lambda\to\infty}P(W>0)=\begin{cases}\left(1+\dfrac{\beta^*\displaystyle\int_{-\infty}^{\beta^*}\Phi\left(\eta^*+(\beta^*-t)\sqrt{\dfrac{\theta}{\mu^*}}\right)d\Phi(t)}{\varphi(\beta^*)\Phi(\eta^*)-\varphi(\sqrt{\eta^{*2}+\beta^{*2}})\exp\dfrac{\eta_1^2}{2}\Phi(\eta_1)}\right)^{-1}, & \beta^*\neq 0;\\ \left(1+\dfrac{\displaystyle\int_{-\infty}^{0}\Phi\left(\eta^*-t\sqrt{\dfrac{\theta}{\mu^*}}\right)d\Phi(t)}{\sqrt{\dfrac{1}{2\pi}}\sqrt{\dfrac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right)}\right)^{-1}, & \beta^*=0;\end{cases}$$

where $\eta_1=\eta^*-\beta^*\sqrt{\frac{\mu^*}{\theta}}$.

•the probability of blocking has the following asymptotic behavior (Theorems 4.6.1 and 4.6.2)

$$\lim_{S\to\infty}\sqrt{S}P(block)=\begin{cases}when\quad \beta^*\neq 0:\\ \dfrac{\nu\varphi(\nu_1)\Phi(\nu_2)+\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\displaystyle\int_{-\infty}^{\beta^*}\Phi\left(\eta^*+(\beta^*-t)\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\frac{\varphi(\beta^*)\Phi(\eta^*)}{\beta^*}-\frac{\varphi(\sqrt{\eta^{*2}+\beta^{*2}})}{\beta^*}e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)};\\ when\quad \beta^*=0:\\ \dfrac{\nu\varphi(\nu_1)\Phi(\nu_2)+\frac{1}{\sqrt{2\pi}}\Phi(\eta^*)}{\displaystyle\int_{-\infty}^{0}\Phi\left(\eta^*-t\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right)};\end{cases}$$

where $\eta_1 = \eta^* - \beta^*\sqrt{\frac{\mu^*}{\theta}}$, $\quad \nu_1 = \frac{\eta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^*}{\sqrt{1+\frac{\mu^*}{\theta}}}$, $\quad \nu_2 = \frac{\beta^*\sqrt{\frac{\mu^*}{\theta}}-\eta^*}{\sqrt{1+\frac{\mu^*}{\theta}}}$, $\quad \nu = \frac{1}{\sqrt{1+\frac{\mu^*}{\theta}}}$.

• the expectation of the waiting time before the agents service (Theorems 4.9.1 and 4.9.2)

$$\lim_{S\to\infty}\sqrt{S}E[W] = \begin{cases} when \quad \beta^* \neq 0: \\ \dfrac{\frac{1}{\mu^*}\left(\frac{1}{\beta^*}\varphi(\beta^*)\Phi(\eta^*) + (\beta^*\frac{\mu^*}{\theta} - \frac{1}{\beta^*} - \eta^*\sqrt{\frac{\mu^*}{\theta}})\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)\right)}{\beta^*\int_{-\infty}^{\beta^*}\Phi(\eta^* + (\beta^* - t)\sqrt{\frac{\theta}{\mu^*}})d\Phi(t) + \varphi(\beta^*)\Phi(\eta^*) - \varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}; \\ when \quad \beta^* = 0: \\ \dfrac{1}{2\mu^*}\cdot\dfrac{\eta^{*2}\frac{\mu^*}{\theta}\Phi(\eta^*) + \eta^*\sqrt{\frac{\mu^*}{\theta}}\left(1+\sqrt{\frac{\mu^*}{\theta}}\right)\varphi(\eta^*)}{\sqrt{\frac{\mu^*}{\theta}}(\eta^*\Phi(\eta^*)+\varphi(\eta^*)) + \sqrt{2\pi}\int_{-\infty}^{0}\Phi(\eta^* - t\sqrt{\frac{\theta}{\mu^*}})d\Phi(t)}; \end{cases}$$

where $\eta_1 = \eta^* - \beta^*\sqrt{\frac{\mu^*}{\theta}}$.

• the conditional density function of the waiting time before the agents service (Theorems 4.9.1 and 4.9.2)

$$\lim_{S\to\infty}\sqrt{S}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \beta^*\mu^* e^{-\beta^*\mu^* t}\dfrac{\Phi(\eta^* - t\sqrt{\mu^*\theta})}{\Phi(\eta^*) - e^{\eta_2}\Phi(\eta_1)} & \beta^* \neq 0; \\ \dfrac{\mu^*\Phi(\eta^* - t\sqrt{\mu^*\theta})}{\eta^*\sqrt{\frac{\mu^*}{\theta}}\Phi(\eta^*) + \sqrt{\frac{\mu^*}{\theta}}\varphi(\eta^*)}, & \beta^* = 0; \end{cases}$$

where $\eta_1 = \eta^* - \beta^*\sqrt{\frac{\theta}{\mu^*}}$, and $\quad \eta_2 = \frac{1}{2}\frac{\mu^*}{\theta}\beta^{*2} - \eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}$.

As we said, we assume that the customers spend in the IVR time which is negligible when compared to the agent's processing time. Mathematically, this means that $\theta \to \infty$. Thus, one obtains that

$$\lim_{\theta\to\infty}\eta_1 = \lim_{\theta\to\infty}\eta^* - \beta^*\sqrt{\frac{\mu^*}{\theta}} = \lim_{\theta\to\infty}\eta\sqrt{\frac{\theta}{\mu}} - \beta\sqrt{\frac{\theta}{\theta-\mu}}\sqrt{\frac{\mu\theta}{\theta(\theta-\mu)}} = \infty; \quad (7.39)$$

$$\lim_{\theta\to\infty}\nu_1 = \lim_{\theta\to\infty}\frac{\eta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^*}{\sqrt{1+\frac{\mu^*}{\theta}}} = \lim_{\theta\to\infty}\frac{\eta\sqrt{\frac{\theta}{\mu}}\sqrt{\frac{\mu\theta}{\theta(\theta-\mu)}}+\beta\sqrt{\frac{\theta}{\theta-\mu}}}{\sqrt{1+\frac{\mu\theta}{\theta(\theta-\mu)}}} = \eta+\beta; \quad (7.40)$$

$$\lim_{\theta\to\infty}\nu_2 = \lim_{\theta\to\infty}\frac{\beta^*\sqrt{\frac{\mu^*}{\theta}}-\eta^*}{\sqrt{1+\frac{\mu^*}{\theta}}} = \lim_{\theta\to\infty}\frac{\beta\sqrt{\frac{\theta}{\theta-\mu}}\sqrt{\frac{\mu\theta}{\theta(\theta-\mu)}}-\eta\sqrt{\frac{\theta}{\mu}}}{\sqrt{1+\frac{\mu\theta}{\theta(\theta-\mu)}}} = -\infty; \quad (7.41)$$

$$\lim_{\theta\to\infty}\nu=\lim_{\theta\to\infty}\frac{1}{\sqrt{1+\frac{\mu^*}{\theta}}}=\lim_{\theta\to\infty}\frac{1}{\sqrt{1+\frac{\mu\theta}{\theta(\theta-\mu)}}}=1; \tag{7.42}$$

$$\lim_{\theta\to\infty}\eta_2=\lim_{\theta\to\infty}\frac{\mu^*\beta^{*2}}{2\theta}-\eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}=\lim_{\theta\to\infty}\frac{\beta^2\mu\theta^2}{\theta(\theta-\mu)^2}-\eta\beta\sqrt{\frac{\mu\theta^3}{\mu\theta(\theta-\mu)^2}}=-\eta\beta; \tag{7.43}$$

$$\lim_{\theta\to\infty}\eta^*=\lim_{\theta\to\infty}\eta\sqrt{\frac{\theta}{\mu}}=\infty; \tag{7.44}$$

$$\lim_{\theta\to\infty}\beta^*=\lim_{\theta\to\infty}\beta\sqrt{\frac{\theta}{\theta-\mu}}=\beta. \tag{7.45}$$

First we find the limit value of $\lim_{\lambda\to\infty}P(W>0)$ when $\theta\to\infty$ and $\beta\neq 0$:

$$\begin{aligned}
&\lim_{\theta\to\infty}\lim_{S\to\infty}P(W>0)= \\
&=\lim_{\theta\to\infty}\frac{\varphi(\beta^*)\Phi(\eta^*)-\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\beta^*\int\limits_{-\infty}^{\beta^*}\Phi\left(\eta^*+(\beta^*-t)\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\varphi(\beta^*)\Phi(\eta^*)-\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)} \\
&=\lim_{\theta\to\infty}\frac{\varphi(\beta^*)-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\eta^{*2}+\beta^{*2})}e^{\frac{1}{2}\eta^{*2}-\eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^2\frac{\mu^*}{\theta}}}{\beta^*\int\limits_{-\infty}^{\beta^*}d\Phi(t)+\varphi(\beta^*)-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\eta^{*2}+\beta^{*2})}e^{\frac{1}{2}\eta^{*2}-\eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^2\frac{\mu^*}{\theta}}} \\
&=\frac{\varphi(\beta)\left(1-e^{-\eta\beta}\right)}{\beta\Phi(\beta)+\varphi(\beta)\left(1-e^{-\eta\beta}\right)}
\end{aligned} \tag{7.46}$$

When $\beta=0$ one obtains:

$$\begin{aligned}
&\lim_{\theta\to\infty}\lim_{S\to\infty}P(W>0)=\lim_{\theta\to\infty}\frac{\sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right)}{\int\limits_{-\infty}^{0}\Phi\left(\eta^*-t\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right)} \\
&=\frac{\sqrt{\frac{1}{2\pi}}\eta}{\frac{1}{2}+\sqrt{\frac{1}{2\pi}}\eta}
\end{aligned} \tag{7.47}$$

Thus, the results coincide with the probability to wait in M/M/S/N queue. Now, consider the approximation for the probability to find the system busy when $\beta \neq 0$:

$$\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}P(block) = \lim_{\theta\to\infty}\frac{\nu\varphi\left(\nu_1\right)\Phi\left(\nu_2\right)+\varphi(\sqrt{\eta^{*2}+\beta^{*2}})\exp^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int\limits_{-\infty}^{\beta^*}\Phi\left(\eta^*+(\beta^*-t)\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\frac{\varphi(\beta^*)\Phi(\eta^*)}{\beta^*}-\frac{\varphi(\sqrt{\eta^{*2}+\beta^{*2}})}{\beta^*}\exp^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}$$

$$= \lim_{\theta\to\infty}\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\eta^{*2}+\beta^{*2})}e^{\frac{1}{2}\eta^{*2}-\eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^{*2}\frac{\mu^*}{\theta}}}{\Phi(\beta^*)+\frac{\varphi(\beta^*)}{\beta^*}-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\eta^{*2}+\beta^{*2})}e^{\frac{1}{2}\eta^2-\eta^*\beta^*\sqrt{\frac{\mu^*}{\theta}}+\beta^2\frac{\mu^*}{\theta}}}$$

$$= \lim_{\theta\to\infty}\frac{\varphi(\beta)e^{-\eta\beta}}{\beta\Phi(\beta)+\varphi(\beta)\left(1-e^{-\eta\beta}\right)} \tag{7.48}$$

When $\beta = 0$ we have

$$\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}P(block) = \lim_{\theta\to\infty}\frac{\nu\varphi\left(\nu_1\right)\Phi\left(\nu_2\right)+\frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int\limits_{-\infty}^{0}\Phi\left(\eta^*-t\sqrt{\frac{\theta}{\mu^*}}\right)d\Phi(t)+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right)}$$

$$= \frac{\sqrt{\frac{1}{2\pi}}}{\frac{1}{2}+\sqrt{\frac{1}{2\pi}}\eta} \tag{7.49}$$

We see, that the approximations for the probability to find the system busy also convert to the probability to find the system busy in M/M/S/N queue. Check the approximation for the expectation for the waiting time in the case $\beta \neq 0$:

$$\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}E[W] =$$

$$= \lim_{\theta\to\infty}\frac{\frac{1}{\mu^*}\left(\frac{1}{\beta^*}\varphi(\beta^*)\Phi(\eta^*)+(\beta^*\frac{\mu^*}{\theta}-\frac{1}{\beta^*}-\eta^*\sqrt{\frac{\mu^*}{\theta}})\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)\right)}{\beta^*\int_{-\infty}^{\beta^*}\Phi(\eta^*+(\beta^*-t)\sqrt{\frac{\theta}{\mu^*}})d\Phi(t)+\varphi(\beta^*)\Phi(\eta^*)-\varphi(\sqrt{\eta^{*2}+\beta^{*2}})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}$$

$$= \frac{\frac{\varphi(\beta)}{\beta}\left[\frac{1-e^{-\eta\beta}}{\beta\mu}-\eta e^{-\eta\beta}\right]}{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}. \tag{7.50}$$

When $\beta = 0$ we obtain:

$$\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}E[W] = \lim_{\theta\to\infty}\frac{\frac{1}{2\mu^*}\eta^{*2}\frac{\mu^*}{\theta}\Phi(\eta^*) + \eta^*\sqrt{\frac{\mu^*}{\theta}}\left(1+\sqrt{\frac{\mu^*}{\theta}}\right)\varphi(\eta^*)}{\sqrt{\frac{\mu^*}{\theta}}\left(\eta^*\Phi(\eta^*)+\varphi(\eta^*)\right) + \sqrt{2\pi}\int_{-\infty}^{0}\Phi(\eta^* - t\sqrt{\frac{\theta}{\mu^*}})d\Phi(t)}$$
$$= \frac{\eta^2}{2\mu(\eta+\sqrt{\frac{\pi}{2}})}. \tag{7.51}$$

Finally, show that the approximation for the conditional density function for the waiting time of the call center with an IVR also coincides with the approximation for the conditional density function for M/M/S/N. First, take a look at the case when $\beta \neq 0$:

$$\begin{aligned}\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) &= \lim_{\theta\to\infty}\beta^*\mu^*e^{-\beta^*\mu^*t}\frac{\Phi(\eta^* - t\sqrt{\mu^*\theta})}{\Phi(\eta^*) - e^{\eta_2}\Phi(\eta_1)} \\ &= \lim_{\theta\to\infty}\beta\sqrt{\frac{\theta}{\theta-\mu}}\frac{\mu\theta}{\theta-\mu}e^{-\beta\sqrt{\frac{\theta}{\theta-\mu}}\frac{\mu\theta}{\theta-\mu}t}\frac{\Phi\left(\eta\sqrt{\frac{\theta}{\mu}} - t\sqrt{\frac{\mu\theta^2}{\theta-\mu}}\right)}{\Phi(\eta\sqrt{\frac{\theta}{\mu}}) - e^{\eta_2}\Phi(\eta_1)} \\ &= \begin{cases}\frac{\beta\mu e^{-\beta\mu t}}{1-e^{-\eta\beta}}, & t < \frac{\eta}{\mu}; \\ 0, & t > \frac{\eta}{\mu}.\end{cases}\end{aligned} \tag{7.52}$$

When $\beta = 0$ we have:

$$\begin{aligned}\lim_{\theta\to\infty}\lim_{S\to\infty}\sqrt{S}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) &= \lim_{\theta\to\infty}\frac{\mu^*\Phi(\eta^* - t\sqrt{\mu^*\theta})}{\eta^*\sqrt{\frac{\mu^*}{\theta}}\Phi(\eta^*) + \sqrt{\frac{\mu^*}{\theta}}\varphi(\eta^*)} \\ &= \lim_{\theta\to\infty}\frac{\frac{\mu\theta}{\theta-\mu}\Phi\left(\eta\sqrt{\frac{\theta}{\mu}} - t\sqrt{\frac{\mu\theta^2}{\theta-\mu}}\right)}{\eta\sqrt{\frac{\theta}{\mu}}\sqrt{\frac{\mu\theta}{\theta(\theta-\mu)}}\Phi\left(\eta\sqrt{\frac{\theta}{\mu}}\right) - \sqrt{\frac{\mu\theta}{\theta(\theta-\mu)}}\varphi\left(\eta\sqrt{\frac{\theta}{\mu}}\right)} \\ &= \begin{cases}\frac{\mu}{\eta}, & t < \frac{\eta}{\mu}; \\ 0, & t > \frac{\eta}{\mu}.\end{cases}\end{aligned} \tag{7.53}$$

Thus, we emphasize that the same result has been obtained for M/M/S/N queue. In conclusion, one can say that when the IVR processing time is negligible when compared to the talking time of the agents, a call center with an IVR may be modelled as the M/M/S/N queue.

7.3 The M/M/S system (Erlang-C)

Note, that the M/M/S queue is an external case of M/M/S/N queue, which is obtained when $N \to \infty$, i.e. there are infinitely many places in the system. Thus, to our previous assumption, that IVR processing time is negligible when compared to the talking time of the agents, i.e. $\theta \to \infty$, we add that the number of trunk lines N, tends to infinity, i.e. $\eta \to \infty$. This case was discovered in Sections 4.5 and 4.11 with a name Case c. It was shown that when $\beta > 0$

• the approximation of the probability to wait has the form (see (4.66)):

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty} P(W>0) = \left(1+\frac{\Phi(\beta)\beta}{\varphi(\beta)}\right)^{-1}.$$

• the approximation of the conditional expectation of the waiting time is following (see (4.113)):

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty} \sqrt{S}E[W|W>0] = \frac{1}{\mu\beta}.$$

• the approximation of the conditional density function of the waiting time it is easy to obtain from (7.52) by letting $\eta \to \infty$:

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty} \sqrt{S} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \beta\mu e^{-\beta\mu t}.$$

These results coincide with the approximations of Halfin and Whitt [17] for M/M/S system. Therefore, one can say that when the IVR processing time is negligible when compared to the talking time of the agents and there are infinity number of trunk lines, the system with an IVR may be modelled as the M/M/S queue.

7.4 The M/M/S/∞/N system

The $M/M/S/\infty/N$ system was represented and analyzed by Vericourt and Jennings in [22]. This system is a particular case of our model, when $\lambda \to \infty$. By the M/M/S/∞/N system it is possible to model a call center with an IVR, when there are exactly N customers in the system. This means that there is no possibility to leave the system after service in the IVR, i.e. $p = 1$, and customers leave the system after agent's service.

Actually, a call center with an IVR is not a good example for the $M/M/S/\infty/N$ system. This model is more efficient for describing hospital problem or machine breakdown problem (see [22]).

In order to illustrate our model as the $M/M/S/\infty/N$ system let us define the states of our system when there is exactly N customers in the system and no possibility to leave the system after IVR. The states will take the following form: $(N-j, j)$, where $0 \le j \le N-S$. Thus, the stationary probabilities will be the following:

$$\pi(N-j,j) = \begin{cases} \pi_0 \dfrac{1}{(N-j)!} \left(\dfrac{\lambda}{\theta}\right)^{N-j} \dfrac{1}{j!} \left(\dfrac{\lambda}{\mu}\right)^{j}, & j \le S; \\ \pi_0 \dfrac{1}{(N-j)!} \left(\dfrac{\lambda}{\theta}\right)^{N-j} \dfrac{1}{S!S^{j-S}} \left(\dfrac{\lambda}{\mu}\right)^{j} & j \ge S; \\ 0 & \text{otherwise,} \end{cases} \tag{7.54}$$

where

$$\pi_0 = \left(\sum_{j=S}^{N} \frac{1}{N!} \left(\frac{\lambda}{\theta}\right)^{N-j} \frac{1}{S!S^{j-S}} \left(\frac{\lambda}{\mu}\right)^{j} + \sum_{j=0}^{S} \frac{1}{(N-j)!} \left(\frac{\lambda}{\theta}\right)^{N-j} \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^{j} \right)^{-1} \tag{7.55}$$

We can also write the stationary probabilities in the equivalent form:

$$\pi(N-j,j) = \begin{cases} \widetilde{\pi}_0 \begin{pmatrix} N \\ j \end{pmatrix} \left(\dfrac{1}{\theta}\right)^{N-j} \left(\dfrac{1}{\mu}\right)^{j}, & j \le S; \\ \widetilde{\pi}_0 \begin{pmatrix} N \\ j \end{pmatrix} \left(\dfrac{1}{\theta}\right)^{N-j} \dfrac{j!}{S!S^{j-S}} \left(\dfrac{1}{\mu}\right)^{j} & j \ge S; \\ 0 & \text{otherwise,} \end{cases} \tag{7.56}$$

where

$$\widetilde{\pi}_0 = \left(\left(\frac{1}{\theta} + \frac{1}{\mu}\right)^{N-j} + \sum_{j=S+1}^{N} \frac{N!}{(N-j)!} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) \left(\frac{1}{\theta}\right)^{N-j} \left(\frac{1}{\mu}\right)^{j} \right)^{-1}. \tag{7.57}$$

The equations (7.56) and (7.57) have the same form as stationary probabilities in [22], and thus all results for $M/M/S/\infty/N$ system are contained in this particular case of our model.

Note, however, that our asymptotic analysis does not cover that in [22], since in our case $\lambda \to \infty$ together with the quantities.

Chapter 8

An algorithm for finding the optimal staffing and trunk level

This chapter is devoted to an algorithm for solution of an optimization problem. The goal is to find the number of trunk lines and agents for a Call Center with an IVR, so that the cost is minimal, but at the same time some constraints are satisfied, for example the probability to wait $P(W > 0) < a$, or the probability of blocking $P(block) < b$. Practically, $P(block) < 0.05$ and $P(W > 0) < 0.1$ mean that at most 5% of the customers at most do not enter the system and up to 10% of the customers wait in queue. Our objective function is taken to be based on statistical estimator of the costs in Call Centers.

8.1 Formulating an optimization problem

[1]The economic performance is affected by costs and revenues. The operational costs of a typical Call Center consist of different components, as shown below:

- Salaries - 63%
- Hiring and training costs - 6%
- Costs for office space - 5%
- Trunk costs - 5%
- IT and telecommunication equipment - 10%

Short term operational planning affects the costs for the agents and telephone costs. As described above, the hourly cost of the agents is the main component. The costs for telephone trunks can be split up into two parts: fixed flat costs per trunk and variable usage costs for each trunk (telephone costs). Among different

[1]Parts of Section 8.1 are adapted from [28].

telephone service numbers we distinguish toll free services, shared cost services, and value added services.

To derive the cost and revenue functions, we use the following notation for a call center with homogeneous agents and customers:

C^A = cost of an agent per time unit,
C^u = telephone cost per trunk and time unit,
S = number of staffed agents,
N = number of telephone trunks,
$\mathbf{E}[u^N]$ = expected trunk utilization,
$\mathbf{E}[L]$ = expected queue length,
$\mathbf{E}[u]$ = expected utilization of the agents.

We describe the cost functions for free services, shared cost services, and value added services below.

(i) Toll free services:

A call center can provide services with toll-free numbers, for example using 1-800-phone numbers in Israel. In this case, the call is free for the customer, and the call canter pays the telephone cost C^u per time unit foe customers on hold and in service. Therefore, the costs per time unit are

$$C(S,N) = C^A S + N\mathbf{E}[u^N]C^u = C^A S + (\mathbf{E}[L] + S\mathbf{E}[u])C^u \tag{8.1}$$

The average number of occupied trunks is $N\mathbf{E}[u^N]$ and can be expressed as the sum of the average number of waiting customers $\mathbf{E}[L]$ and the average number of customers in service $S\mathbf{E}[u]$.

(ii) Shared cost services:

In some Call Centers a calling customer pays a price to the telecommunication provider dependent on the length of the call or a fixed price per call.

If a customer has to bear a part of the telephone cost per time unit, the call center pays C^u per time unit of waiting and talk time, depending on the telecommunication provider and the used number. In these cases costs per time unit are given by (8.1). In fact, the structure of the cost function is identical to that in the case (i), but usually the cost C^u for a call providing shared cost service is lower then the cost C^u of the toll free service.

(iii) Value added services:

Sometimes, call centers can provide value added services, so a calling customer bears the whole phone cost only, i.e.

$$C(S,N) = C^A S$$

We consider the cost function (ii) and try to find the optimal number of trunk

lines N and the number of agent S so as to pay minimal cost under the following conditions:

- the probability to wait will be less than a
- the probability of blocking will be less than b.

Note that the values a and b are any value between 0 and 1. So we have the next problem:

$$\begin{gathered}\min_{S,N} C(S,N) = C^A S + NE(u^N)C^U; \\ subject\ to \\ P(W>0) < a; \\ P(block) < b; \\ where\ S\ and\ N\ are\ integer\ and\ 0 \le S \le N;\end{gathered} \tag{8.2}$$

where

$$P(W>0) = 1 - \sum_{k=1}^{N} \sum_{j=0}^{min(k,S)-1} \chi(k,j);$$

$$P(block) = \pi_0 \left(\frac{\lambda^N}{N!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^N + \sum_{j=S+1}^{N} \frac{1}{(N-j)!} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{N-j} \left(\frac{p\lambda}{\mu} \right)^j \right);$$

here $\chi(k,j)$ is from formula (4) and π_0 is from formula (2).

8.2 Performance measures

Let us consider in what way the changing in the number of trunk lines N and the number of agents S influence the behaviour of the probability to wait $P(W>0)$ and the probability to find the system busy $P(block)$. Note, that

- decrease in the number of trunk lines causes decrease of the probability to wait $P(W>0)$ and growth of the probability to find the system busy $P(block)$;
- growth of the number of trunk lines causes growth of the probability to wait $P(W>0)$ and decrease of the probability to find the system busy $P(block)$;
- decrease in the number of agents causes growth of the probability to wait $P(W>0)$ and growth of the probability to find the system busy $P(block)$;
- growth of the number of trunk lines causes decrease of the probability to wait $P(W>0)$ and decrease of the probability to find the system busy $P(block)$;

As confirmation of these facts, we wrote programs for calculating the probability to wait and the probability to find the system busy (see Appendix A) and consider as an example the case when

- the number of agents S is between 20 and 30;

- the number of trunk lines N is between 20 and 60;
- the arrival rate λ is equal to 20;
- the probability to be served by an agent p is equal to 1;
- the agent's service rate μ is equal to 1;
- the IVR's service rate θ is equal to 1;
- the problem's restrictions are:

$$P(W > 0) < 0.2;$$
$$P(block) < 0.05.$$

The results are in Table 8.1 and Table 8.2 below. In these tables yellow cells correspond to appropriative values of N and S and red cells correspond to optimal solutions.

N\S	S=20	S=21	S=22	S=23	S=24	S=25	S=26	S=27	S=28	S=29	S=30
20	0										
21	4.97E-07										
22	5.49E-06	2.4E-07									
23	3.18E-05	2.7E-06	1.13E-07								
24	0.000129	1.7E-05	1.36E-06	5.4E-08							
25	0.000409	6.9E-05	8.52E-06	6.7E-07	2.6E-08						
26	0.001087	0.00023	3.72E-05	4.4E-06	3.3E-07	1.2E-08					
27	0.002509	0.00063	0.000127	2E-05	2.2E-06	1.6E-07	5.7E-09				
28	0.005173	0.0015	0.000359	7E-05	1E-05	1.1E-06	8E-08	2.7E-09			
29	0.009714	0.00318	0.000883	0.0002	3.8E-05	5.5E-06	5.7E-07	3.9E-08	1.3E-09		
30	0.016859	0.00614	0.001931	0.00051	0.00011	2E-05	2.8E-06	2.9E-07	1.9E-08	6E-10	
31	0.027355	0.01095	0.003832	0.00116	0.0003	6.3E-05	1.1E-05	1.5E-06	1.4E-07	9E-09	3E-10
32	0.041879	0.01821	0.007004	0.00236	0.00069	0.00017	3.5E-05	5.8E-06	7.5E-07	7E-08	4E-09
33	0.060949	0.02852	0.011931	0.00442	0.00143	0.0004	9.6E-05	1.9E-05	3.1E-06	4E-07	4E-08
34	0.084849	0.04241	0.019113	0.0077	0.00275	0.00086	0.00023	5.4E-05	1E-05	2E-06	2E-07
35	0.11358	0.06024	0.029015	0.0126	0.0049	0.00169	0.00051	0.00013	3E-05	6E-06	8E-07
36	0.146843	0.08214	0.042004	0.01951	0.00818	0.00307	0.00102	0.0003	7.5E-05	2E-05	3E-06
37	0.184066	0.10802	0.058297	0.02877	0.01291	0.00522	0.00189	0.00061	0.00017	4E-05	9E-06
38	0.224445	0.13754	0.077922	0.04062	0.01937	0.0084	0.00328	0.00115	0.00036	1E-04	2E-05
39	0.267022	0.17014	0.1007	0.05513	0.02778	0.01281	0.00537	0.00203	0.00069	0.0002	6E-05
40	0.310764	0.20509	0.126258	0.07225	0.03826	0.01865	0.00832	0.00338	0.00124	0.0004	0.0001
41	0.354651	0.24154	0.15405	0.09173	0.0508	0.02604	0.01229	0.00531	0.00209	0.0007	0.0002
42	0.397748	0.27863	0.183414	0.11319	0.06525	0.035	0.01738	0.00795	0.00333	0.0013	0.0004
43	0.439259	0.31551	0.213622	0.1361	0.08135	0.04544	0.02362	0.01137	0.00504	0.002	0.0008
44	0.47857	0.35144	0.243945	0.15987	0.09869	0.05718	0.03098	0.01562	0.00729	0.0031	0.0012
45	0.515255	0.38579	0.273707	0.18391	0.11681	0.06993	0.03931	0.02066	0.01011	0.0046	0.0019
46	0.549074	0.41811	0.302329	0.20761	0.13522	0.08332	0.04841	0.02641	0.01348	0.0064	0.0028
47	0.579947	0.4481	0.329355	0.23047	0.15342	0.09697	0.05801	0.03273	0.01735	0.0086	0.004
48	0.607922	0.47561	0.354466	0.25205	0.17098	0.11048	0.06781	0.03941	0.02161	0.0111	0.0054
49	0.633144	0.50061	0.377477	0.27207	0.18753	0.12349	0.07752	0.04624	0.02612	0.0139	0.007
50	0.655817	0.52317	0.398322	0.29032	0.20281	0.13571	0.08684	0.05299	0.03074	0.0169	0.0088
51	0.676177	0.54344	0.417029	0.30673	0.21663	0.14692	0.09555	0.05946	0.0353	0.0199	0.0106
52	0.694468	0.56161	0.433698	0.3213	0.22893	0.15697	0.10349	0.06547	0.03964	0.0229	0.0126
53	0.71093	0.57787	0.448477	0.33411	0.23971	0.16582	0.11054	0.0709	0.04366	0.0257	0.0145
54	0.725782	0.59243	0.461538	0.34529	0.24904	0.17346	0.11667	0.07568	0.04726	0.0283	0.0163
55	0.739226	0.60549	0.47306	0.35499	0.25703	0.17996	0.12188	0.07979	0.0504	0.0306	0.0179
56	0.751438	0.61722	0.483221	0.36336	0.26381	0.18541	0.12624	0.08323	0.05306	0.0326	0.0193
57	0.762569	0.6278	0.492186	0.37057	0.26952	0.18993	0.12983	0.08606	0.05527	0.0343	0.0206
58	0.772753	0.63736	0.500106	0.37678	0.27431	0.19365	0.13274	0.08835	0.05706	0.0357	0.0216
59	0.782103	0.64603	0.507115	0.38211	0.27831	0.19668	0.13507	0.09016	0.05847	0.0368	0.0224
60	0.790716	0.65392	0.513328	0.38669	0.28164	0.19913	0.13692	0.09159	0.05958	0.0377	0.0231

Table 8.1: $P(W > 0)$ when $20 \leq S \leq 30$, $20 \leq N \leq 60$, $\lambda = 20$, $p = 1$, $\mu = 1$, $\theta = 1$.

N\S	S=20	S=21	S=22	S=23	S=24	S=25	S=26	S=27	S=28	S=29	S=30
20	0.52131										
21	0.49824	0.4982									
22	0.47531	0.4753	0.4753								
23	0.45256	0.4525	0.4525	0.4525							
24	0.43001	0.43	0.43	0.43	0.43						
25	0.40775	0.4076	0.4076	0.4076	0.4076	0.4076					
26	0.38588	0.3855	0.3854	0.3854	0.3854	0.3854	0.38538				
27	0.36456	0.3637	0.3635	0.3634	0.3634	0.3634	0.36343	0.3634			
28	0.34401	0.3424	0.3419	0.3418	0.3418	0.3418	0.34175	0.3418	0.3418		
29	0.32445	0.3217	0.3208	0.3205	0.3204	0.3204	0.32037	0.3204	0.3204	0.3204	
30	0.30612	0.3019	0.3001	0.2995	0.2994	0.2993	0.29931	0.2993	0.2993	0.2993	0.2993
31	0.28918	0.283	0.2802	0.2791	0.2787	0.2786	0.27861	0.2786	0.2786	0.2786	0.2786
32	0.2737	0.2652	0.2611	0.2593	0.2586	0.2584	0.25832	0.2583	0.2583	0.2583	0.2583
33	0.25965	0.2488	0.2429	0.2401	0.239	0.2386	0.23848	0.2384	0.2384	0.2384	0.2384
34	0.24684	0.2336	0.2259	0.2219	0.2201	0.2194	0.21916	0.2191	0.2191	0.2191	0.2191
35	0.23501	0.2196	0.21	0.2046	0.202	0.2009	0.20043	0.2003	0.2002	0.2002	0.2002
36	0.22381	0.2067	0.1952	0.1884	0.1848	0.1831	0.18237	0.1821	0.182	0.182	0.182
37	0.21289	0.1946	0.1816	0.1733	0.1686	0.1662	0.16507	0.1646	0.1645	0.1644	0.1644
38	0.20191	0.183	0.1688	0.1592	0.1534	0.1502	0.14861	0.1479	0.1477	0.1476	0.1475
39	0.19065	0.1717	0.1568	0.1462	0.1393	0.1352	0.1331	0.1321	0.1317	0.1315	0.1314
40	0.17899	0.1605	0.1453	0.1339	0.1262	0.1213	0.11858	0.1172	0.1166	0.1163	0.1162
41	0.16692	0.1492	0.1342	0.1224	0.114	0.1084	0.10511	0.1033	0.1024	0.102	0.1019
42	0.15458	0.1378	0.1233	0.1115	0.1027	0.0966	0.0927	0.0905	0.0893	0.0887	0.0885
43	0.14216	0.1264	0.1125	0.101	0.0921	0.0856	0.08135	0.0787	0.0773	0.0765	0.0762
44	0.12993	0.115	0.1019	0.0909	0.0822	0.0756	0.071	0.0681	0.0663	0.0654	0.0649
45	0.11815	0.1038	0.0916	0.0812	0.0728	0.0663	0.06162	0.0585	0.0565	0.0554	0.0548
46	0.10706	0.0931	0.0816	0.072	0.0641	0.0578	0.05313	0.0499	0.0477	0.0464	0.0457
47	0.09684	0.083	0.0721	0.0632	0.0558	0.05	0.04547	0.0422	0.04	0.0386	0.0378
48	0.0876	0.0737	0.0632	0.0549	0.0482	0.0428	0.03859	0.0354	0.0332	0.0318	0.0309
49	0.07941	0.0653	0.0551	0.0473	0.0412	0.0363	0.03244	0.0295	0.0274	0.0259	0.025
50	0.07226	0.0578	0.0477	0.0404	0.0348	0.0304	0.02698	0.0243	0.0223	0.0209	0.02
51	0.06608	0.0512	0.0412	0.0342	0.0291	0.0252	0.02219	0.0198	0.018	0.0167	0.0158
52	0.06079	0.0455	0.0355	0.0288	0.0241	0.0207	0.01803	0.016	0.0144	0.0132	0.0124
53	0.05629	0.0406	0.0306	0.0241	0.0198	0.0168	0.01448	0.0127	0.0114	0.0103	0.0096
54	0.05246	0.0364	0.0264	0.0202	0.0162	0.0134	0.01148	0.01	0.0089	0.008	0.0073
55	0.04919	0.0328	0.0228	0.0168	0.0131	0.0107	0.009	0.0078	0.0068	0.0061	0.0055
56	0.04639	0.0297	0.0198	0.014	0.0106	0.0084	0.00698	0.006	0.0052	0.0046	0.0041
57	0.04396	0.0271	0.0173	0.0117	0.0085	0.0066	0.00536	0.0045	0.0039	0.0034	0.003
58	0.04185	0.0249	0.0152	0.0098	0.0069	0.0051	0.00408	0.0034	0.0029	0.0025	0.0022
59	0.03998	0.0229	0.0134	0.0083	0.0055	0.004	0.00308	0.0025	0.0021	0.0018	0.0016
60	0.03832	0.0212	0.0119	0.007	0.0045	0.0031	0.00232	0.0018	0.0015	0.0013	0.0011

Table 8.2: $P(block)$ when $20 \leq S \leq 30$, $20 \leq N \leq 60$, $\lambda = 20$, $p = 1$, $\mu = 1$, $\theta = 1$.

We can see that the level line of the function $P(W > 0) = F_1(S, N)$ determines a monotone increasing function $S = f_1(N)$, and the level line of the function $P(block) = F_2(S, N)$ determines monotone decreasing function $S = f_2(N)$. The area of admissible values of S and N for problem (8.2) is shaded area on the figure below. The number of solutions is infinite. We would like to consider the case, when the staffing cost much more than in other cases. It means that the optimal point is when S is the smallest.

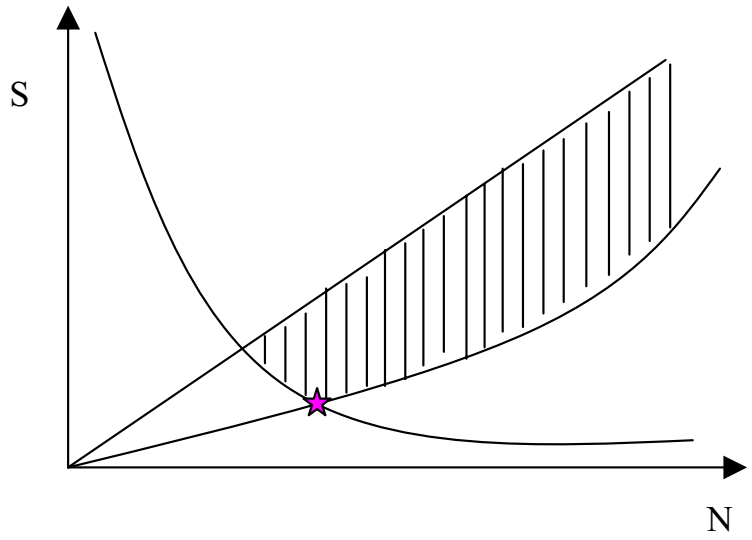

Figure 8.1: The domain of pairs (S, N) which satisfy the problem (8.2) and the optimal solution of this problem.

So, we start with the state when the probability to wait $P(W > 0)$ is smaller than a and the probability to find the system busy $P(block)$ is bigger than b. Further, by adding trunk lines we increase $P(W > 0)$ up to the desired level and thus check $P(block)$. If it is bigger than the desired value, we add one more agent and continue the process. Otherwise, we increase $P(block)$ up to the value b by subtracting trunk lines and stop. In this way, we receive the algorithm which is formulated in the next section.

8.3 Algorithm

As we have said, after the analysis of the performance measure's behavior, we can build a method for calculation of an optimal (S, N) pair under our problem's

constraints. But first, we need to define the initial data for our algorithm. We want to increase the probability to wait, so the best way to start is to put $N_0 = S_0$, because in this case $P(W > 0) = 0$. Also we want to decrease the probability to find the system busy. It seems that we can take any number of agents which implements the condition $P(block) \geq b$. There are many such numbers and we want the one that helps us to reduce the quantity of iteration in our algorithm. It is easy to see that in these cases $P(block) \geq b$. So, which initial number of agents do we need? As we see in Section 4.8 when the number of trunk lines goes to infinity and $\beta < 0$ the probability to find the system busy tends to $-\beta$. Moreover, when the number of trunk lines is growing the probability to find the system busy is growing as well. Thus, we can say that $\sqrt{S}P(block) \geq -\beta$. A more detailed graphical confirmation of this fact we will be given in Chapter 9. Now using this condition we get:

$$P(block) \approx b \;\Rightarrow\; \sqrt{S}\, b \approx -\beta \;\Rightarrow\; b \geq -\frac{\beta}{\sqrt{S}} \approx -(1 - \frac{\lambda p}{\mu S}) \;\Rightarrow\; S \approx \frac{\lambda p}{\mu(1+b)}.$$

So, our initial agent's number approximately equals

$$S_0 = \left[\frac{\lambda p}{\mu(1+b)}\right]$$

Now we can formulate the algorithm for finding the optimal (S, N) pair under our problem's constraints.

- Step 1. Let $S = [\lambda p/\mu(1+b)]$, $N = S$ and go to Step 2.
- Step 2. Add trunk lines till $P(W > 0) < a$, and then go to Step 3.
- Step 3. If $P(block) > b$ add one agent and go to Step 2, else go to Step 4.
- Step 4. If $P(block) < b$ subtract trunk lines till $P(block) < b$, else stop.

Chapter 9

Recommendations to a manager of a call center with an IVR

In previous chapters we were analyzing a call center with an IVR. We found approximations for its performance measures and performed graphical analysis of these measures. Now we would like to answer the following question :"What can we recommend to a manager of such a call center?" A central goal of each call center's manager is to establish an appropriate balance between cost and service level. In this chapter we demonstrate the methods that can help reach sound decisions. We do this in the following way. First, we recall the operational performance measures that one can calculate. Next, we analyze some ways to reduce the cost of a call center. Then, we investigate the effect of changes in parameters of the system on the optimal solution of the problem, which was analyzed in the previous chapter. We also consider the effect of a call center's size on its service level. In addition, we demonstrate the behaviour of performance measures when the system's parameters are changing.

9.1 Calculating operational performance measures

In principle, we know how to calculate exact and approximate performance measures. However, the exact calculation of these measures can take a long time because of two reasons: complicated expressions and numerical instability. Moreover, in Chapter 5 it was shown that the approximating values are very close to the exact values. So, the use of the approximations is an easy and convenient way to calculate performance measures of call centers with an IVR.

Let us recall the operational performance measures that we have been calcu-

lating. Suppose we know the values of the following parameters:

- λ - the average arrival rate;
- θ - the average rate of the service in IVR;
- μ - the average rate of the service by an agent;
- p - the probability that a customer would like to receive the service by an agent;
- S - the number of agents;
- N - the number of trunk lines.

Then it is possible to find the exact and approximate values of the following operational characteristics:

• **the probability to find the system busy**:

exact calculation (3.15):

$$P(block) = \frac{\frac{\lambda^N}{N!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^N + \sum\limits_{j=S+1}^{N}\frac{1}{(N-j)!}\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{p\lambda}{\mu}\right)^j}{\sum\limits_{i=0}^{N-S}\sum\limits_{j=S}^{N-i}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j + \sum\limits_{i+j\le N, j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j};$$

approximate calculation (Theorems 4.6.1 and 4.6.2):

$$\lim_{S\to\infty}\sqrt{S}P(block) = \begin{cases} \dfrac{\nu\varphi(\nu_1)\Phi(\nu_2) + \varphi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int\limits_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t) + \frac{\varphi(\beta)\Phi(\eta)}{\beta} - \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}, & \beta\neq 0 \\ \dfrac{\nu\varphi(\nu_1)\Phi(\nu_2) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int\limits_{-\infty}^{0}\Phi\left(\eta-t\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(t) + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right)}, & \beta=0, \end{cases}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu}{p\theta}}$, $\quad \nu_1 = \frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}$, $\quad \nu_2 = \frac{\beta\sqrt{\frac{\mu}{p\theta}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}$, $\quad \nu = \frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}$.

• **the probability to wait before the agent's service:**

exact calculation (3.12)

$$P(W>0) = \left(1 + \frac{\sum\limits_{i+j\le N-1, j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j}{\sum\limits_{i=0}^{N-S-1}\sum\limits_{j=S}^{N-i-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j}\right)^{-1};$$

approximate calculation (Theorems 4.3.1 and 4.3.2):

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1+\dfrac{\beta\int\limits_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\dfrac{p\theta}{\mu}}\right)d\Phi(t)}{\varphi(\beta)\Phi(\eta)-\varphi(\sqrt{\eta^2+\beta^2})\exp\dfrac{\eta_1^2}{2}\Phi(\eta_1)}\right)^{-1}, & \beta\neq 0 \\ \left(1+\dfrac{\int\limits_{-\infty}^{0}\Phi\left(\eta-t\sqrt{\dfrac{p\theta}{\mu}}\right)d\Phi(t)}{\sqrt{\dfrac{1}{2\pi}}\sqrt{\dfrac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right)}\right)^{-1}, & \beta=0, \end{cases}$$

where $\eta_1=\eta-\beta\sqrt{\frac{\mu}{p\theta}}$.

• the probability to wait less than t units of time before the agent's service

exact calculation (3.10):

$$P(W\le t)=1-\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)\sum_{l=0}^{j-S}\frac{(\mu St)^l e^{-\mu St}}{l!}. \tag{9.1}$$

approximate calculation (Corollaries 6.2.1 and 6.2.2):

$$\lim_{\lambda\to\infty} P(W\le\frac{t}{\sqrt{S}}) = \begin{cases} \int\limits_0^t \beta\mu e^{-\beta\mu u}\frac{\Phi(\eta-u\sqrt{p\mu\theta})}{\Phi(\eta)-e^{\eta_2}\Phi(\eta_1)}du, & \beta\neq 0 \\ \int\limits_0^t \frac{\mu\Phi(\eta-u\sqrt{p\mu\theta})}{\eta\sqrt{\frac{\mu}{p\theta}}\Phi(\eta)+\sqrt{\frac{\mu}{p\theta}}\varphi(\eta)}du, & \beta=0, \end{cases}$$

where $\eta_1=\eta-\beta\sqrt{\frac{p\theta}{\mu}}$, and $\eta_2=\frac{1}{2}\frac{\mu}{p\theta}\beta^2-\eta\beta\sqrt{\frac{\mu}{p\theta}}$.

• the average speed of answer (ASA) by agents to the customers who would like to be served by an agent

exact calculation (3.11):

$$E[W]=\frac{\frac{1}{\mu S}\sum\limits_{m=0}^{N-S-1}\sum\limits_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j(j-S)}{\sum\limits_{m+j\le N-1,j<S}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j+\sum\limits_{m=0}^{N-S-1}\sum\limits_{j=S}^{N-m-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j};$$

approximate calculation (Theorems 4.9.1 and 4.9.2):

$$\lim_{S\to\infty}\sqrt{S}E[W]=\begin{cases}\dfrac{\frac{1}{\mu}\left(\frac{1}{\beta}\varphi(\beta)\Phi(\eta)+(\beta\frac{\mu}{p\theta}-\frac{1}{\beta}-\eta\sqrt{\frac{\mu}{p\theta}})\varphi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)\right)}{\beta\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}})d\Phi(t)+\varphi(\beta)\Phi(\eta)-\varphi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}, & \beta\neq 0\\ \dfrac{\eta^2\frac{\mu}{p\theta}\Phi(\eta)+\eta\sqrt{\frac{\mu}{p\theta}}\left(1+\sqrt{\frac{\mu}{p\theta}}\right)\varphi(\eta)}{\frac{1}{2\mu}\cdot\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right)+\sqrt{2\pi}\int_{-\infty}^{0}\Phi(\eta-t\sqrt{\frac{p\theta}{\mu}})d\Phi(t)}, & \beta=0,\end{cases}$$

where $\eta_1=\eta-\beta\sqrt{\frac{\mu}{p\theta}}$.

• **the number of customers in the queue:**

$$L_q=p\lambda\left(1-P(block)\right)E[W],$$

where $P(block)$ and $E[W]$ can be calculated by the formulae which were derived before; see (3.15) and (3.11) in the case of exact calculation and Theorems 4.6.1 and 4.6.2 and Theorems 4.9.1 and 4.9.2 in the case of approximate calculation.

• **occupancy of the agents:**

$$u=\frac{p\lambda\left(1-P(block)\right)}{S\mu},$$

where $P(block)$ can be calculated by (3.15) in the case of exact calculation and by formulae from Theorems 4.6.1 and 4.6.2 in the case of approximate calculation.

9.2 Analyzing performance measures as functions of β, η and $\frac{p\theta}{\mu}$

Let us rewrite the problem for finding the optimal solution, that was formulated in Section 8.1 as follows:

$$\begin{aligned}&\min_{\beta,\eta}C(S(\beta),N(\beta,\eta));\\ &\quad subject\ to\\ &\quad P(W>0)\le a;\\ &\sqrt{S}P(block)\le b;\\ &\quad -\infty<\beta,\eta<\infty,\end{aligned}\tag{9.2}$$

As we can see from the approximations for performance measures the probability to wait and the probability to find the system busy can be presented as functions on β, η and $\frac{p\theta}{\mu}$ (see Theorems 4.3.2-4.6.2). So, restriction (9.2) can be rewritten as follows

$$\begin{gathered} \min_{\beta,\eta} C(\beta,\eta,c); \\ subject\ \ to \\ f_1(\beta,\eta,c) \leq a; \\ f_2(\beta,\eta,c) \leq b; \\ -\infty < \beta,\eta < \infty,\ \ 0 < c < \infty, \end{gathered} \tag{9.3}$$

where

$$\lim_{\lambda\to\infty} P(W>0) = f_1(\beta,\eta,c), \qquad \lim_{\lambda\to\infty} \sqrt{S}P(block) = f_2(\beta,\eta,c),$$

$$c = \frac{p\theta}{\mu}.$$

We may interpret the fraction c as a ratio of the average service time by agents vs. average service time by IVR. It is important to remark that this average service time by an agent includes the service time of customers that do not continue service by agents, i.e. service time that equals 0. Now let us consider the behaviour of $f_2(\beta,\eta,c)$ when we change its parameters.

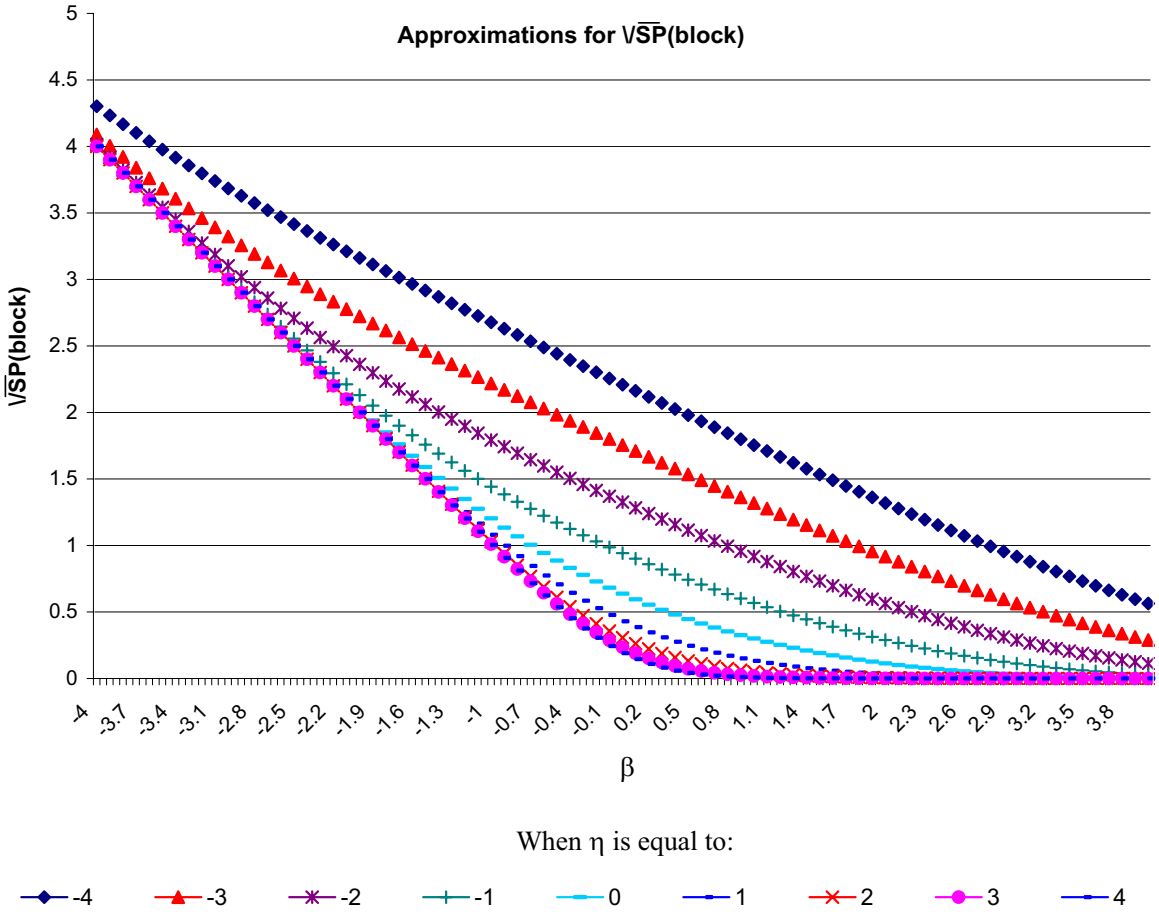

Figure 9.1: The illustration of the changing of the approximation for $\sqrt{SP}(block)$ when the parameters η and β are changing and $c = 1$.

Figure 9.1 shows the behaviour of $\sqrt{SP}(block)$ in the changing patterns of η and β from -4 to 4, when $c = 1$. We can see that with the growing of β, the function decreases, even if η is small. For $\eta > 1$ the values of the function f_1 are close to each other.

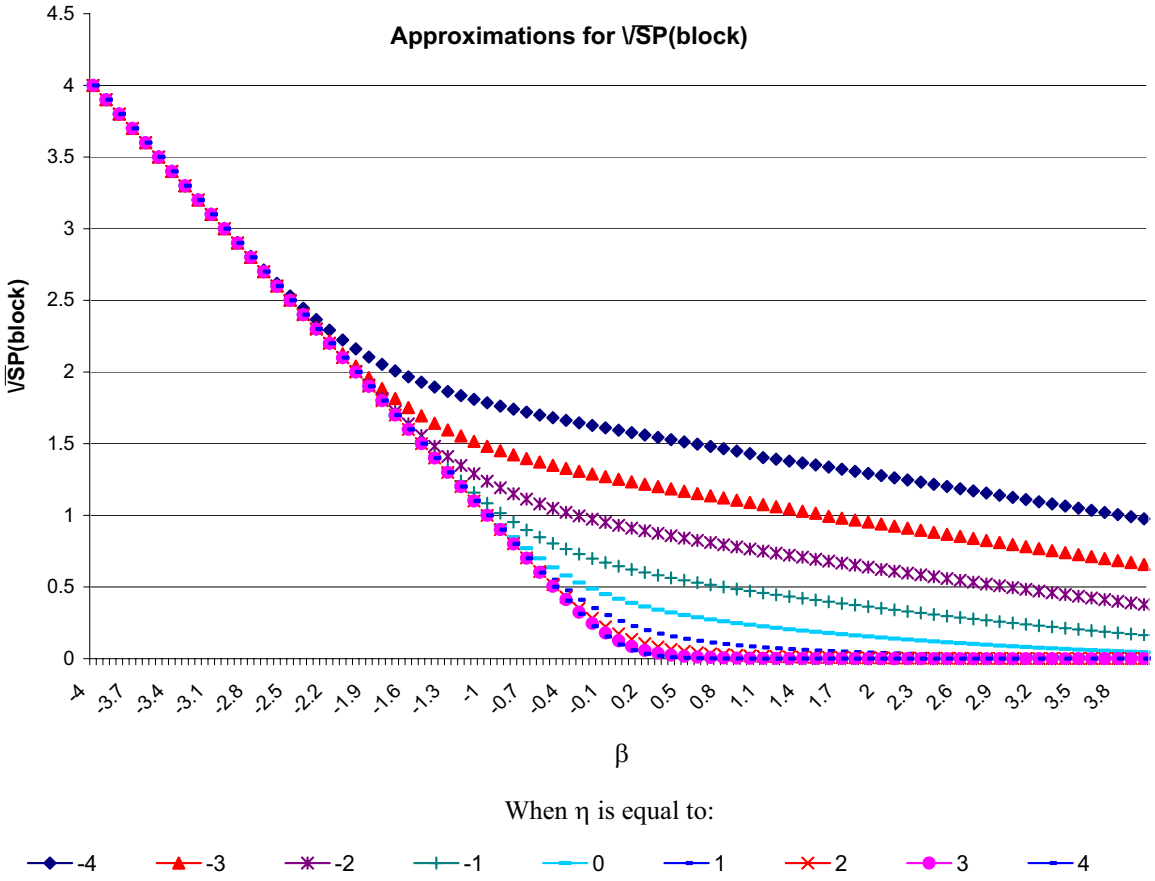

Figure 9.2: The illustration of the changing of the approximation for $\sqrt{SP}(block)$ when the parameters η and β are changing and $c = 0.2$.

Figure 9.2 illustrates the case for $c = 0.2$, meaning that the average service time in IVR is significantly higher than the average service time by agents for all customers. For such a call center, while $\beta \leq -2$ the values of the function f_2 are almost the same, however after this point the values of the functions highly differ for $\eta < 0$. Like in the above case with $c = 1$, when $\eta \geq 1$ the values of the function f_2 are once again close to each other.

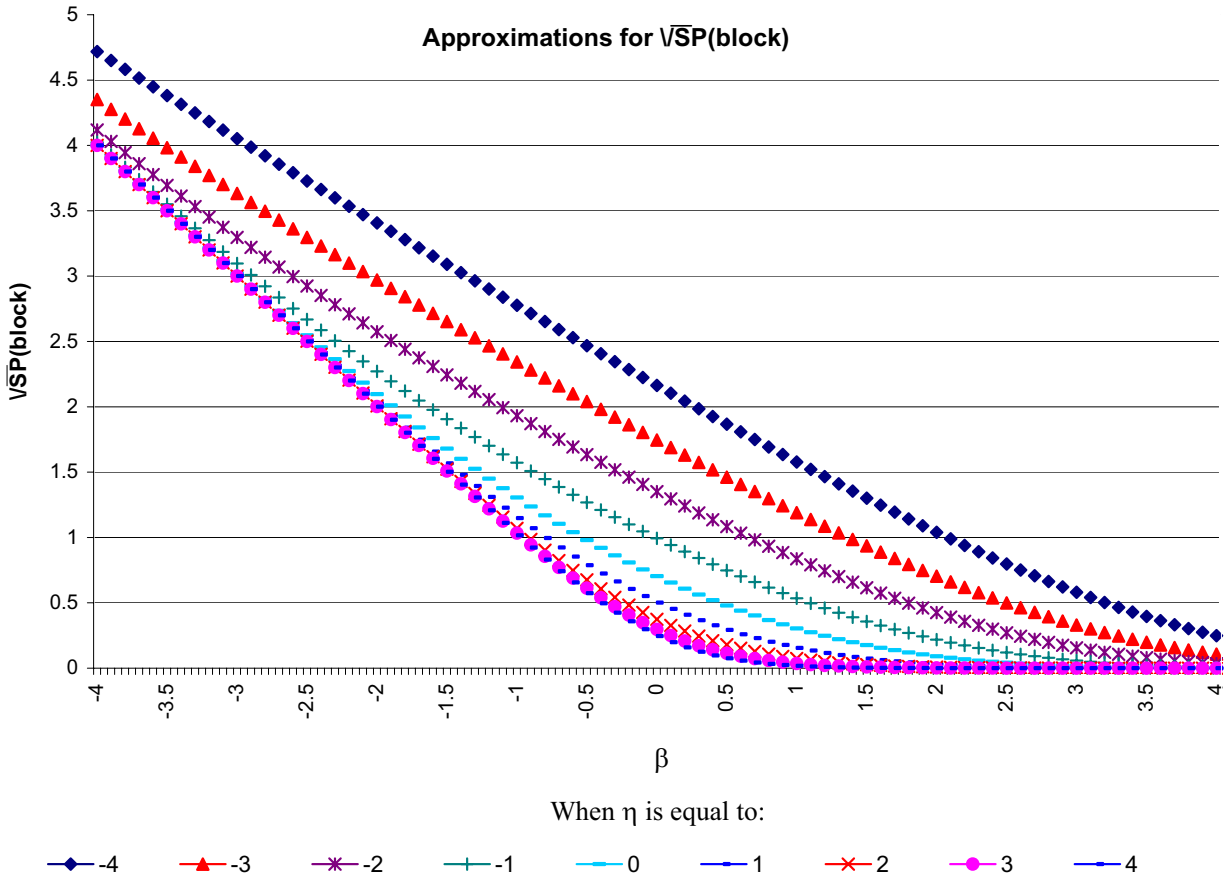

Figure 9.3: The illustration of the changing of the approximation for $\sqrt{SP}(block)$ when the parameters η and β are changing and $c = 5$.

Taking the same parameters of the call center, but now with $c = 5$, we receive Figure 9.3. As before, when $\eta > 1$ the values of the function f_2 are very similar. So, we can conclude that in many cases such values of η, i.e. $\eta > 1$, will not correspond to the optimal number of trunk lines. All the cases analyzed above (Figures 9.1-9.3) demonstrate that with the growing of β, the function decreases, even if η is small. When β is close to -4, the values of the functions are close to each other, and it happens because of the system being overloaded, meaning too many customers are waiting to be served while there are not enough agents. For such lack of the staff, the increase of the trunk lines number will not be efficient and this is what we call "system explosion". Actually, the above figure supports

the conclusion of Case C in Section 4.8, i.e.

$$f_2(\beta, \eta, c) = \lim_{\lambda \to \infty} \sqrt{S} P(block) \longrightarrow \begin{cases} -\beta, & \beta < 0, \\ 0, & \beta > 0 \end{cases}$$

Also using the fact that when η decreases the probability to find the system busy grows we can also say that $\sqrt{S}P(block) \geq -\beta$. This fact helps to formulate the following rule of thumb:"Under the given average arrival rate λ, the average service time $\frac{1}{\mu}$, and the probability that the customer will be served by an agent p, we can find the minimal number of agents S, which satisfies the condition $P(block) \leq b$. This number is equal to $S = \frac{\lambda p}{\mu(1+b)}$".

The function is going to be 0, when β is positive and growing. It can be explained for η is large enough when compared with the number of arriving calls. In case when $\eta < 0$, this is a less comprehensible prima facie. But on closer examination, we can see that a small η means a small queue size. At the same time, big values of β cause short waiting times. In terminology of this thesis that means the system is working in an efficiency-driven operational regime. So, the probability to find the system busy is really decreasing when β grows.

Now let us look at the behaviour of the approximation of the probability to wait, which was found in Theorems 4.3.1 and 4.3.2. As we said this approximation is a function of β, η and $c = \frac{\mu}{p\theta}$. So, let us see how the function f_1 changes with the various values of the parameters.

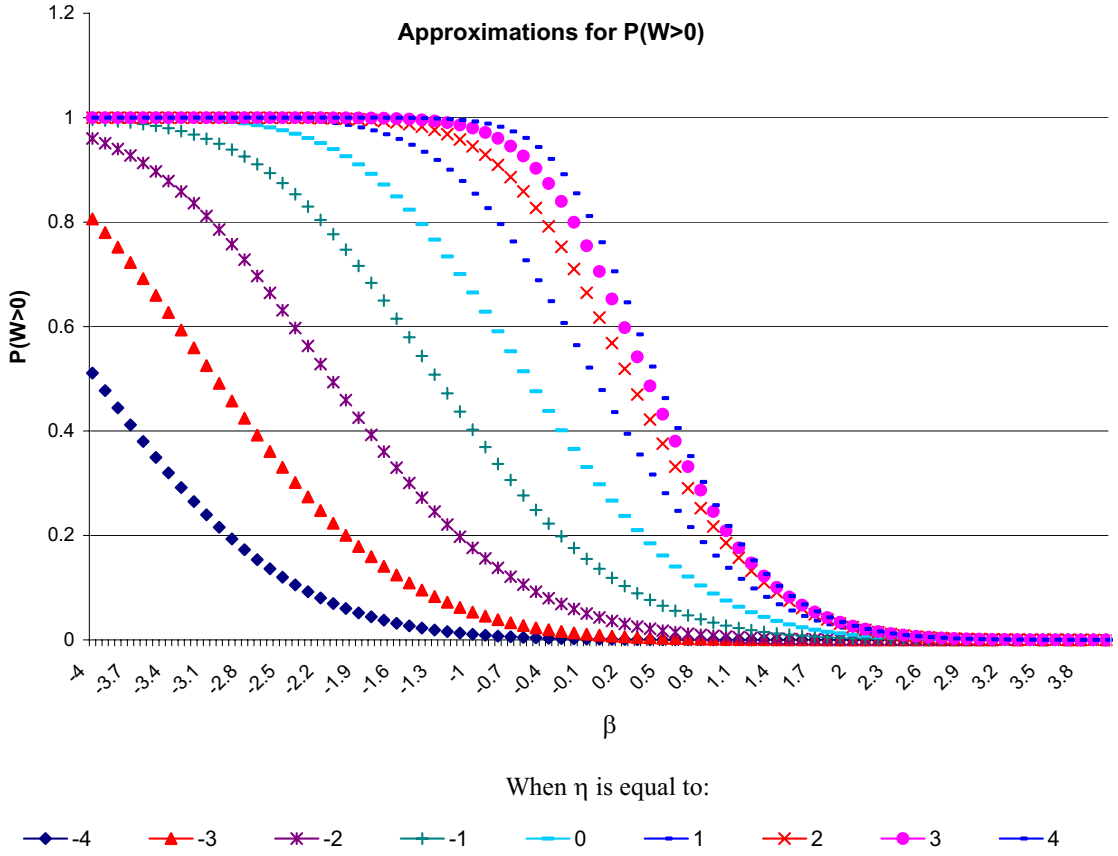

Figure 9.4: The illustration of the changing of the approximation for $P(W > 0)$ when the parameters η and β are changing and $c = 1$.

Figure 9.4 demonstrates changes in the function $f_2(\beta, \eta, c)$ while $c = 1$ and $\eta,\ \beta \in [-4, 4]$. For every $\eta \geq 2$ the values of the functions are close to each other. For smaller values of η $f_1(\beta, \eta, c)$ decreases and this confirms supposition regarding to the relation between the trunk lines and the probability to wait.

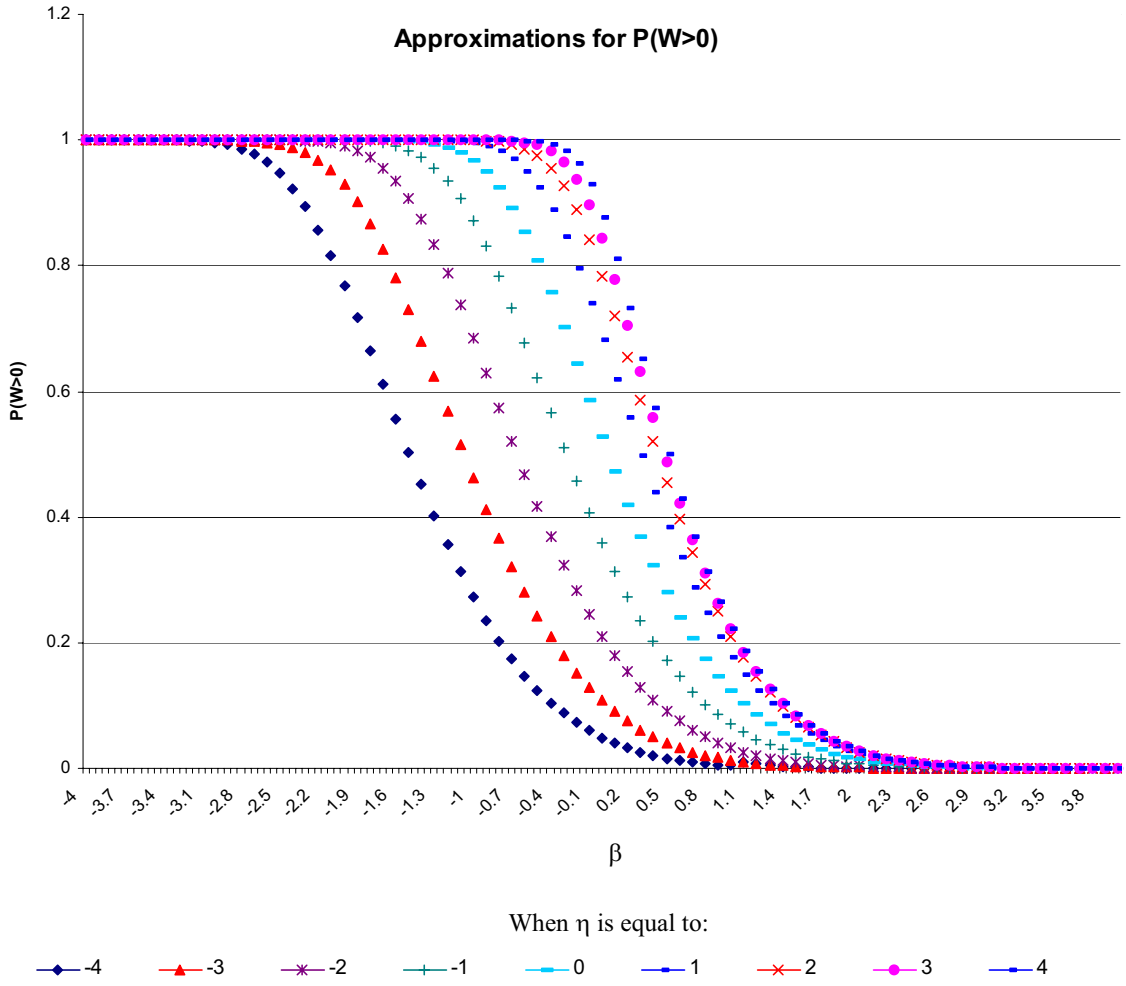

Figure 9.5: The illustration of the changing of the approximation for $P(W > 0)$ when the parameters η and β are changing and $c = 0.2$.

A call center with $c = 0.2$ is demonstrated in Figure 9.5. Here we see that the differences between the values of the functions $f_1(\beta, \eta, c)$ for various η are less significant than in the case of $c = 1$. This can be explained by the smaller probability of $P(block)$.

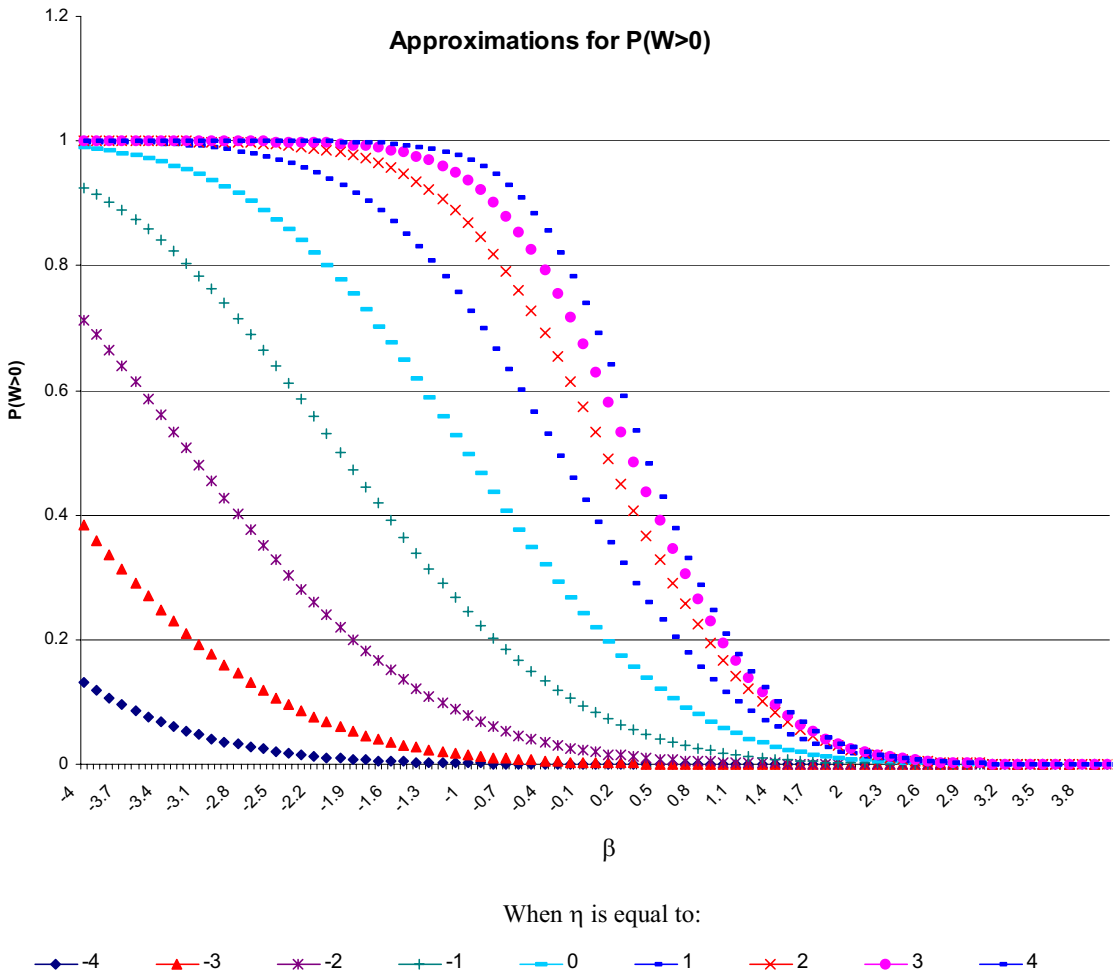

Figure 9.6: The illustration of the changing of the approximation for $P(W > 0)$ when the parameters η and β are changing and $c = 5$.

The case of $c = 5$ is similar to the case $c = 1$. The difference is in the speed of decreasing of the function with decreasing of β and this is caused by the same tendency in the probability to find the system busy.

When we solve the problem (9.3) we are searching for the smallest β and η for which the restriction are true. Suppose that $c = 1$ and plot the function f_1 and f_2 in the same figure.

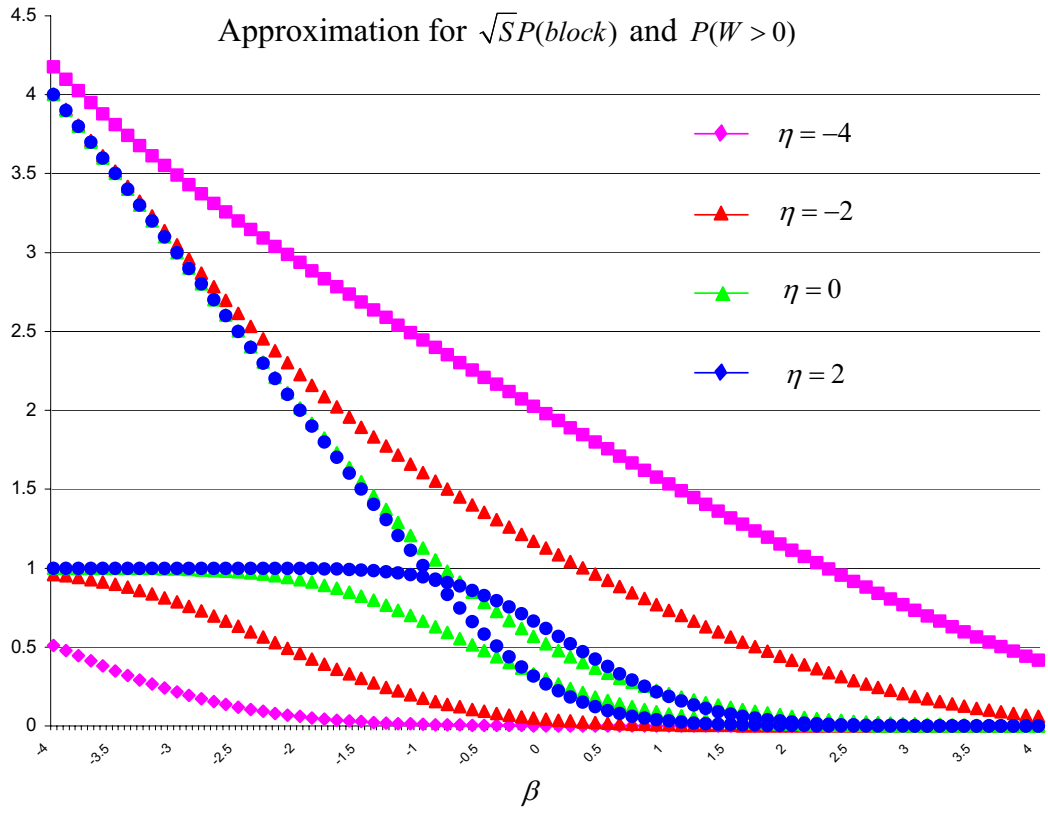

Figure 9.7: The illustration of the changing of the approximation for $\sqrt{S}P(block)$ and $P(W > 0)$ when the parameters η and β are changing and $c = 1$.

Suppose we are looking for an optimal solution of the problem (9.3), when $a = 0.5$ and $b/\sqrt{S} = 0.02$. Consider first the function f_1. Figure 9.7 shows that the smallest β and η for which the restriction is true are $\beta = -4$ and $\eta = -4$. On the other side, for this pair $f_2 = 4.2$. This means that if we wish to satisfy the second restriction in the problem (9.3) we need the number of agents in a call center to be at least $S = \left(\frac{4.2}{0.02}\right)^2 = 44100$. Thus, the pair $(\beta = -4, \eta = -4)$ can be the optimal pair for only a huge call center. If we want the number of agents in a call center to be about 1000, we need the value of the function f_2 to be about 0.65, and the optimal value of β, which corresponds to $\eta = -4$ will be 3.2. Taking into account that the cost of the trunk lines constitutes about 7% of the staffing cost, it is easy to see that this solution cannot be optimal. A more appropriate way is to take the pair $(\beta = -0.25, \eta = 0)$. By analyzing Figure 9.7 we can conclude that the optimal solution of the problem (9.3) will probably be a pair (β, η), where β and η are between -3 and 3. Of course, these values depend on the size of a call center.

We consider the case when $c = \frac{p\theta}{\mu} = 1$, but, of course, this value can be less or more than 1. Figures 9.2-9.6 show that when $c \neq 1$ we receive the same range for optimal pair (β, η). The explanations of this fact are as follows:

- In Section 7.3 it was shown that when $\eta \to \infty$ the system with an IVR may be modelled as the M/M/S queue. Borst, Mandelbaum and Rieman showed in [5], that in the case of M/M/S queue the β^*, which corresponds to the optimal solution $S_{opt} = \frac{\lambda}{\mu} + \beta^* \sqrt{\frac{\lambda}{\mu}}$ is less than 3. Taking into account the fact that the limitation of a number of trunk lines decreases the probability to wait we can also say that our $\beta < 3$.

- In Section 4.8 we proved and in this section we illustrated that when $\beta < 0$ the probability to find the system busy satisfies the following restriction

$$\sqrt{S} P(block) \approx -\beta.$$

  Usually, a manager of a call center wishes that the probability to find the system busy is less than 2%. Then, if $\beta = -3$ the size of a call center must be at least $S = \left(\frac{3}{0.02}\right)^2 = 22500$. There are not many call centers in the world of such a big size. Thus, we can suppose that usually $\beta > -3$.

- Analyzing Figures 9.2-9.6 we can see that when $\eta > 3$ the values of performance measures are almost not changing. On the other side, when $\eta < -3$ we can see that the probability to find the system busy got too big values and needs a very big value of β in order to satisfy the needed restriction. And what is more, sometimes it is even impossible to achieve a needed value, because as we showed in Section 4.8 $f_2 \to \infty$ when $\eta \to -\infty$.

Thus, we conclude that for a regular call center the optimal solution (S, N) of the problem (8.2) has the following domain:

$$\begin{aligned} &(i) \quad N - S \approx \frac{\lambda}{\theta} + \eta \sqrt{\frac{\lambda}{\theta}}; \\ &(ii) \quad S \approx \frac{\lambda p}{\mu} + \beta \sqrt{\frac{\lambda p}{\mu}}; \end{aligned} \tag{9.4}$$

where $-3 < \beta < 3$ and $-3 < \eta < 3$.

9.3 Some ways to reduce operating costs

As has been noted several times, staffing costs (salary, training, etc.) account for over 65% of the operating costs of a typical call center. Hence, the main way to reduce costs is to reduce the number of agents. In this section we consider two methods of doing that, as well as some of their advantages and disadvantages.

9.3.1 Reducing the number of trunk lines

The first way to reduce operating costs is to reduce the number of trunk lines and the number of agents. The effectiveness of this method was discussed in Section 5.2. There we saw that even minor reduction in the number of trunk lines caused the improvement of some characteristics, such as the probability to wait and the average waiting time. It is thus possible to reduce the number of agents so that the system still satisfies a desired service level requirement.

The disadvantage of this approach is the increase of the probability to find the system busy. The following example illustrates what can be achieved by reducing the number of trunk lines and what can be lost as well.

Consider a call center with an IVR in which:

- the average arrival rate $\lambda = 80$,
- the number of agents S equals 82,
- the number of trunk lines N equals 200,
- the average service rate in the IVR θ equals 1,
- the average agent's service rate μ equals 1,
- the probability to be served by an agent p equals 1.

Thus, the operational performance measures are as follows:

- $P(W > 0) = 0.65$,
- $P(block) = 0.01$.

Suppose that we wish to reduce $P(W > 0)$ to 0.4. There are two ways to do that:

(1) reduce 30 trunk lines, from 200 down to 170;

(2) add 4 agents, from 82 to 86.

At first glance it may seem that the more appropriative way to achieve the goal is to reduce 30 trunk lines. But in this case the probability to find the system busy increases to 0.03. If we would like to reduce this probability back to the previous value 0.01 we need to add 7 agents. The values associated with this example are illustrated in Table 9.1:

	S	P(W>0)	P(block)
N=200			
	82	0.65	0.01
	86	0.4	0.004
	89	0.22	0.0015
N=170			
	82	0.4	0.03
	86	0.2	0.015
	89	0.01	0.01

Table 9.1: Operational performance measures for a call center with an IVR, when $p = \mu = \theta = 1$ and the arrival rate equals 80.

As was said in Section 8.1, the trunk costs constitute 5% and the staffing costs - about 65% of a call center's operational costs. If we suppose that these proportions are not changing with adding or reducing agents or trunk lines, then the solution (1) will cost approximately 5% more than the current costs and the solution (2) will cost around 3% more. Thus we conclude that the second way is more appropriate.

9.3.2 Adding functionality to the IVR

The effective way to reduce costs without making the service level worse is to extend the IVR's capabilities. Adding functions to the IVR will decrease the probability p to be served by an agent. Indeed, the operations which previously only an agent could do are now carried out routinely by the "self-service" customer. Therefore, the number of customers wishing to be served by agents will decrease and, as a result, the number of agents S will decrease as well.

As an example, consider a call center with the following parameters:

(i) the average arrival rate $\lambda = 1000$,

(ii) the average service rate in the IVR θ equals 1,

(iii) the average agent's service rate μ equals 1.

Suppose that the performance constraints are as follows:

$$\begin{aligned} P(W > 0) &< 0.4; \\ P(block) &< 0.02. \end{aligned} \tag{9.5}$$

We can find the optimal pair (S, N), where N is the number of trunk lines and S is the number of agents $0 \leq S \leq N$. This pair (S, N) minimizes the costs and provides the desired level of service. The algorithm for solving this problem was described in Chapter 8. In this algorithm we used the exact formulas of performance measures. It was easy, because we consider a relatively small call center (with the arrival rate $\lambda = 20$). In the current example and in all the other examples in this chapter we will consider big call centers (the arrival rate λ is about 1000). The calculation of exact performance measures in such a big call center takes a lot of time and needs a complicated programming process. Thus, we will use approximations for performance measures for finding the optimal solution.

Let us change p - the probability to be served by an agent over the range from 0 to 1, and for each case find the optimal solution (S, N).

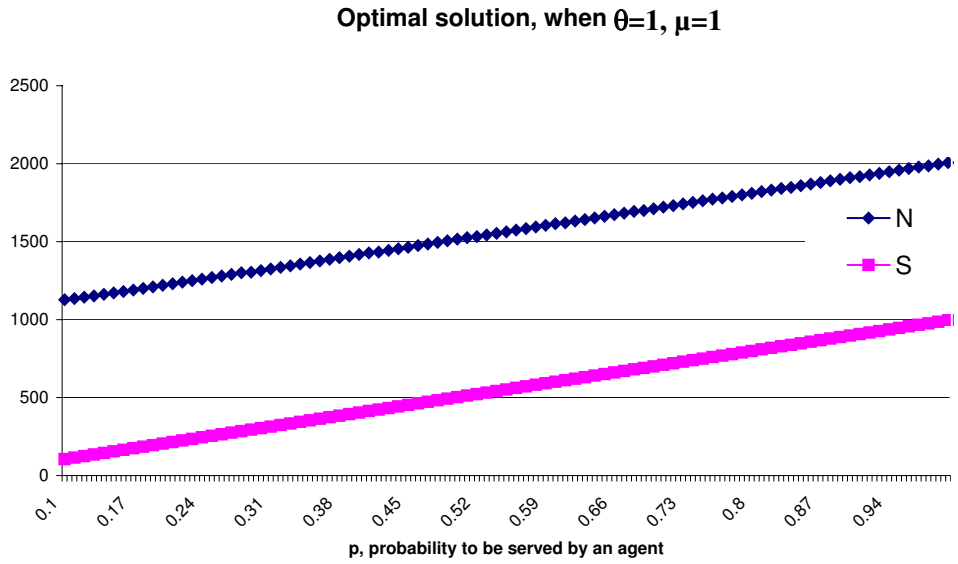

Figure 9.8: The optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's and the IVR's service rates equals 1 and p changes from 0 to 1 .

The dependence between the probability p to be served by an agent and the optimal number of trunk lines N and the optimal number of agents S, as in Figure 9.8 looks linear. Is it really a linear dependence? The domain of (S, N) values is given in (9.4). Consider, for example, the relationship between S and p: $S \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}}$. The values of β are small and so the influence of $\beta\sqrt{\frac{\lambda p}{\mu}}$ is negligible in comparison to $\frac{\lambda p}{\mu}$. Using this fact we can do rough approximation of the optimal number of agents. For example, if $p = 0.5$ we can predict that the optimal number of agents S will be equal to 500. We can see in Figure 9.8 that it is almost true, but because of a small resolution we cannot see the value exactly. Note, that such an approximation is too rough and the optimal number of agents can be actually equal to 550, i.e. the error is 10%.

The relationship between N and p are similar to the relationship between S and p plus the item, which does not depend on p:

$$N \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}.$$

Because of this, the line N is parallel to the line S and the difference is equal to $\frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}$. Moreover, we can roughly say that the difference is equal to $\frac{\lambda}{\theta}$, i.e. 1000 in our example, because the item $\eta\sqrt{\frac{\lambda}{\theta}}$ is negligible compared with $\frac{\lambda}{\theta}$.

Now, let us continue to analyze changing of the optimal solution when changing the parameters of the system. It is reasonable to assume that along with changing the probability to be served by an agent the service time in the IVR is changing as well. Unfortunately, we do not know how those changes occur, but first let us consider what happens when we change the parameter θ. Let the average service rate in the IVR be equal to 0.2, i.e. the average service time in the IVR is 5 times more what it was before. We then get the following figure:

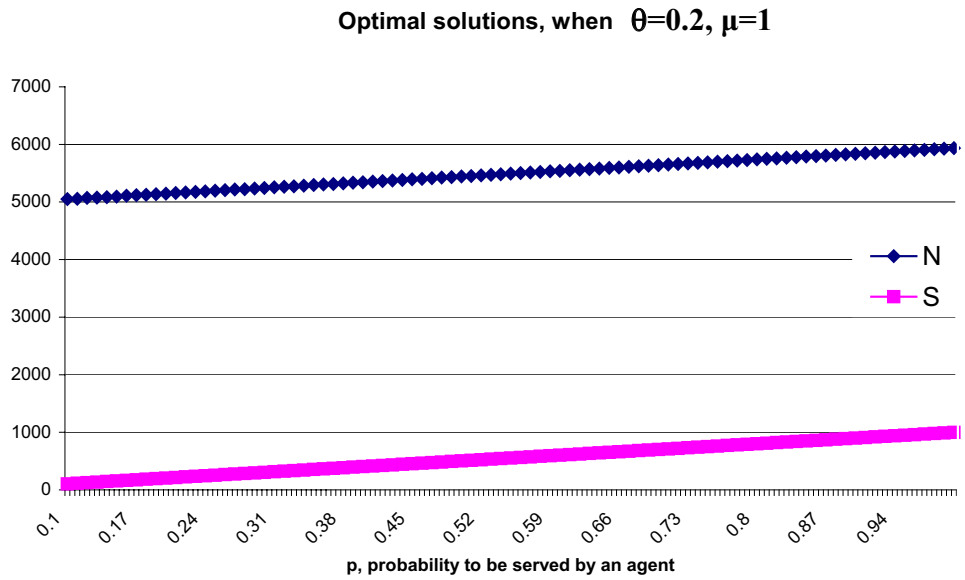

Figure 9.9: The optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's service rate equals 1, the IVR's service rate equals 0.2 and p changes from 0 to 1 .

Let us compare Figure 9.8 with Figure 9.9. When $\theta = 0.2$, the optimal number of agents is exactly the same as when $\theta = 1$, and only the number of trunk lines N changes. Actually, it is not surprising, because we saw that the optimal value of S does not depend on θ. When $\theta = 0.2$, the values of N are about 5000 trunk lines more than when $\theta = 1$, and it happens because now the difference between N and S is about $\frac{\lambda}{\theta} = \frac{1000}{0.2} = 5000$.

Let us see what happens when $\theta = 5$, i.e. the average service time in IVR is 5 times less than in the first case and 10 times less than in the second case. We can hypothesize that the optimal number of agents will be the same as in the previous cases and the optimal number of trunk lines will be less, and the difference found will be as follows: $N_{opt}(\theta = 5) \approx \frac{1}{5}(N_{opt}(\theta = 1) - S_{opt}(\theta = 1)) + S_{opt}(\theta = 5)$. Our intuition is that the optimal number of agents will not change and therefore the

optimal number of trunk lines will be about 900 trunk lines less than in the case when $\theta = 1$. Figure 9.10 supports this intuition:

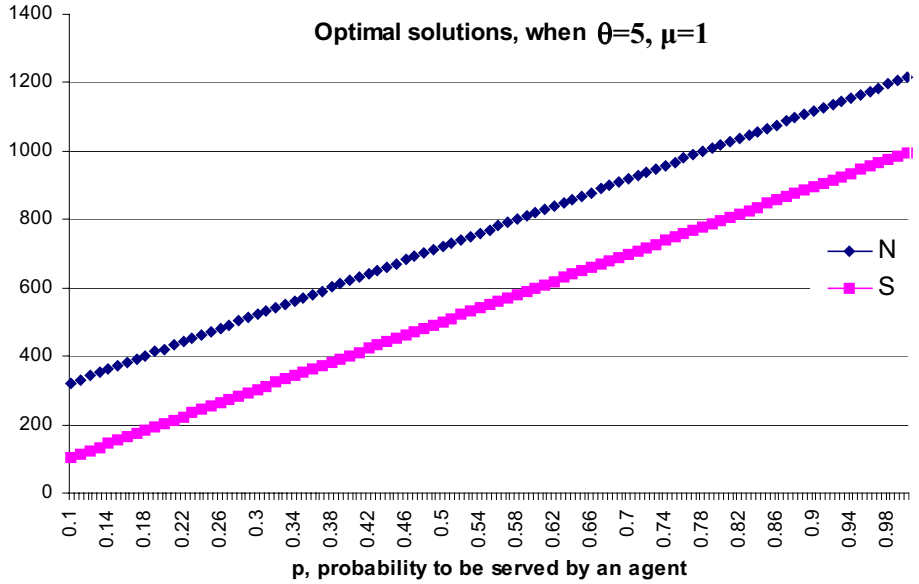

Figure 9.10: The optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's service rate equals 1, the IVR's service rate equals 5 and p changes from 0 to 1.

Now, let us assume that the service rate in the IVR is a function of p. Intuitively this function must be an increasing function, because when the number of customers wishing to be served by an agent increases the time that these customers spend in the IVR is decreasing, therefore the service rate in the IVR is increasing. For simplicity, suppose that this function is linear such that when nobody wishes to be served by an agent ($p = 0$), the average service rate in the IVR equals 1, and when everyone wishes to be served by an agent ($p = 1$), the average service rate equals 5. Thus, this function has the following form:

$$\theta(p) = 4 \cdot p + 1. \tag{9.6}$$

According to the previous analysis we can guess that changes in the service time in the IVR will not influence the optimal agent's number. Now let us look at the optimal solution (S, N) to the problem (9.5), but in the case when θ is given by (9.6).

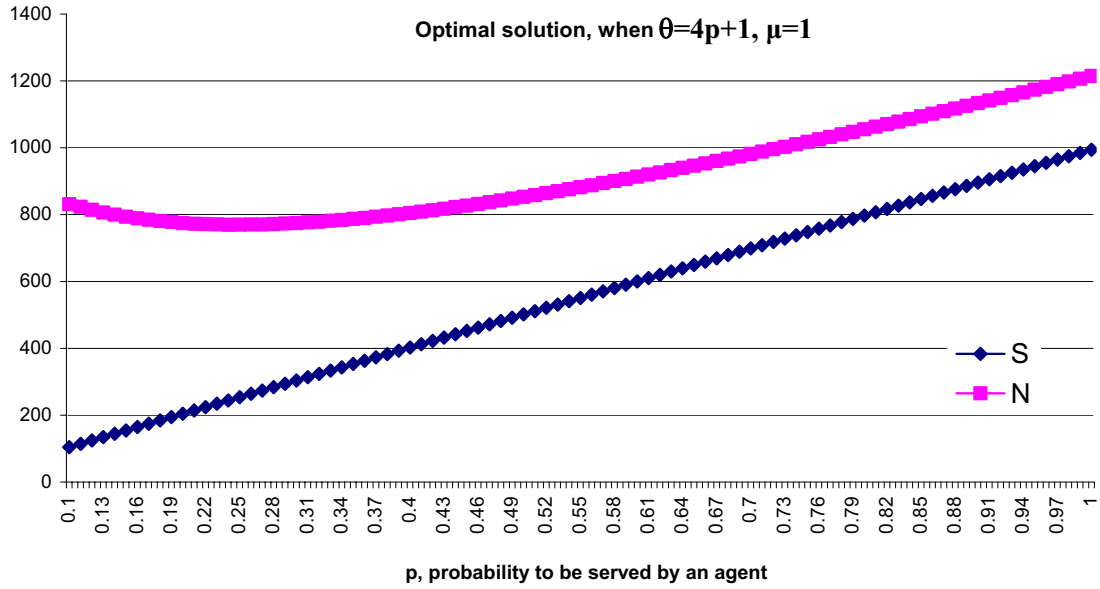

Figure 9.11: The optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's service rate equals 1, the IVR's service rate depends on p, and p changes from 0 to 1.

Figure 9.11 shows that our intuition was right. The optimal number of agents did not change. This fact is very important, because we can see once again that adding functions to an IVR is a good way to reduce costs of a call center. The line N values for this now does not look linear. So, it is a line similar to $N \approx \frac{\lambda}{\theta} + \frac{\lambda p}{\mu}$. This is a rough approximation, because we do not take into account the items $\eta\sqrt{\frac{\lambda}{\theta}}$ and $\beta\sqrt{\frac{\lambda p}{\mu}}$, but the values of these items are negligible and do not influence the form of the line N.

Another property that can be manipulated with the addition of functions to the IVR is the service time at the agents' pool. Indeed, if the IVR has more functionality, then the agents do not need to do part of the functions. Thus, it is natural to suppose that there will be a decrease in the average agent's service time and, as a result, it will lead to a decrease in the optimal number of agents. Another average agent's service time might increase as a result of additional functions to an IVR. Customers might have questions as for the IVR usage, since it is more complicated now. Moreover, the customers who will be served by an agent after adding more functions to an IVR are expected to have more complicated requests which take longer to be satisfied. Thus, the relationship between the probability p to be served by an agent and the rate μ of an agent's service is not easy to predict. Consider two scenarios of changing the rate of the

agent's service:

$$\mu = 1 + 2 \cdot p \cdot (1 - p) \tag{9.7}$$

and

$$\mu = 1 - 2 \cdot p \cdot (1 - p). \tag{9.8}$$

Let us now see the illustration comparing the optimal number of agents in the two scenarios.

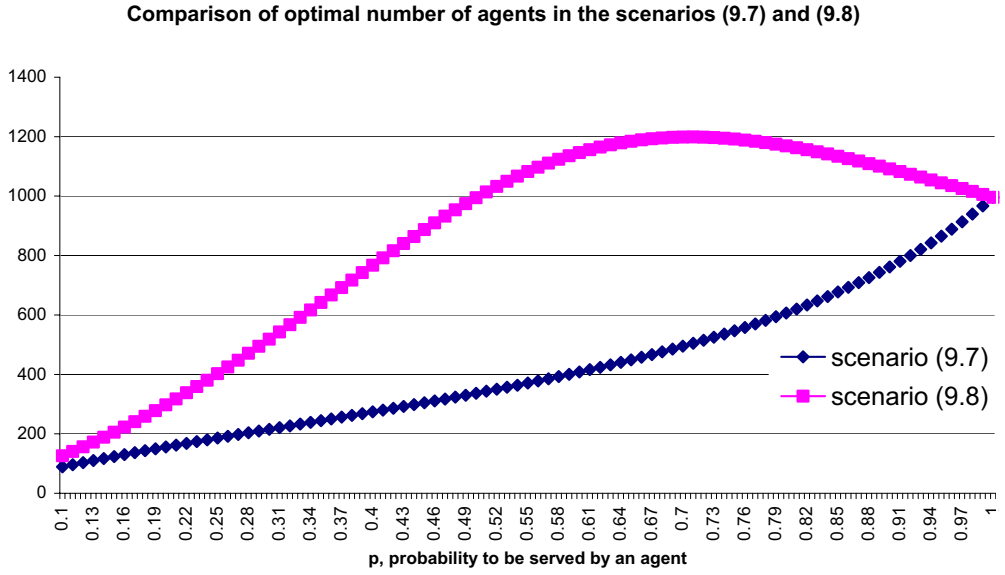

Figure 9.12: A comparison of the optimal number of agents for a call center with an IVR, when the arrival rate equals 1000, the IVR's service rate depends on p, p changes from 0 to 1, and the agent's service rate μ in the first scenario it equals $1 + 2p(1 - p)$ and in the second scenario equals $1 - 2p(1 - p)$.

Figure 9.13 shows that the first scenario is preferable to the second. We can come to the same conclusion using the picture below, which shows the comparison of the optimal number of trunk lines for both scenarios.

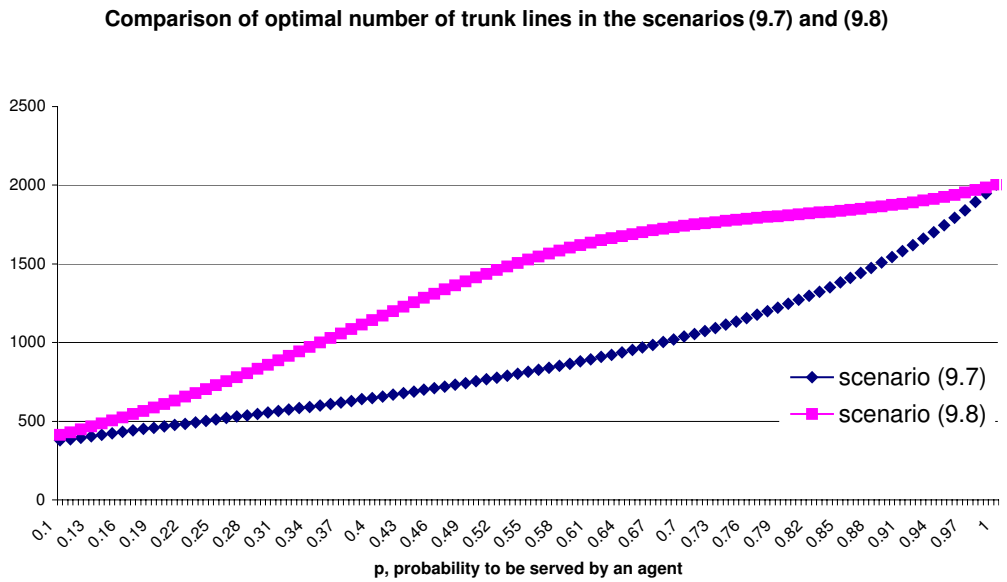

Figure 9.13: A comparison of the optimal trunk lines numbers for a call center with an IVR, when the arrival rate equals 1000, the IVR's service rate depends on p, p changes from 0 to 1, and the agent's service rate μ in the first scenario equals $1 + 2 \cdot p \cdot (1 - p)$ and in the second scenario it equals $1 - 2 \cdot p \cdot (1 - p)$.

Thus, we can see that adding functions to an IVR can provide a very good solution for the costs reduction. However, sometimes this may bring undesirable changes. Such a result can be seen in the second scenario. Indeed, before adding functions to the IVR, i.e. when p was equal to 1, the optimal agent's number was 503. After the addition of some functions to the IVR, for example, when $p = 0.7$, the optimal number of agents in the second scenario was 605. This is almost by 20% more than it was before. As said in Section 8.1, the trunk costs constitute 5% and the staffing costs - about 65% of a call center's operational costs. If we suppose that these proportions are not changing with adding or reducing agents or trunk lines, then after adding functions to an IVR, the call center's costs increase by about 14%. Such a scenario can happen as a result of unsuccessful design of an IVR, which shows how important it is to be careful while designing it. As we said in the Section 1.1, when we deploy an IVR we need to take into consideration not only a call center's interests, but also the wishes, needs and possibilities of customers in order to make it an easy to use, understandable and convenient application.

9.4 Investigation of the effect of changes in p, θ and μ on the optimal solution (S, N)

Let us now return to the original problem (8.2) with restriction (9.5) in order to consider the effect of changes in p, η and μ and their influence on the optimal solution (S, N). Actually, we began this analysis in the previous section. Figure 9.8-9.10 showed that the dependence between the probability p to be served by an agent and the optimal pair (S, N) looks linear. We explained this fact by using the domain of (S, N) values:

(i) $N - S \approx \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}$;

(ii) $S \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}}$;

As can be seen, the main influence on S and N is that of the first term in (ii), thus the dependence is basically linear. The following questions arise now:

- How large is the value of the last term in (ii), i.e how strong is this term's influence on S and as a result on N?
- Is there really a linear dependence between S and p?
- What is the relationship between μ and S?
- What is the relationship between θ and $N - S$?

9.4.1 Effect of p

Let us try to answer the first and second questions. For this purpose we will consider β and η as functions of p and plot the values of β and η, which correspond to the original pair (S, N).

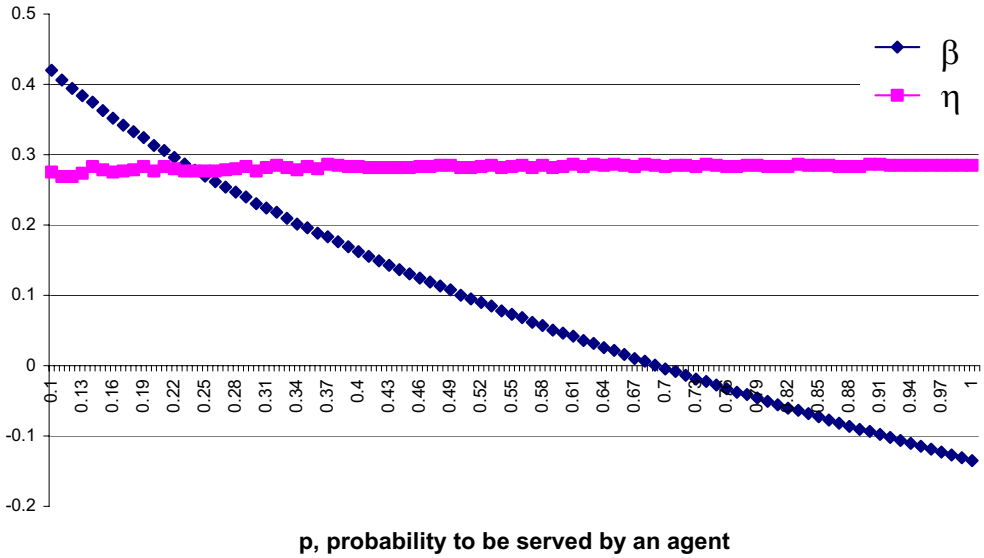

Figure 9.14: The values of β and η, that correspond to the optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's and the IVR's service rates equal 1 and p changes from 0 to 1.

Figure 9.14 shows that η is almost constant and β changes in a manner close to linear. Also, the plot shows that all values are close to zero, then the last terms in (i) and (ii) from the domain of (S, N) values, do not have a profound effect on the optimal pairs (S, N). For the reasons given above the dependence of values (S, N) on p seems linear (see Figures 9.8-9.10).

9.4.2 Effect of μ

Now let us consider in what way μ influences the optimal solution. From the definition of the domain of (S, N) values we can see that μ influences the number of agents and not the potential number of places in the queue $N - S$. So, we can suppose that η will be constant. The next plot confirms our assumption.

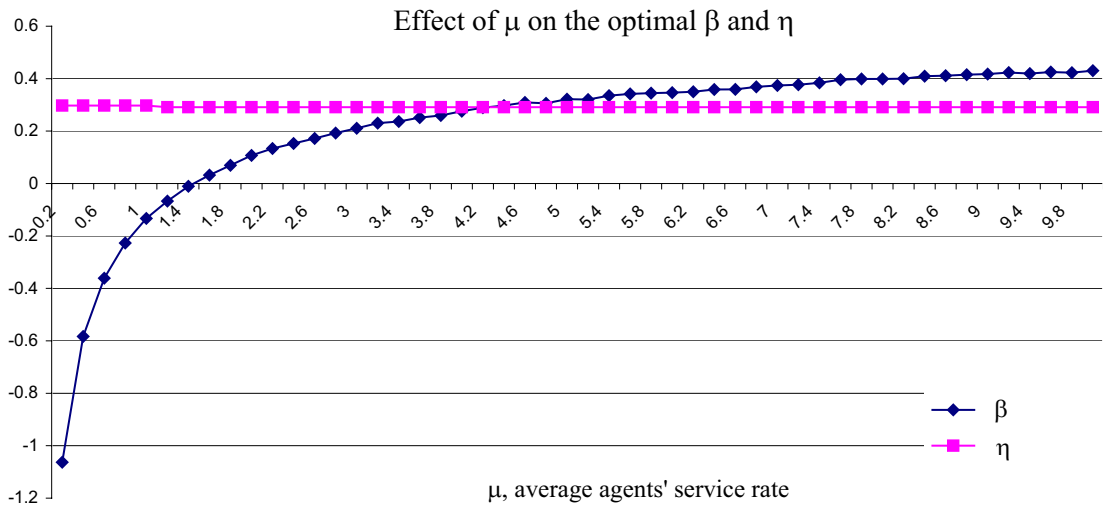

Figure 9.15: The values of β and η, that correspond to the optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, p and the IVR's service rates equals 1 and the agent's service rate changes from 0.2 to 10.

Figure 9.15 shows that η is almost constant, moreover, it equals the same value as in the case when p was changed. Thus, the optimal value of η does not depend on p and μ, but the optimal value of β depends on p and μ as well. A more detailed analysis of the domain of (S, N) values shows that μ is included into the same items as p. So, the optimal value of beta depends on the fraction $\frac{p}{\mu}$. Figure 9.14 shows the optimal values β and η as functions of $\frac{p}{\mu}$.

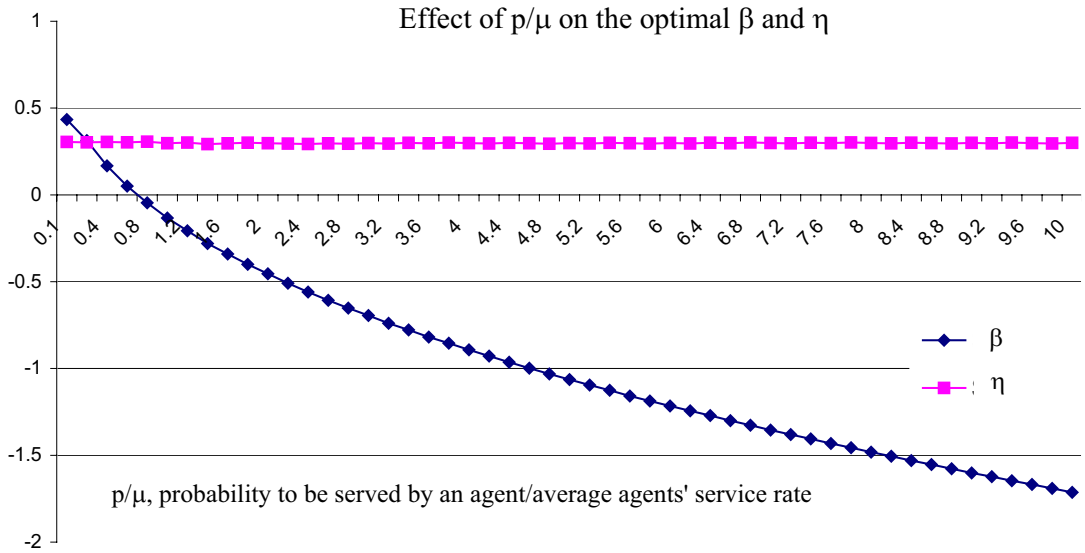

Figure 9.16: The values of β and η, that correspond to the optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, average service rate in the IVR θ equals 1 and the fraction $\frac{p}{\mu}$ changes from 0.1 to 10.

We can once again make sure that the values of η did not change. Furthermore, Figure 9.14 is a particular case of the Figure 9.16.

9.4.3 Effect of θ

In the domain of (S, N) values we can see that S is not dependent on θ. So, it is natural to suppose that with the changes of θ the value of β will not change either.

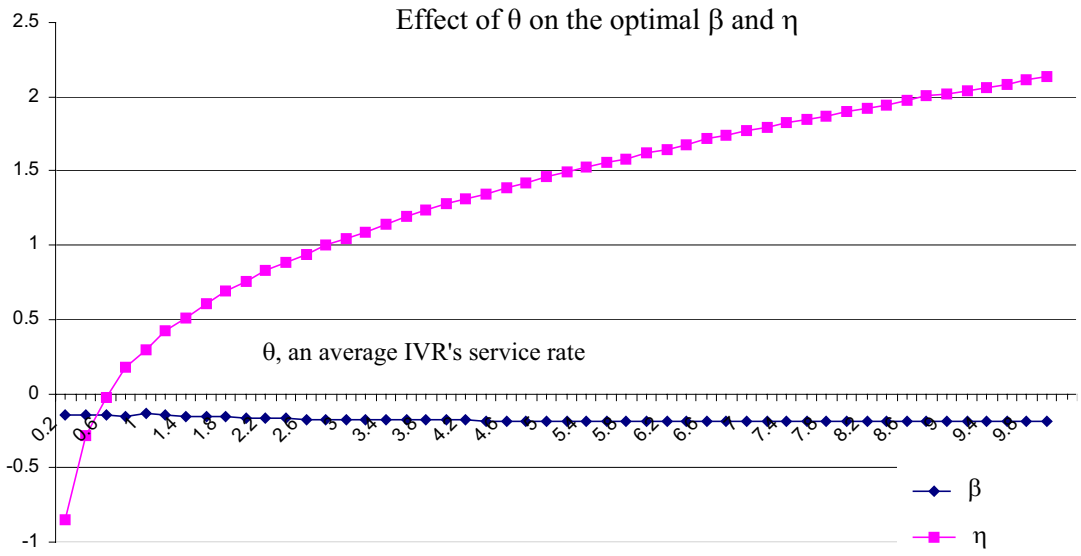

Figure 9.17: The values of β and η, that correspond to the optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, p and the agents' service rates equals 1 and the IVR's service rate changes from 0.2 to 10.

Figure 9.17 supports our assumption, i.e. an average service rate in an IVR does not depend on the optimal number of agents. The value of η grows with growing of the average service rate in the IVR θ.

9.4.4 Conclusions

After this section's analysis we can formulate the following conclusions:

- The items $\beta\sqrt{\frac{p\lambda}{\mu}}$ and $\eta\sqrt{\frac{\lambda}{\theta}}$ are small values in comparison with $\frac{p\lambda}{\mu}$ and $\frac{\lambda}{\theta}$ respectively, so sometimes we can neglect these values and predict the optimal solution in the following way: $S \approx \frac{p\lambda}{\mu}$ and $N - S \approx \frac{\lambda}{\theta}$.

- Growth of the probability p to be served by an agent or average agent's service time $\frac{1}{\mu}$ causes growth of the optimal number of trunk lines, but the value of η does not change.

- Growth of the probability p to be served by an agent causes growth of the optimal number of agents, but the value of β decreases.

- Growth of an average agent's service rate μ causes decrease of the optimal number of agents, but the value of β grows.

- If the values of the probability p to be served by an agent or an average agent's service time μ are changing, then it is convenient to consider the changing of fraction $p\mu$ instead of each of them taken separately. Growth of this fraction causes growth of the optimal pair (S, N), but the value of β decreases and the value of η remains constant.

- Growth of an average service rate θ in an IVR causes growth of the optimal agents' number S_{opt} and growth of β as well.

- Growth of an average service rate θ in an IVR causes decrease of the optimal number of trunk lines, but the value of η grows. Changes in the value of θ have no effect on β and as a result on the optimal number of agents S_{opt}.

9.5 Effect of the call center's size

First, let us have a look at the behaviour of $P(block)$. This characteristics depends on β and η, S and $\frac{p\theta}{\mu}$, but in this case we assume that $p = \theta = \mu = 1$.

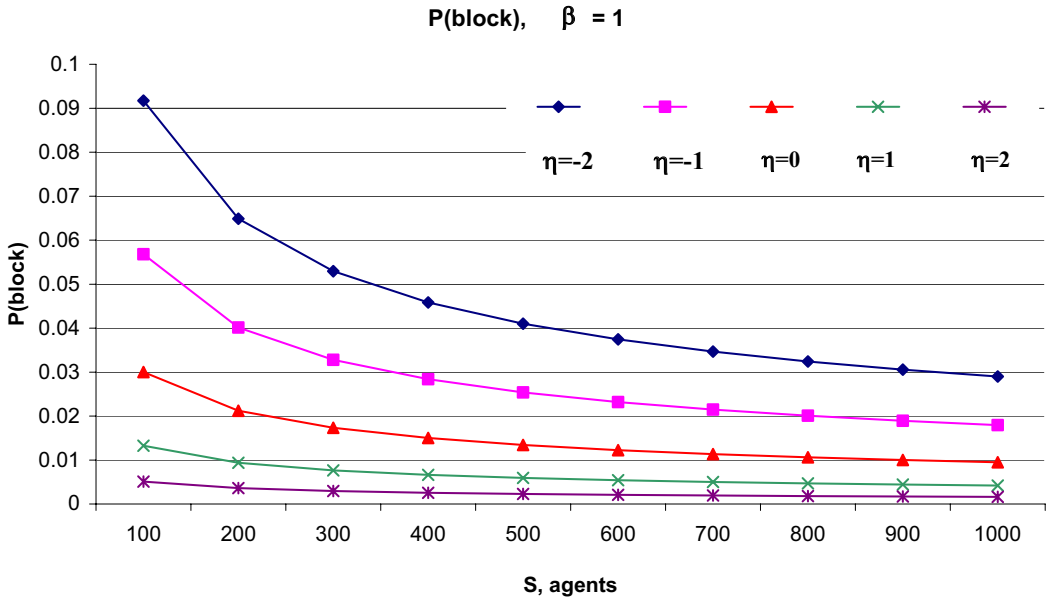

Figure 9.18: The illustration of the changing of the approximation for $P(block)$ when the parameter η is changing and β is equal to 1.

In the Figure 9.18 we plot changing of $P(block)$ when β is equal to 1, η is -2, -1, 0, 1 and 2. The number of agents S is changing from 100 to 1000. Usually, we say

a call center with the number of agents S=100 is mid-sized and when $S = 1000$ a call center is big-sized. In this graph we really see $P(block)$ decreasing with growing of S when all the other parameters are fixed. We can also see that with the growth of η this decrease is less distinctive. This fact can be explained by a slight changing of $\lim \sqrt{S}P(block)$ in this case. On the other side, when η is decreasing these changes are more visible. Thus, when $\eta = -2$ and $S = 100$ the probability to find the system busy is more than 9%, and when $S = 1000$ this probability is less than 3% under the same parameters. This is one more piece of evidence in the service level increasing with growing of a call center's size. But this increase is not always remarkable, even in the case when the arrival rate λ is big. For better understanding let us see the following graph:

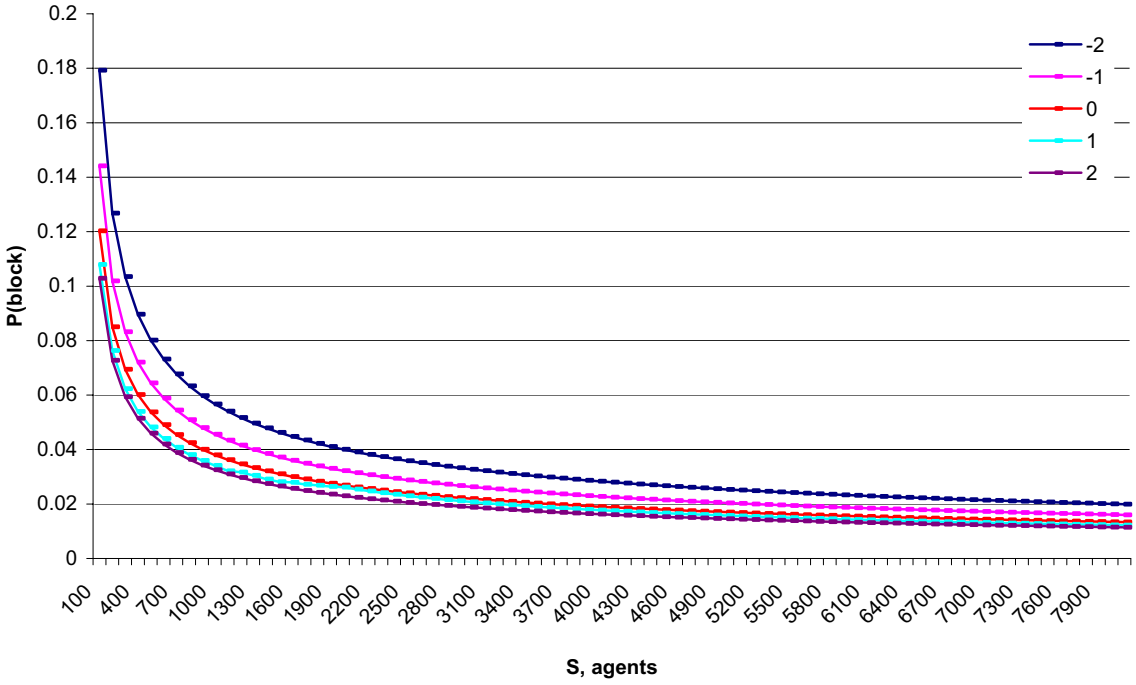

Figure 9.19: The illustration of changing of the approximation for $P(block)$ when the parameter η is changing and $\beta = -1$.

Figure 9.19 shows significance of the parameter β. We can see that even when our call center is huge, for example $S = 10000$, the probability to find the system busy still cannot be less 1% under the given system's parameters. Thus, this example also shows that parameters β and η are important even for a big call center.

Now, we would like to plot the effect of a call center's size on the values of β and η. First, let us look at the changes in β for call centers with the arrival rate

being one of the following:

$$\lambda = 500, \quad \lambda = 1000, \quad \lambda = 5000, \quad \lambda = 10000, \quad \lambda = 50000.$$

For each case β corresponds to the optimal solution (S, N).

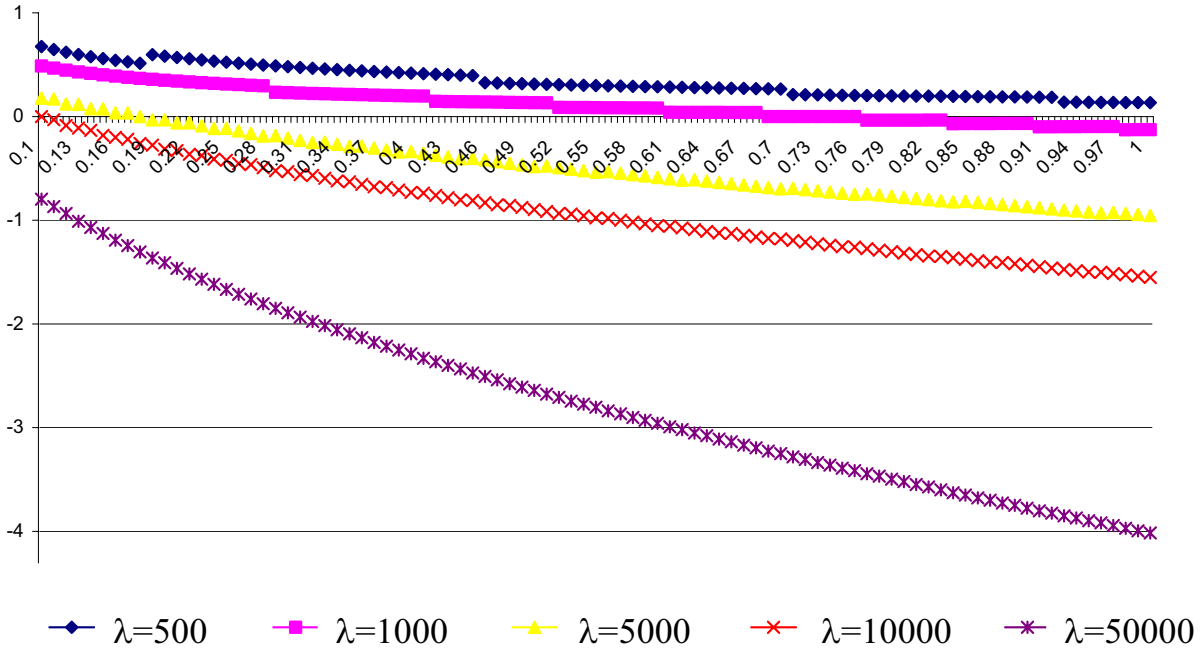

Figure 9.20: The values of β, that correspond to the optimal values S for a call center with an IVR, when the arrival rates are different, the agent's and the IVR's service rates equal 1 and p changes from 0 to 1.

The results show that all the values are close to zero, while β decreases as the call center's size increases. This once again supports the well known fact that a large call center works faster that a smaller one.

Next, we look at the changes in η for the same cases described above (see figure 9.21).

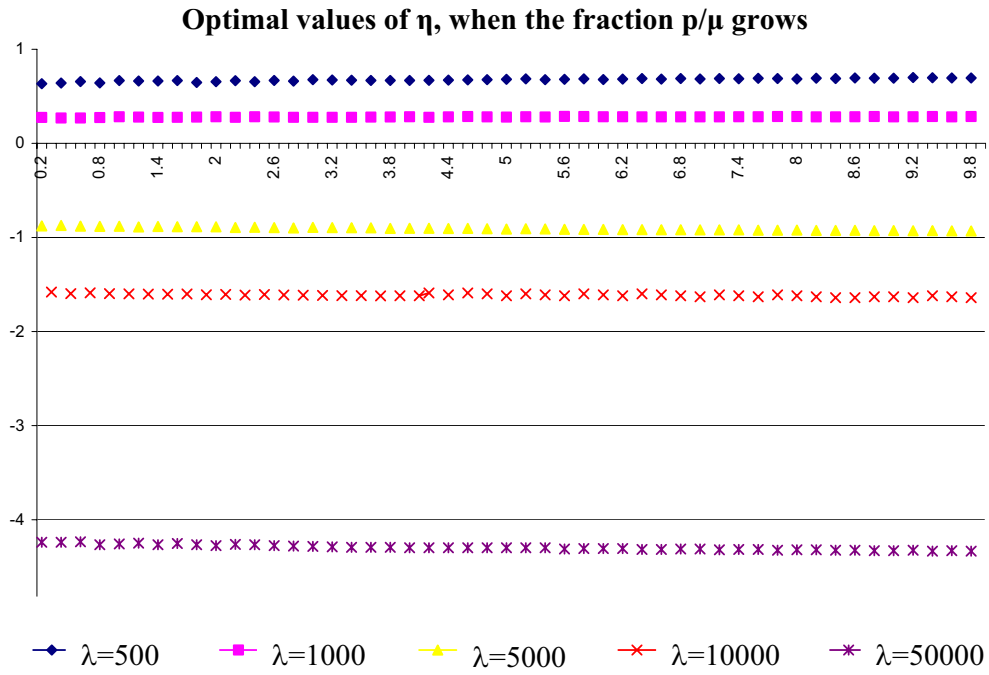

Figure 9.21: The values of η, that correspond to the optimal pairs (S, N) for a call center with an IVR, when the arrival rates are different, the agent's and the IVR's service rates equal 1 and p changes from 0 to 1.

Here, the values of η are almost constant and they decrease as a call center's size increases. Also, we can see that these values are small enough. Even for the call center with the arrival rate $\lambda = 50000$ the value of η is about -4.

Figure 9.20-9.21 approves the range for β and η, which we receive in Section 9.2. Recall, that for a regular call center β and η, which correspond to the optimal solution of the problem (8.2), are between -3 and 3.

Chapter 10

Future research

Finally, we outline some directions worthy of further research.

Adding abandonment and retrials to the model. The subject of this thesis is a Markovian model for a call center with an IVR. We have tried to make the problem as realistic as possible. However, some more work is required to understand the effects of retrials and abandonments. Analogous problems were analyzed by Garnett, Mandelbaum and Reiman in [13] for M/M/N queue and by Zeltyn and Mandelbaum in [25] for M/M/n+G queue.

Mixed customer population. One more realistic problem is the issue of different service requirements for different classes of customers. Such problems are called Skills-Based Routing and they were already investigated by Gurvich, Armony and Mandelbaum in [2] and Atar, Mandelbaum and Shaikhet in [4]. It is advisable (interesting) to investigate the models of Skill-Based Routing for call centers with an IVR.

Dimensioning the model of call center with an IVR. In our context, the term dimensioning was introduced in Borst, Mandelbaum and Reiman [5]. 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. Increased competition, deregulation and rising customer acquisition costs highlight the importance of both high-quality customer service and effective management of operating costs, and [5] developed a formal framework for this problem. 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 offered load, and $y^*(\cdot)$ is a function that is easily computable. It is of interest to analyze analogous problem for a call center with an IVR, where in addition to staffing and waiting costs, IVR's costs arise.

Chapter 11

Appendix

11.1 Proof of Lemma 6.1.1

PROOF. In the table about inverse Laplace transform we find the following:

$$L^{-1}_{\frac{1}{b+x}}(t) = e^{-bt}, \tag{11.1}$$

and

$$L^{-1}_{\frac{1}{(a+x)^n}}(t) = \frac{t^{n-1}}{(n-1)!}e^{-at}. \tag{11.2}$$

Now let us find the inverse Laplace transform for the function

$$g(x) = \frac{1}{b+x} \cdot \left(\frac{1}{a+x}\right)^n.$$

For this purpose, decompose this function into common fractions

$$\frac{1}{b+x} \cdot \left(\frac{1}{a+x}\right)^n = \frac{B}{b+x} + \frac{A_1}{a+x} + ... + \frac{A_k}{(a+x)^n}. \tag{11.3}$$

If we multiply the left and right sides of equation (11.3) first by $b+x$ and second by $a+x$, we get the following equations:

$$\left(\frac{1}{a+x}\right)^n = \frac{B}{b+x} + \left(\frac{A_1}{a+x} + ... + \frac{A_n}{(a+x)^n}\right)(b+x), \tag{11.4}$$

$$\frac{1}{b+x} = \frac{B}{b+x}(a+x)^k + A_1(a+x)^{k-1} + ... + A_n. \tag{11.5}$$

By substituting $x = -b$ and $x = -a$ in (11.4) and (11.5) respectively, ones obtains

$$B = \frac{1}{(a-b)^n}, \qquad A_n = \frac{1}{b-a}. \tag{11.6}$$

Now let us look more carefully at equation (11.5). By differentiating n-1 times the two sides of the equation and substituting each time $x = -a$, we get that

$$A_j = \frac{(-1)^{k-j}}{(b-a)^{k-j+1}}, \qquad \forall \quad 0 < k < n.$$

Thus, the inverse Laplace transform for the function $g(x)$ is equal to

$$\begin{aligned}
L^{-1}_{g(x)} &= \frac{1}{(a-b)^k}e^{-bt} + \frac{(-1)^{k-1}}{(b-a)^k}e^{-at} + ... - \frac{1}{(b-a)^2}\cdot\frac{t^{k-2}}{(k-2)!e^{-at}} + \frac{1}{b-a}\cdot\frac{t^{k-1}}{(k-1)!e^{-at}} \\
&= \frac{1}{(a-b)^k}e^{-bt} - e^{-at}\left[\frac{1}{(a-b)^k} + ... + \frac{1}{(a-b)^2}\cdot\frac{t^{k-2}}{(k-2)!} + \frac{1}{a-b}\cdot\frac{t^{k-1}}{(k-1)!e^{-at}}\right] \\
&= \frac{1}{(a-b)^k}\left[e^{-bt} - e^{-at}\left(1 + (a-b)t + \frac{(a-b)^2t^2}{2!} + ... + \frac{(a-b)^{k-1}t^{k-1}}{(k-1)!}\right)\right] \\
&= \frac{e^{-at}}{(a-b)^k}\left[e^{(a-b)t} - 1\left(1 + (a-b)t + ... + \frac{(a-b)^{k-1}t^{k-1}}{(k-1)!}\right)\right]
\end{aligned}$$

Finally, the inverse Laplace transform of the function

$$q(x) = \frac{b}{b+x}\left(1 - \left(\frac{a-b}{a+x}\right)^n\right)$$

is equal to

$$\begin{aligned}
L^{-1}_{q(x)} &= be^{-bt} - b\left[e^{-bt} - e^{-at}\left(1 + (a-b)t + \frac{(a-b)^2t^2}{2!} + ... + \frac{(a-b)^{n-1}t^{n-1}}{(n-1)!}\right)\right] \\
&= be^{-at}\left(1 + (a-b)t + \frac{(a-b)^2t^2}{2!} + ... + \frac{(a-b)^{n-1}t^{n-1}}{(n-1)!}\right),
\end{aligned}$$

which yields the required result. □

Bibliography

[1] AFII, Invest in France Agency, "A high potential for growth", available at http://www.investinfrance.org/France/KeySectors/Operations/?p=call$_c entersl = en$

[2] Armony M., Gurvich I. and Mandelbaum A., "Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers", Draft, November 2004, available at http://iew3.technion.ac.il/serveng/References/references.html

[3] Aspect Communications Corporation, "Why Your Customers Hate Your IVR Systems", White paper, 2003, available at http://www.aspect.com/mm/pdf/products/interactive

[4] Atar R., Mandelbaum A. and Shaikhet G. "Queueing systems with many servers: null controllability in heavy traffic", (2005). Submitted to Annals of Applied Probability, available at http://iew3.technion.ac.il/serveng/References/references.html

[5] Borst S., Mandelbaum A. and Reiman M. (2004). Dimensioning large call centers. *Operations Research*, 52(1), 17-34.

[6] Brandt A., Brandt M., Spahl G. and Weber D., "Modelling and Optimization of Call Distribution Systems", Elsevier Science B.V., 1997.

[7] Chernova N., "Theory of Pobability", Lecture notes, 2003, available at http://www.nsu.ru/mmf/tvims/chernova/tv/lec/lec.html

[8] Delorey E., "Correlating IVR Performance and Customer Satisfaction", May 2003, available at http://www.easyivr.com/tech-ivr-applications$_1$08.*htm*

[9] Erlang A.K., "On the rational determination of the number of circuits". In "The life and works of A.K.Erlang." E.Brochmeyer, H.L.Halstrom and A.Jensen, eds.Copenhagen: The Copenhagen Telephone Company, 4.1.1, 4.2.1, 1948.

[10] Feller W., "An Introduction to Probability Theory and Its Applications", Vol.1, Ed.3. John Wiley and Sons, New York.

[11] Gans N., Koole G. and Mandelbaum A. "Telephone Call Centers: Tutorial, Review, and Research Prospects", Invited review paper by *Manufacturing and Service Operations Management (MSOM)*, 5(2), pp. 79-141, 2003.

[12] Garnett O., "Designing a telephone call center with impatient customers", M.Sc. thesis, Technion - Israel Institute of Technology, 1998.

[13] Garnett O., Mandelbaum A. and Reiman M., "Designing a Call Center with Impatient Customers", *Manufacturing and Service Operations Management (MSOM)*, 4(3), pp. 208-227, 2002.

[14] Gilson K.A. and Khandelwal D.K., "Getting more from call centers", The McKinsey Quarterly, Web exclusive, April 2005, available at http://www.mckinseyquarterly.com/article$_page.aspx?ar = 1597L2 = 1L3 = 106$

[15] Grassmann W.K., "Is the fact that the emperor wears no clothes a subject worthy of publication?", *Interfaces*, 16:2, pp. 43-51, 1986.

[16] Grassmann W.K., "Finding the right number of servers in real-world queueing systems", *Interfaces*, 18:2, pp. 94-104, 1988.

[17] Halfin S. and Whitt W. "Heavy-Traffic Limits for Queues with Many Exponential Servers",Operations Research, 29, pp. 567-587, 1981.

[18] Horovitz B., "Whatever happened to customer service?", USA TODAY, posted 25.09.2003, available at http://www.usatoday.com/money/economy/services/2003-09-25-services-frontcover$_x.htm$

[19] Jagerman D.L., "Some properties of the Erlang loss function", Bell Systems Technical Journal, **53:3**, pp. 525-551, 1974.

[20] Jelenkovic P., Mandelbaum A. and Momcilovic P., "Heavy Traffic Limits for Queues with Many Deterministic Servers", QUESTA 47, pp. 53-69, 2004.

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

[22] de Viricout F. and Jennings O.B., "Large-Scale Membership Services", Submitted to *Operations Research*, 2006.

[23] Klenke M., "Endless Loops: Are Your Customers Getting Stuck?", EasyIVR Tech Library, available at
http://www.easyivr.com/tech-ivr-message-on-hold$_1$04.*htm*

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

[25] Mandelbaum A. and Zeltyn S., "Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue", Draft, June, 2005, available at http://iew3.technion.ac.il/serveng/References/references.html

[26] 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*.

[27] 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).

[28] Stolletz R., "Performance Analysis and Optimization of Inbound Call Centers", Springer-Verlag Berlin Heidelberg, 2003.

[29] Srinivasan R., Talim J. and Wang J., "Performance Analysis of a Call Center with Interacting Voice Response Units", contributed paper for the First Madrid Conference on Queueing Theory, Madrid, July 2-5, 2002.

[30] Sze D.Y., "A queueing model for telephone operator staffing", *Operation Research*, 32, pp. 229-249, 1984.

[31] Whitt W., "Understanding the efficiency of multi-server service systems", *Mgmt. Sc.*, 38, 708-723, 1992.

[32] Wolff R.L., "Poisson Arrivals See Time Averages", *Oper. Res.*, vol. 30, pp. 223-231, 1982.

[33] WordiQ.com. Definition of Call center - wordIQ Dictionary and Encyclopedia, http://www.wordiq.com/definition/Call_center