TEL-AVIV UNIVERSITY
RAYMOND AND BEVERLY SACKLER
FACULTY OF EXACT SCIENCES
SCHOOL OF MATHEMATICAL SCIENCES ,
DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH

Scheduling arrivals to queues: a model with no-shows

Thesis submitted in partial fulfillment of requirements for the
M.Sc. degree in the school of Mathematical Sciences,
Tel-Aviv University

By
Sharon Mendel

The research work for this thesis has been carried out at
Tel-Aviv University under the supervision of Prof. Rafael Hassin

July 2006

Contents

1 Abstract

Queueing systems with scheduled arrivals, i.e. appointment systems, are typical for many frontal service systems, e.g. health clinics. When considering the optimal schedule to this class of queueing systems, two competing interests must be considered: if arrivals are scheduled too closely together, long waiting times develop for customers; if they are scheduled too far apart then the server is underutilized.

We consider the problem of obtaining an optimal schedule of n arrivals of independent customers to a single server system with exponential service times, where each customer shows up with probability $p \in (0, 1]$. Our objective is to determine the schedule that minimizes the sum of the expected customers' and server's costs.

Based on a model presented by Pegden and Rosenshine [8], we consider this variation with no-shows. We develop a recursive formulation for computing the value of the expected customers' waiting and server's availability times in the general case of n arrivals. We then use numeric optimization methods to determine an optimal schedule. Following the work of Stein and Cote [10] we also compare this optimal solution to the optimal solution obtained when constraining the schedule to equally spaced arrivals.

2 Introduction and literature review

Queueing systems with scheduled arrivals are known as appointment systems. Appointment systems are typical for many frontal service systems, of which the most studied are outpatient health clinics and services. National health services all over the world are a target of criticism due to mal operation of outpatient clinics. Not only NHS are under such focus, nowadays when many health services have turned into somewhat economical highly competitive markets, various health care providers are under a great deal of pressure on the one hand to improve the quality of service and on the other hand to reduce costs. A better designed appointment system can reduce waiting time for patients and increase the utilization of expensive personnel and other medical resources.

Studies concerning appointment scheduling of outpatient services exist for over 50 years. The first to present an extensive work dealing with this problem was Bailey [1] in 1952. Bailey was among the firsts to state the importance of a well designed appointment policy and presented quantitative tools to improve the performance of appointment-based systems in terms of controlling both customers' waiting time and server's idle time.

Following his work came many others, approaching the topic from different points of view. Different appointment policies have been studied under different scenarios, e.g. different service time distributions, patterns of customer behavior towards appointments etc'. A literature review on appointment policies, focusing on the objective function used and the assumptions made regarding the behavior of the customers, can be found in the work of Mondschein and Weintraub [7] from 2003.

The studies in this field can be classified into three main groups according to the research methodology used: analytical, simulation-based and case study. Cayirli and Veral [2] provide in their paper from 2003 a comprehensive survey of research on appointment scheduling in outpatient services, including the general problem formulation and modeling condition of each research. They do so while classifying the reviewed works by the research methodology used.

A more general bibliography of queues in health care was provided by Preater [9] in 2001. This bibliography lists dozens of papers, all dealing with applications of queueing theory to health care.

One aspect of behavior of customers that influences the overall efficiency of such systems is the phenomenon of no-shows. Despite the extensive work done on appointment systems in health care, the issue of no-shows has been studied very little. The first to address the subject in an analytic approach was Mercer [5] in 1960 and then again in 1973 [6]. In his paper from 1960 he studies the non-equilibrium distribution of the queue length and gives also the results for the equilibrium distribution. He assumes that customers are scheduled to arrive at a queue on a equally spaced schedule, but may arrive at any time after the start of the interval on which they are scheduled, or not arrive at all. In his later work from 1973 he also studies the distribution of the queue length for other models differing in the arrival process and service process, i.e., batch scheduling and general service time. More recent is the work of Kaandrop and Koole [4] from 2006. In this work they define a local search scheduling algorithm and prove that it converges to the optimal schedule in respect to their defined objective function - a weighted sum of the average expected waiting times of customers, idle time of server and tardiness in schedule. The algorithm is flexible and can

incorporate no-shows.

As a base point for our study we use an analytical model presented by Pegden and Rosenshine [8] in 1990 and add to it the assumption that no-shows are allowed. A summary of our results is presented in Section 3. A full descriptions of Pegden and Rosenshine's model and of our variation allowing no-shows can be found in Section 4. Section 5 details our study of small systems designed for two and three customers. In Section 6 we present the optimal schedule for larger systems with no-shows, including an analysis of the results. Others who based their study on this work of Pegden and Rosenshine are Stein and Cote [10] in 1994. They use the base model to study larger systems designed for more customers than Pegden and Rosenshine did, and also study and compare the results obtained when adding a constraint of equally spaced scheduled appointments. The study of the optimal equally spaced schedule with no-shows is presented in Section 7.

When presenting the results of the study, we focus on the range of no-show probabilities found in outpatient services. Empirical studies of no-shows cited in [2] indicate that in many clinics the volume of the phenomenon of no-shows is in the range of 5% - 30% of the scheduled appointments. Some medical web-sites such as `http://www.medicalnewstoday.com` quote figures of about 40% of no-shows. Hence when discussing and analyzing specific cases of no-shows we give a closer attention to this range of no-show probabilities.

3 Summary of results

Our study indicates that no-shows affects the optimal schedule and should be taken into account when designing an appointment system. We study two main types of appointment systems:

1. Systems in which the inter-arrival times between scheduled arrivals are not necessarily equal. I.e, a customer can be scheduled to arrive at any time after, or even at the same time, as the customer scheduled to arrive before her.

2. Systems with equal inter-arrival times between scheduled arrivals.

We study each type of system separately, and also compare the optimal results obtained for each type. The objective we use in our study is minimizing the sum of expected customers' waiting cost and server's availability cost. The main results of the study, in a nutshell, are presented below in Sections 3.1 to 3.3:

3.1 Optimal schedule

- The time between consecutive arrivals has an exponential pattern as a function of the relative cost between the server's availability cost and the customers' waiting cost. (See Sections 5.1, 5.2).

- The optimal schedule schedules the first few customers close together, then the inter-arrival time between scheduled arrival increases, but is nearly constant for the latter scheduled customers, and then again, the inter-arrival time between scheduled arrivals decreases and the last few customers are scheduled to arrive closer together. (See Section 6.2).

- As the showing up probability decreases, the optimal schedule is to schedule appointments closer together. From a certain no-show probability some customers, at the beginning of the schedule, are scheduled to arrive at the same time. As the probability of showing up further decreases, customers at the end of the schedule are also scheduled to arrive together. (See Section 6.2).

- The inter-arrival time between scheduled arrivals in the equally spaced model is approximately the average of the inter-arrival times of the equivalent unrestricted model (a model with the same parameters but with no constraints on the size of the intervals between the scheduled arrivals). (See Section 7.2.1). This inter-arrival time and the showing up probability are somewhat linear dependent. (See Section 7.2.2).

- The optimal equal spacing decreases as the showing probability decreases, i.e. if all the customers who do not show up would have notified in advanced that they are not going to show up, an optimal schedule could have been designed only for those who show up, resulting with a more spacious schedule. (See Section 7.2.3).

3.2 Customers' waiting times

- The expected waiting time for customers who show up can be of order of magnitude higher than the expected service time. The expected waiting time is monotonously increasing from the first scheduled customer to the last. Thus of all scheduled customers, the customer scheduled last has the maximal expected waiting time of all customers, if he shows up. (See Sections 6.2.1, 6.2.2).

- The expected waiting time of customers who show up is affected by the phenomenon of no-shows. As more customers are expected not to show up the expected waiting time for customers who show up decreases.
  Nevertheless these customers are waiting much longer than they would have waited if they were served by a system that serves the same expected number of customers and is designed and operates as an appointment system where all customers show up. (See Sections 6.2.2, 6.2.3).

- The expected waiting times in the equally spaced model behave in a similar manner to the expected waiting times in the unrestricted model. (See Sections 7.2.4, 7.2.5).

3.3 Costs

- The operating cost of a system when the server's availability is significantly higher than the customers' waiting cost, is lower when some of the customers do not show up than when all customers show up. (See Section 6.2.4).

- The operating cost per customer who shows up varies as a function of the probability of showing up. If the customers' waiting cost is significantly higher than the servers' availability cost, the operating cost per customer when all customers show up is lower than the operating cost per customer who shows up when some of the customers do not show up. (See Section 6.2.5).

- The no-shows phenomenon increases the operating cost. Operating a system with no-shows costs more than operating a system that serves the same expected number of customers, but is designed and operates as an appointment system where all customers show up. The effect of no-shows increases as the portion of no-shows increases. (See Section 6.2.6).

- The expected costs for an appointment system with equally spaced arrivals are not significantly different from the costs of an equivalent unrestricted appointment system. (See Section 7.2.6).

4 Formulation of the scheduling problem objective

4.1 Base Model

We study scheduling arrivals to queues with no-shows based on the work of Pegden and Rosenshine [8] and the work of Stein and Cote [10], which used the same model; in this section we present this base model. The objective is to determine the schedule for a fixed number of customers, that minimizes the sum of the expected customers' waiting costs and the expected server's availability cost. It is assumed that customers show up on time and that the server provides a Markovian service and cannot take a vacation; i.e., the server remains available all time, till the last customer leaves the system.

Notation 1

n Number of customers to be scheduled.

c_w Customer's waiting cost per unit of time.

c_s Server's availability cost per unit of time.

$1/\mu$ Mean service time (service exponentially distributed).

x_i Time interval between the scheduled arrival times of the ith and the $(i+1)$st customers.

$\bar{x}$ A vector of intervals between scheduled arrivals $\bar{x} = (x_1, x_2, \ldots, x_{n-1})$.

t_i Time of the ith scheduled arrival, obviously $t_i = t_1 + \sum_{j=1}^{n-1} x_j$.

w_i Expected waiting time of the ith scheduled customer in the queue.

$N(t_i)$ Number of customers in the system just prior to the time of the ith scheduled arrival; thus, $Pr\{N(t_i) = j\}$ is the probability for j customers in the system just prior to the time of the ith scheduled arrival.

The objective is to determine t_1, time of first arrival, and a vector $x^* = (x_1, x_2, \ldots, x_{n-1})$ of intervals between scheduled arrivals, that minimizes the sum of the expected customer waiting cost and the expected server availability cost. Thus the objective is to minimize the function:

$$\Phi^{01}(\bar{x}) = c_w \sum_{i=1}^{n} w_i + c_s \left[t_1 + \sum_{i=1}^{n-1} x_i + w_n + \frac{1}{\mu} \right]. \tag{4.1}$$

It is obvious that in any optimal solution $t_1 = 0$ and that $w_1 = 0$. Moreover, $\frac{c_s}{\mu}$ is a constant term, and so may be omitted. Hence (4.1) can be written as:

$$\Phi^{02}(\bar{x}) = c_w \sum_{i=2}^{n-1} w_i + c_s \sum_{i=1}^{n-1} x_i + (c_s + c_w) w_n. \tag{4.2}$$

Defining the relative server's availability cost:

$$\gamma = \frac{c_s}{c_s + c_w}. \tag{4.3}$$

Dividing (4.2) by $(c_s + c_w)$, the objective function $\Phi^0(\bar{x})$ to be minimized is :

$$\Phi^0(\bar{x}) = (1-\gamma)\sum_{i=2}^{n-1} w_i + \gamma \sum_{i=1}^{n-1} x_i + w_n. \tag{4.4}$$

Pegden and Rosenshine give an explicit optimal solution for two and for three customers, and develop a recursive algorithm, for computing the value of (4.4) in the general case of $n \in \mathbb{N}$ arrivals. They do not manage to give a proof that (4.4) is convex for a any general $n \in \mathbb{N}$, but if for a specific n it is convex, the algorithm guarantees that the global minimum can be found using a gradient search.

Stein and Cote formulate the algorithm defined by Pegden and Rosenshine in a matrix form using transition matrixes. They assume the function is convex for every $n \in \mathbb{N}$ and obtain the optimal spacing for various n number of arrivals using optimization software that implements reduced-gradient search. They also compare the optimal solution for (4.4) to the optimal solution obtained for equally spaced arrivals and to the solution obtained using a $D/M/1$ queue model where the deterministic inter-arrival time is the optimization variable.

4.2 The model with no-shows

We consider a variation where the customers do not necessarily show up, but only a proportion $p \in (0, 1]$ of the customers show up. I.e., assuming that customers are independent, each customer has a probability of p of showing up. If a customer shows up he does so exactly at his scheduled arrival time.

The server does not have any prior knowledge which of the customers will show up, hence he has to remain available at least till t_n, when the last customer is scheduled to arrive.

Customers are served in the order of their scheduled appointments. Since we allow customers to be scheduled to arrive together we define a queue discipline. If a few customers are scheduled to arrive at the same time, the one with the lower scheduled place in line is served first if he arrives. If $t_i = t_j$ and $i < j$, if they both show up, customer i is served before customer j.

Notation 2

p Probability of a customer to show up.

$S(n,p)/M/1$ A modified form of Kendall's queueing notation, based on the notation used by Pegden and Rosenshine. Denoting a queueing system with n scheduled independent customers, each showing up with probability $p \in (0, 1]$ according to a specified schedule $S(n, p)$, to a Markovian service process provided by one server.

w_i^s Expected waiting time in the queue of the ith customer if he shows up.

4.2.1 Objective function

Rewriting (4.1) to reflect the required modifications from the model presented by Pegden and Rosenshine to this variation:

$$\Phi^1(\bar{x}) = c_w \sum_{i=1}^{n} w_i + c_s \left(t_1 + \sum_{i=1}^{n-1} x_i + E\left[\text{server's time after } t_n\right] \right). \tag{4.5}$$

In order to later obtain the optimal solution for (4.5) we first compute w_i and E[server's time after t_n] for this model:

$$w_i = p \cdot E[i\text{th customer's waiting time in queue} \mid \text{customer } i \text{ shows up}] = p \cdot w_i^s. \tag{4.6}$$

$$\begin{aligned} E[\text{server's time after } t_n] &= p \cdot E[\text{server's time after } t_n \mid \text{last customer shows up}] \\ &+ (1-p) \cdot E[\text{server's time after } t_n \mid \text{last customer does not show up}] \\ &= p \cdot \left(w_n^s + \frac{1}{\mu} \right) \\ &+ (1-p) \cdot E[\text{server's time after } t_n \mid \text{last customer does not show up}]. \end{aligned}$$

If the last customer does not show up the server still has to stay in the system and serve all the customers who showed up before t_n and are still in the system at t_n. The amount of time after t_n that would take the server to do so, is equal to the amount of time the last customer would have waited if he had showed up. Thus the expected server's time after t_n is:

$$E[\text{server's time after } t_n] = p \cdot \left(w_n^s + \frac{1}{\mu} \right) + (1-p) \cdot w_n^s = w_n^s + \frac{p}{\mu}. \tag{4.7}$$

Substituting (4.7) and (4.6) in (4.5) we obtain:

$$\Phi^2(\bar{x}) = c_w p \sum_{i=1}^{n} w_i^s + c_s \left[t_1 + \sum_{i=1}^{n-1} x_i + w_n^s + \frac{p}{\mu} \right] \tag{4.8}$$

The objective function as presented in (4.8) can be simplified, since $\frac{c_s \cdot p}{\mu}$ is a constant term, and so may be omitted. Also, in any optimal solution $t_1 = 0$ and $w_1^s = 0$, hence (4.8) can be rewritten as:

$$\Phi^3(\bar{x}) = c_w p \sum_{i=2}^{n-1} w_i^s + c_s \sum_{i=1}^{n-1} x_i + (c_s + c_w p) w_n^s. \tag{4.9}$$

We redefine the relative server's availability cost for this model as:

$$\tilde{\gamma} = \frac{c_s}{c_s + c_w p}. \tag{4.10}$$

Nevertheless, when describing a specific case of the system it is more intuitive to evaluate the relative cost as defined in (4.3). As mentioned by Stein and Cote, the ratio γ is a measure of the relative importance of the two costs. As γ increases the cost of the server is considered

more important whereas as γ decreases more emphasis is placed on minimizing the cost of customer waiting. Note that $\tilde{\gamma}$ can be obtained from γ and p by:

$$\tilde{\gamma} = \frac{\gamma}{\gamma + p \cdot (1 - \gamma)}. \tag{4.11}$$

By dividing (4.9) by $(c_s + c_w p)$, the objective function $\Phi(\bar{x})$ to be minimized can be written as:

$$\Phi(\bar{x}) = (1 - \tilde{\gamma}) \sum_{i=2}^{n-1} w_i^s + \tilde{\gamma} \sum_{i=1}^{n-1} x_i + w_n^s. \tag{4.12}$$

4.2.2 Recursive Expression for w_i^s

Our objective is to determine $x^* = (x_1, x_2, \ldots, x_{n-1})$ that minimizes (4.12). Hence we develop a general expression for w_i^s, the expected waiting time of the ith customer if he shows up, as a function of x^*.

Since the service is Markovian, the expected waiting time in the queue for a customer who shows up depends upon the number of customers he encounters when arriving at the system, i.e.,

$$w_i^s \equiv \sum_{j=1}^{i-1} \left(\frac{j}{\mu} \cdot Pr\{N(t_i) = j\} \right). \tag{4.13}$$

The probability that there are j customers in the system at t_i depends upon whether or not the $(i-1)$th customer shows up:

$$\begin{aligned} Pr\{N(t_i) = j\} &= p \cdot Pr\{N(t_i) = j | (i-1)\text{th customer shows up}\} \\ &+ (1-p) \cdot Pr\{N(t_i) = j | (i-1)\text{th customer does not show up}\}. \end{aligned} \tag{4.14}$$

The probability that there are $0 < j < i$ customers in the system just prior to t_i can be computed from the state probabilities at time t_{i-1}. Thus, for $1 \leq j < i$ and $i \geq 2$ (4.14) can be obtained by:

$$\begin{aligned} Pr\{N(t_i) = j\} &= p \cdot \sum_{k=0}^{i-j-1} (Pr\{N(t_{i-1}) = j + k - 1\} \cdot Pr\{k \text{ departures between } t_{i-1} \text{ and } t_i\}) \\ &+ (1-p) \cdot \sum_{k=0}^{i-j-2} (Pr\{N(t_{i-1}) = j + k\} \cdot Pr\{k \text{ departures between } t_{i-1} \text{ and } t_i\}). \end{aligned}$$

Since the service is Markovian the probability of k departures between the $(i-1)$st and the ith scheduled arrivals (assuming there are at least k customers in the system at t_{i-1}), is the

probability of exactly k events in a Poisson process with the rate of μ. Thus:

$$\begin{aligned}
Pr\{N(t_i)=j\} &= p\cdot\sum_{k=0}^{i-j-1} Pr\{N(t_{i-1})=j+k-1\}\cdot\frac{(\mu x_{i-1})^k}{k!}e^{-\mu x_{i-1}} \\
&+ (1-p)\cdot\sum_{k=0}^{i-j-2} Pr\{N(t_{i-1})=j+k\}\cdot\frac{(\mu x_{i-1})^k}{k!}e^{-\mu x_{i-1}} \\
&= p\cdot\sum_{k=0}^{i-j-1} Pr\{N(t_{i-1})=j+k-1\}\cdot\frac{(\mu x_{i-1})^k}{k!}e^{-\mu x_{i-1}} \\
&+ (1-p)\cdot\sum_{k=1}^{i-j-1} Pr\{N(t_{i-1})=j+k-1\}\cdot\frac{(\mu x_{i-1})^{k-1}}{(k-1)!}e^{-\mu x_{i-1}} \\
&= \sum_{k=1}^{i-j-1} Pr\{N(t_{i-1})=j+k-1\}\cdot\frac{(\mu x_{i-1})^{k-1}}{(k-1)!}e^{-\mu x_{i-1}}\cdot\Big(\frac{p\mu x_{i-1}}{k}+1-p\Big) \\
&+ p\cdot Pr\{N(t_{i-1})=j-1\}\cdot e^{-\mu x_{i-1}} \quad 1\le j<i, \quad i\ge 2. \qquad (4.15)
\end{aligned}$$

The last term in (4.15) drives from $k=0$.

Similarly, the probability that the system is empty just prior to t_i, depends upon whether or not the $(i-1)$st customer shows up. Suppose that the $(i-1)$st customer shows up and finds $k-1$ customers in the system. For the system to be empty prior to t_i, the service times of these $k>0$ customers must be such that their sum is less than the time between the $(i-1)$st and ith scheduled arrivals. If the $(i-1)$st customer does not show up, for the system to be empty prior to t_i, either the service times of $k>0$ customers that are in the system prior to t_{i-1} is such that their sum is less than the time between the $(i-1)$st and ith scheduled arrivals, or the system is empty already prior to t_{i-1}. Thus the probability that the system is empty just prior to t_i, where $2\le i$, can be obtained by:

$$\begin{aligned}
Pr\{N(t_i)=0\} &= p\cdot\sum_{k=1}^{i-1} Pr\{N(t_{i-1})=k-1\} \\
&\quad\times Pr\{\text{time between } t_{i-1} \text{ and } t_i \text{ is sufficient for at least } k \text{ departures}\} \\
&+ (1-p)\cdot\sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \\
&\quad\times Pr\{\text{time between } t_{i-1} \text{ and } t_i \text{ is sufficient for at least } k \text{ departures}\}.
\end{aligned}$$

Using the Markovian attribute of the service for $i \geq 2$ we obtain:

$$
\begin{aligned}
Pr\{N(t_i)=0\} &= p \cdot \sum_{k=1}^{i-1} Pr\{N(t_{i-1})=k-1\} \cdot \sum_{l=k}^{\infty} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}} \\
&+ (1-p) \cdot \sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \cdot \sum_{l=k}^{\infty} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}} \\
&= p \cdot \sum_{k=1}^{i-1} Pr\{N(t_{i-1})=k-1\} \cdot \left(1 - \sum_{l=0}^{k-1} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}}\right) \\
&+ (1-p) \cdot \sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \cdot \left(1 - \sum_{l=0}^{k-1} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}}\right) \\
&= p \cdot \sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \cdot \left(1 - \sum_{l=0}^{k} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}}\right) \\
&+ (1-p) \cdot \sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \cdot \left(1 - \sum_{l=0}^{k-1} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}}\right) \\
&= \sum_{k=0}^{i-2} Pr\{N(t_{i-1})=k\} \cdot \left(1 - \sum_{l=0}^{k-1} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}} - p \cdot \frac{(\mu x_{i-1})^k}{k!} e^{-\mu x_{i-1}}\right).
\end{aligned}
\tag{4.16}
$$

5 $S(n,p)/M/1 \quad n=2,3$

We first consider $S(n,p)/M/1$ for relatively small appointment systems, of $n=2,3$ customers. Forming the expected waiting time in these small appointment systems is relatively easy, and demonstrates the idea behind the somewhat cumbersome recursive formulation presented in Section 4.2.2.

5.1 $S(2,p)/M/1$

Consider two independent arrivals, such that each occurs with probability $p \in (0,1]$.

Based on (4.13) and (4.15) the expected waiting time is:

$$w_2^s = \frac{1}{\mu} Pr\{N(t_2)=1\} = \frac{p}{\mu} e^{-\mu x_1}. \tag{5.1}$$

Hance the objective (4.12) becomes:

$$\Phi(\bar{x}) = \tilde{\gamma} x_1 + \frac{p}{\mu} e^{-\mu x_1}. \tag{5.2}$$

The value of x_1 that minimizes (5.2) must satisfy:

$$\frac{\mathrm{d}\Phi(\bar{x})}{\mathrm{d}x_1} = \tilde{\gamma} - pe^{-\mu x_1} = 0.$$

Hence:

$$x_1 = -\frac{1}{\mu} \ln \frac{\tilde{\gamma}}{p}. \tag{5.3}$$

Since $x_1 \geq 0$ and $\mu > 0$, the probability p that a customer shows up, must satisfy:

$$p \geq \tilde{\gamma}. \tag{5.4}$$

By substituting $\tilde{\gamma}$ as defined in (4.11) we obtain:

$$p \geq \frac{-\gamma + \sqrt{\gamma^2 + 4\gamma(1-\gamma)}}{2(1-\gamma)}. \tag{5.5}$$

Or in terms of costs, by substituting $\tilde{\gamma}$ as defined in (4.10):

$$p \geq \frac{-c_s + \sqrt{{c_s}^2 + 4c_s c_w}}{2c_w}. \tag{5.6}$$

If the condition in (5.5) is not satisfied, the two customers are scheduled to arrive together on time $t_1 = t_2 = 0$, we denote this critical value of γ as γ_0^2. Intuitively, this condition is not satisfied when p, the showing up probability, is relatively low. So scheduling the two customers to arrive together will ensure zero server's availability cost due to server's waiting times for customers; and yet we are able to schedule the two customers to arrive together without significantly increasing the expected customers' waiting cost, since the probability of customers actually showing up is relatively low.

Obviously, for the case of $p=1$ we get to the same solution as Pegden and Roshenshine did. The schedule for $\mu = 1$ for various values of p and γ is presented in Figures 1 and 2. In these graphs, the relative server's availability cost is stated on the X-axis, denoted by γ, and the scheduled arrivals on the Y-axis, denoted by T.

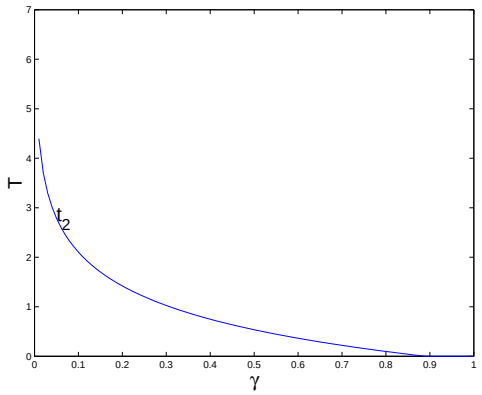

Figure 1: $S(2, 0.90)/M/1$ Schedule

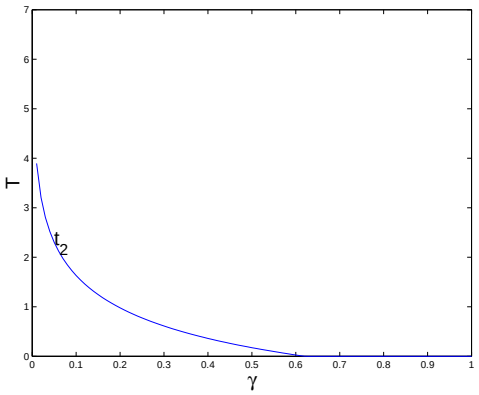

Figure 2: $S(2, 0.70)/M/1$ Schedule

5.2 $S(3, p)/M/1$

Another relatively small case is the case of three independent customers, each showing up with probability $p \in (0, 1]$.
Based on (4.13) the expected waiting times are:

$$w_2^s = \frac{1}{\mu} Pr\{N(t_2) = 1\} = \frac{p}{\mu} e^{-\mu x_1}.$$

$$w_3^s = \frac{1}{\mu} Pr\{N(t_3) = 1\} + \frac{2}{\mu} Pr\{N(t_3) = 2\}. \tag{5.7}$$

To obtain w_3^s we compute the probabilities $Pr\{N(t_3) = k\}$ where $k = 1, 2$ based on (4.15):

$$\begin{aligned} Pr\{N(t_3) = 1\} &= Pr\{N(t_3) = 1 | N(t_2) = 0\} Pr\{N(t_2) = 0\} \\ &+ Pr\{N(t_3) = 1 | N(t_2) = 1\} Pr\{N(t_2) = 1\} \\ &= pe^{-\mu x_2}[(1-p) + p(1 - e^{-\mu x_1})] \\ &+ [p\mu x_2 e^{-\mu x_2} + (1-p)e^{-\mu x_2}] pe^{-\mu x_1}. \qquad (5.8) \\ Pr\{N(t_3) = 2\} &= p \cdot pe^{-\mu(x_1+x_2)} = p^2 e^{-\mu(x_1+x_2)}. \qquad (5.9) \end{aligned}$$

After substituting (5.8) and (5.9) in (5.7):

$$w_3^s = \frac{pe^{-\mu x_2} + pe^{-\mu(x_1+x_2)} + p^2 e^{-\mu(x_1+x_2)} \mu x_2}{\mu}.$$

Thus (4.12) becomes:

$$\Phi(\bar{x}) = (1 - \tilde{\gamma}) \frac{pe^{-\mu x_1}}{\mu} + \tilde{\gamma}(x_1 + x_2) + \frac{pe^{-\mu x_2}}{\mu} (1 + e^{-\mu x_1} + p\mu x_2 e^{-\mu x_1}). \tag{5.10}$$

Again, for the case of $p = 1$ we arrive, as expected, to the same result as Pegden and Roshenshine did.

The values of x_1 and x_2 which minimize (5.10) must satisfy $\frac{\partial \Phi(\bar{x})}{\partial x_i} = 0 \quad i = 1, 2$. Taking the partial derivatives of $\Phi(\bar{x})$ with respect to x_1 and x_2 yields:

$$\begin{aligned} \frac{\partial \Phi(\bar{x})}{\partial x_1} &= -(1 - \tilde{\gamma}) pe^{-\mu x_1} + \tilde{\gamma} - pe^{-\mu x_2}(e^{-\mu x_1} + p\mu x_2 e^{-\mu x_1}) = 0. \qquad (5.11) \\ \frac{\partial \Phi(\bar{x})}{\partial x_2} &= \tilde{\gamma} - pe^{-\mu x_2}(1 + e^{-\mu x_1} + p\mu x_2 e^{-\mu x_1} - pe^{-\mu x_1}) = 0. \qquad (5.12) \end{aligned}$$

Rewriting Equations (5.11) and (5.12) we obtain:

$$\tilde{\gamma} = p\left[(1-\tilde{\gamma})e^{-\mu x_1} + e^{-\mu(x_1+x_2)}(1+p\mu x_2)\right]. \tag{5.13}$$

$$\tilde{\gamma} = p\left[e^{-\mu x_2} + e^{-\mu(x_1+x_2)}(1+p\mu x_2 - p)\right]. \tag{5.14}$$

Equating the two right-hand side expressions of (5.13) and (5.14) we obtain:

$$\begin{aligned}(1-\tilde{\gamma})e^{-\mu x_1} + e^{-\mu(x_1+x_2)}(1+p\mu x_2) &= e^{-\mu x_2} + e^{-\mu(x_1+x_2)}(1+p\mu x_2 - p)\\ (1-\tilde{\gamma})e^{-\mu x_1} &= e^{-\mu x_2} - pe^{-\mu(x_1+x_2)}\\ e^{-\mu x_2} &= (1-\tilde{\gamma})\frac{e^{-\mu x_1}}{1-pe^{-\mu x_1}} \equiv A(x_1). \end{aligned}\tag{5.15}$$

Thus

$$x_2 = -\frac{1}{\mu}\ln\left[(1-\tilde{\gamma})\frac{e^{-\mu x_1}}{1-pe^{-\mu x_1}}\right] = -\frac{1}{\mu}\ln A(x_1). \tag{5.16}$$

In order to find the optimal x^*, we substitute x_2 as defined in (5.16) in either (5.13) or (5.14), and solve the resulting equation for x_1.
The solution obtained is a global minimum since the objective function (5.10) is convex.

By substituting for x_2 in (5.14) we obtain:

$$\tilde{\gamma} = p\left[e^{\ln A(x_1)} + e^{-\mu x_1 + \ln A(x_1)}(1 - p\ln A(x_1) - p)\right]. \tag{5.17}$$

That is:

$$f(x_1) \equiv pA(x_1)\left[1 + e^{-\mu x_1}(1 - p\ln A(x_1) - p)\right] - \tilde{\gamma} = 0. \tag{5.18}$$

We must constrain our solution by $x_i \geq 0 \quad i = 1, 2$. Hence if we obtain $x_1 < 0$ we force the feasible optimal solution $x_1 = 0$ and obtain x_2 by substituting $x_1 = 0$ in (5.10):

$$f(x_2) \equiv \Phi(x_1 = 0, x_2) = (1-\tilde{\gamma})\frac{p}{\mu} + \tilde{\gamma}x_2 + \frac{pe^{-\mu x_2}}{\mu}(2 + p\mu x_2). \tag{5.19}$$

The value of x_2 that minimizes (5.19) must satisfy:

$$\frac{\mathrm{d}f(x_2)}{\mathrm{d}x_2} = \tilde{\gamma} - pe^{-\mu x_2}(2 + p\mu x_2 - p) = 0. \tag{5.20}$$

In the same manner applied for x_1, if we obtain $x_2 < 0$ we force the feasible optimal solution $x_2 = 0$.

The equations for x_1 and x_2 can be simplified by normalizing, without lost of generality, $\mu = 1$. The solution found for this simplified case, is actually the solution for μx_1 and μx_2 in the general case; hence x_1 and x_2 can be found for any value of μ, based on the solution for the simplified case of $\mu = 1$.

Apparently a closed-form solution does not exist, thus the solution must be obtained numerically. We obtain the solution numerically for various values of $p \in (0, 1]$ and $\gamma \in (0, 1]$ by using the Newton-Raphson method, for finding approximations to the zeros of a real-valued function. For details see Appendix A.1. The x_1 and x_2 found define the optimal schedule by using the definition of t_i from Notation 1.

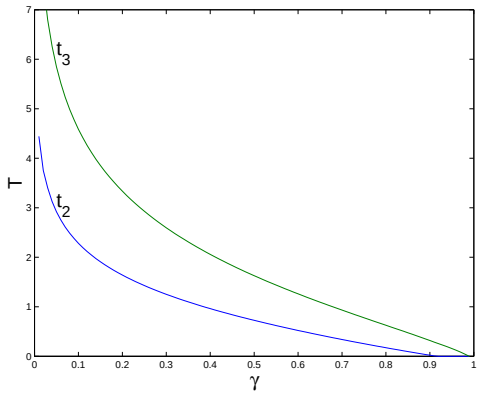

Figure 3: $S(3, 0.90)/M/1$ Schedules

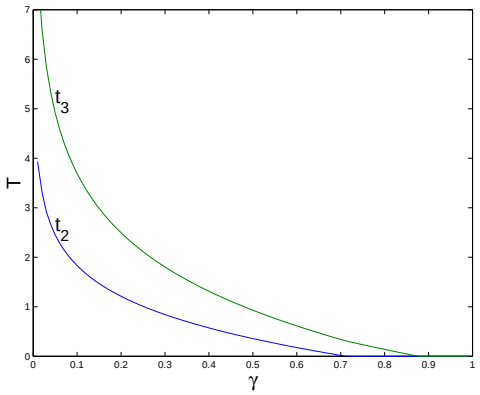

Figure 4: $S(3, 0.70)/M/1$ Schedules

The results for various values of p and γ are presented in Figures 3 and 4. As before, in these graphs the relative server's availability cost is stated on the X-axis, denoted by γ, and the scheduled arrivals on the Y-axis, denoted by T.

We observe that if γ, the relative server's availability cost, is low, the three customers are scheduled to arrive on distinct times. As γ increases, the inter-arrival time between the first and the second customer decreases till $x_1 = 0$ and so $t_1 = t_2 = 0$ and the first two customers are scheduled to arrive together. We denote this critical value of γ as γ_{01}^3. Finally, for high values of γ also $x_2 = 0$, and all three customers are scheduled to arrive together at $t_1 = t_2 = t_3 = 0$. We denote this critical value of γ as γ_{02}^3. This behavior of the optimal solution is somewhat intuitive. All three scheduled customers have a probability of not showing up, and they make their decision independently. By scheduling the first two customers closer together we reduce the server's idle time in case one of them, or more, does not show up; even if both the first two scheduled customers do show up, there is a chance the third scheduled customer would not show up at t_3 as scheduled, but by then, or soon after, the server will finish serving them both and be "dismissed". Since the last customer is not scheduled so close to the first two, even if all three show up, this schedule reduces his waiting. If we would set the last two customers closer together (in oppose to setting the first two together), we would risk the server waiting long idle for them to show up in case the first customer does not show up, and then if both of these last customers show up, the server will have to stay long after t_3 to serve them both.

Comparing the results of $S(3, p)/M/1$ with those of $S(2, p)/M/1$, for $\gamma \geq \gamma_{01}^3$, i.e. analyzing whether from this point the three customers model behaves like a two customers model, i.e. x_2 follows (5.3), reveals that this is not the case. Figure 5 presents for various values of the showing up probability p the critical values of the relative server's availability cost γ for $S(3, p/M/1)$ and for $S(2, p/M/1)$. Examining Figure 5 we note that $\gamma_0^2 < \gamma_{01}^3 < \gamma_{02}^3 \quad \forall p$, i.e. in a system of two customers, the two are scheduled to arrive together for server's relative availability costs for which in a system of three customers all three are scheduled in distinguished times. This can be explained intuitively by the fact, that in $S(3, p)/M/1$ we need to be more careful before scheduling customers to arrive together, because by doing so we potentially cause waiting times for two customers and not just for one as in $S(2, p)/M/1$.

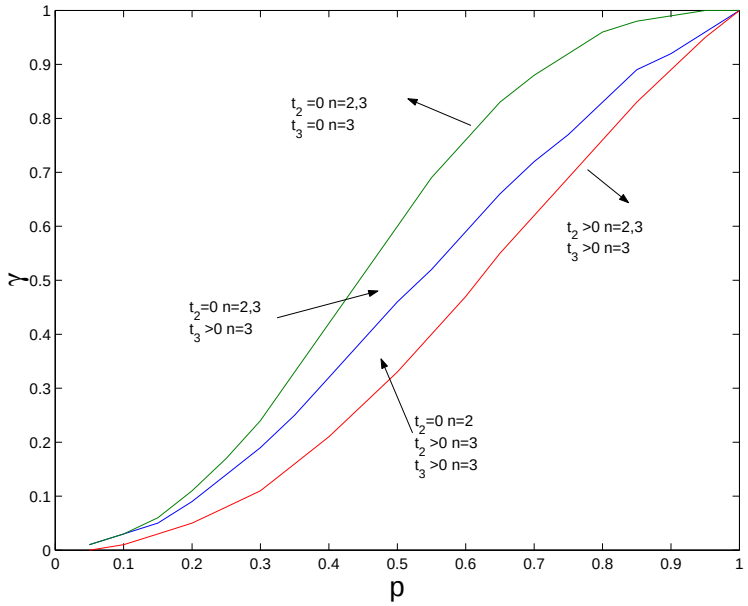

Figure 5: Critical γ values

6 $S(n,p)/M/1 \quad n \geq 4$

6.1 General solution method

We next consider queueing systems where there are at least four independent customers to schedule. To form the expected waiting time of a customer who shows up, to be substituted in the objective function (4.12), we use the recursive formulation presented in Section 4.2.2.

To form the objective function for a given number of customers n, we produce a probability matrix $P(\bar{x})$ of the size of $n \times n$ where $P(\bar{x})_{ij} = Pr\{N(t_i) = j|\bar{x}\} \quad \forall 0 \leq j < i \leq n$, i.e. the probability for j customers in the system just prior to the time of the ith scheduled arrival, for a given solution $\bar{x}$.

As in $S(3,p)$, we must obtain the optimal solution numerically using an approximation algorithm. To obtain the optimal solution for a given $n \in \mathbb{N}$ we assume that (4.12) is convex. As stated by Pegden and Rosenshine the objective function (4.4) is believed to be convex despite the appearance of non-convex terms. However a general, most likely inductive, proof to this assumption, has not been found, even though looked by them and by others. Showing the convexity for a given number of customers $n \geq 5$ is somewhat preposterous, since the number of terms in w_n^s grows very rapidly with the dimension of the solution space, and the additional terms become more unwieldy as well. Each member in the sum that forms the no-shows model objective function (4.12) is a linear combination of a member of (4.4). Hence (4.4) is convex if and only if (4.12) is convex. Based on this assumption we obtain the optimal solution by applying Sequential Quadratic Programming on the objective function, using Matlab optimization toolbox. For further details see Appendix A.2.

6.2 Analysis of the optimal solution

The optimal solutions found for $n \geq 4$ conforms with the findings of Stein and Cote. For $p = 1$ they find that instead of a continued trend of wider intervals between successive scheduled arrivals which may have been inferred from Pegden and Rosenshine, the optimal interval width increases for the first few customers, then stays almost constant till it decreases for the last few customers. We find that in the model with no-shows this phenomenon expands, and as the probability of showing up decreases and the relative server's availability cost increases, not only the first few customers are scheduled to arrive together but also so do the last few customers. The latter phenomenon, of the last customers scheduled to arrive together, occurs for relatively low showing up probabilities. For example, in a system with ten customers we begin to observe this pattern only for $p \leq 0.30$ and $\gamma \geq 0.90$. Figures 6 to 9 present the optimal spacing between scheduled arrivals for various values of γ in a system of ten customers. In these graphs the interval numbers are noted on the X-axis and the spacing between scheduled arrivals are noted on the Y-axis. Point (i, x_i) on a certain γ graph in these figures is the optimal value of x_i in the relevant model, i.e., the scheduled inter-arrival time between customer i and customer $i+1$ for a system of $S(10,p)/M/1$ with p and γ as stated in the figure.

The intuitive explanation to these spacings is similar to the one given by Stein and Cote. In the probabilistic steady state approximately equally spaced arrivals are scheduled, however the last few customers are scheduled to arrive closer, or even together, to avoid the

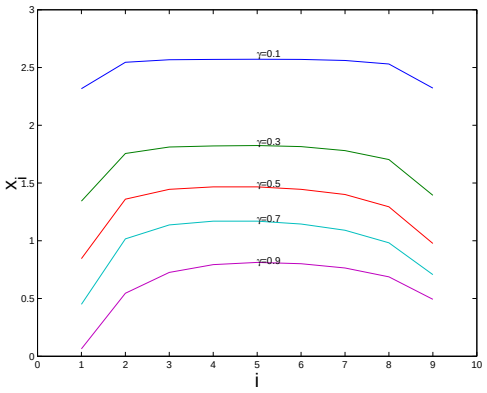

Figure 6: Spacing $S(10, 0.90)/M/1$

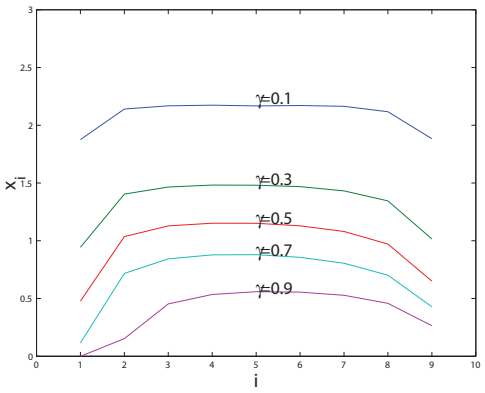

Figure 7: Spacing $S(10, 0.70)/M/1$

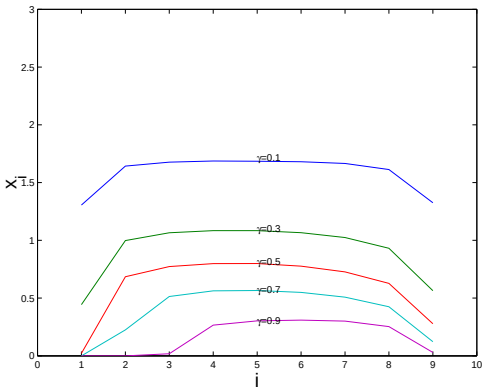

Figure 8: Spacing $S(10, 0.50)/M/1$

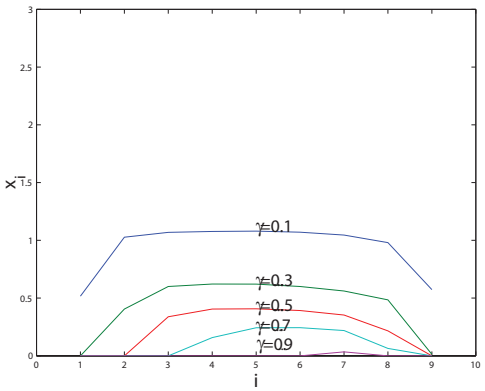

Figure 9: Spacing $S(10, 0.30)/M/1$

server being idle while only a few customers remain to arrive. The scheduling of the first few customers to arrive close or even together fits what is known as Bailey's Low [1] which recommends to schedule the first customers together in order to later reduce the server's idle time.

For instance, examining the optimal spacing between scheduled arrivals for $\gamma = 0.70$ in Figure 9 where $n = 10$ and $p = 0.30$, we note that the first four customers are scheduled to arrive together since $x_1 = x_2 = x_3$, then the 6th, 7th and 8th customers are scheduled to arrive on more or less equal inter-arrival times as $x_5 \approx x_6 \approx x_7$, and then again the inter-arrival time between the scheduled arrivals of the 8th, 9th and 10th customers decrease: $x_7 > x_8 > x_9$. For $\gamma = 0.90$ on the same figure, there are only two appointment times, the first seven customers are scheduled to arrive together at t_1 and then later at $t_8 > 0$ are scheduled the last three customers.

6.2.1 Resulting customers' expected waiting times

Though our objective function (4.12) is a combination of both customers' expected waiting times and server's expected availability time, it could be, for practical reasons, of special interest to note the resulting customers' expected waiting times. If the expected customers waiting times are too high, the system's management may consider reevaluating the relative cost, so greater consideration would be given to customers' waiting cost.

Figures 10 to 13 present expected waiting times of customers who show up for five and ten customers for various showing up probabilities. In these graphs the relative server's availability cost is stated on the X-axis, noted by γ, and the expected waiting times are stated on the Y-axis.

We note that the variance of the expected waiting times of customers who show up, is quite high, and that the expected waiting time of a customer who shows up increases as his scheduled place in line increases, i.e. $w^s_{i+1} > w^s_i \quad \forall i = 1, ..., n-1$. This could be explained by the queue's discipline as defined in Section 4.2. If $t_i = t_j$ and $i < j$, customer i is served first if he arrives. In this case, if customer j arrives as well he waits with probability of 1.00. As more customers are scheduled to arrive together the more they wait if more than one of them arrives.

We also note that the expected waiting time for customers who show up decreases as the probability of no-shows increases, i.e. for a given n and γ, $w^s_i \quad \forall i = 1, ..., n$ decreases as p decreases. This indicates that the optimal schedule exploits the no-shows phenomenon to decrease the expected waiting time of customers who do show up. Since less customers are expected to arrive, the customers who show up are less likely to wait due to long service times of earlier customers.

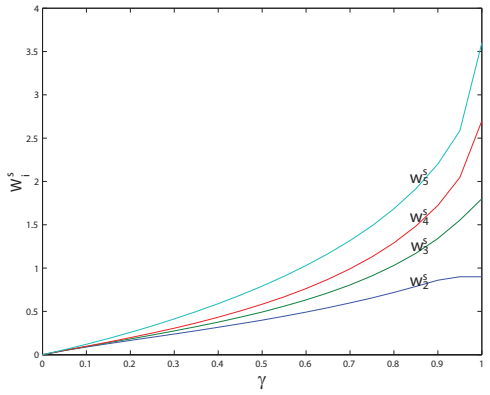

Figure 10: Waiting $S(5, 0.90)/M/1$

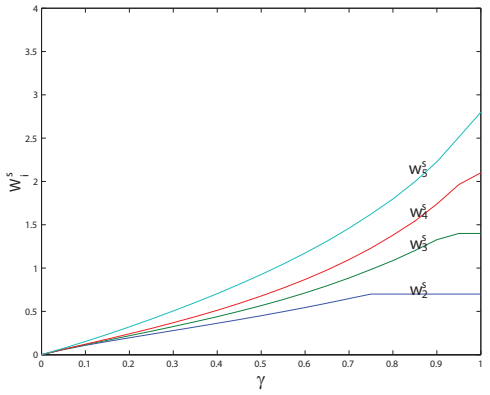

Figure 11: Waiting $S(5, 0.70)/M/1$

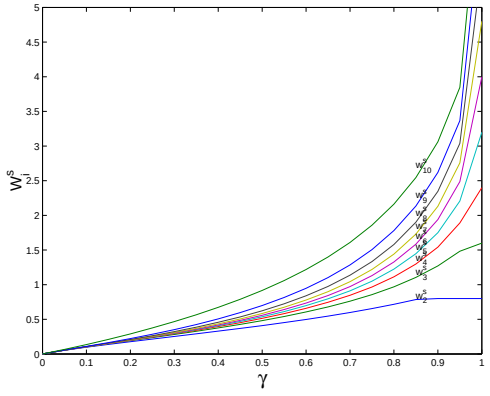

Figure 12: Waiting $S(10, 0.80)/M/1$

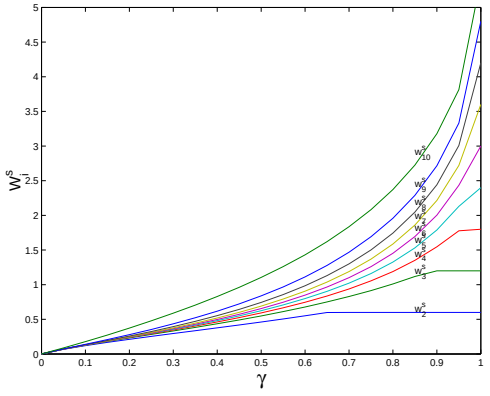

Figure 13: Waiting $S(10, 0.60)/M/1$

6.2.2 Average and maximal customers' expected waiting times

Figures 14 to 17 present the average and the maximal expected waiting time of a customer who shows up, for five and ten customers for various showing up probabilities. In these figures, w_{mean} refers to $\frac{1}{n}\sum_{i=1}^{n} w_i^s$, and w_{max} refers to $\max\{w_i^s\}_{i=1}^{n}$. As mentioned in Section 6.2.1 the maximum expected waiting time of a customer who shows up is obtained by the last customer as $\max\{w_i^s\}_{i=1}^{n} = w_n^s$. We note that for large values of γ, i.e., low relative customers' waiting time cost, the maximum expected waiting time of a customer who shows up is up to twice the average expected waiting time and more than five times the average service time (which is in our model considered to be $\frac{1}{\mu} = 1$).

In Section 6.2.1 we note a decrease in $w_i^s, \quad \forall i = 1, ..., n$, as p decreases. Here we note, as expected, that the average and maximal expected waiting time for customers who show up, decreases as the probability of no-shows increases.

6.2.3 Quantifying the impact of no-shows on customers' expected waiting times

Studying the impact of the no-shows phenomenon we compare the expected average and maximal waiting times of systems where all customers show up to systems with no-shows with the same expected number of customers who show up. We compare the expected waiting times for customers who show up of $S(n, p < 1.0)/M/1$ models to the ones of $S(n', 1.0)/M/1$

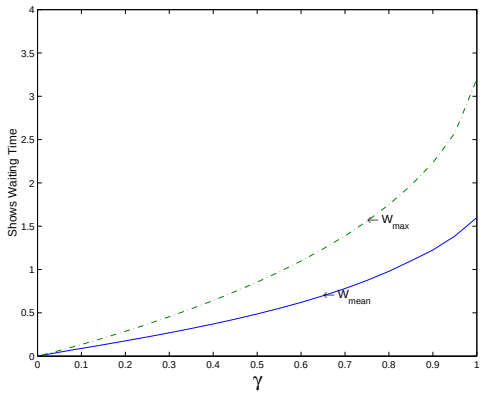

Figure 14: Waiting $S(5, 0.80)/M/1$

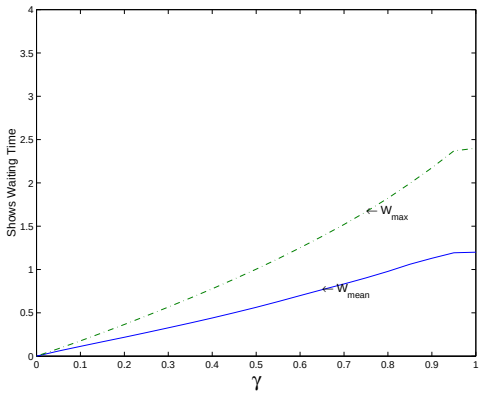

Figure 15: Waiting $S(5, 0.60)/M/1$

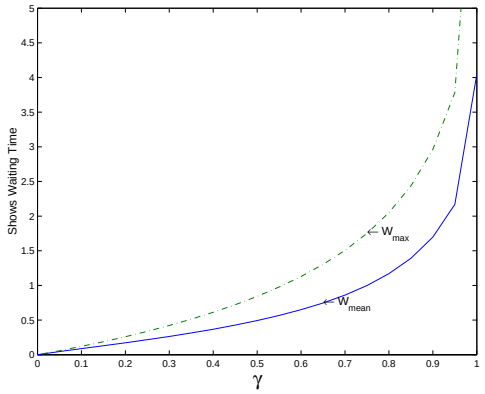

Figure 16: Waiting $S(10, 0.90)/M/1$

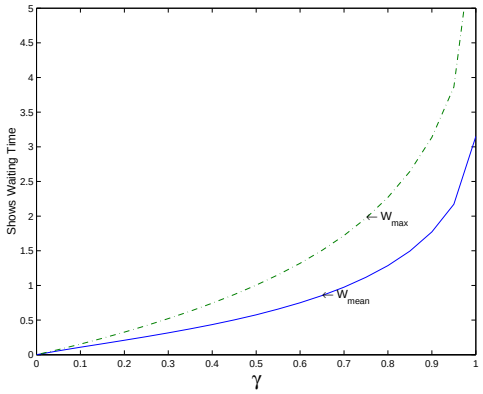

Figure 17: Waiting $S(10, 0.70)/M/1$

models where $np = n'$. If all the customers who do not show up would have notified in advanced that they are not going to show up, an optimal schedule could have been designed only for those who show up, resulting with smaller expected waiting times for the customers who show up.

Each of the Figures 18 to 21 presents a comparison between the expected waiting times for customers who show up of a $S(n, p < 1.00)/M/1$ model and the same expected waiting times of the equivalent $S(n', 1.00)/M/1$ model where $np = n'$. In these graphs the relative server's availability cost is stated on the X-axis, noted by γ, and the expected waiting times measures of customers who show up are stated on the Y-axis. The waiting times of the models with no-shows are drawn in solid lines, whereas the ones of the models where all customers show up are drawn in a dashed lines.

We note that as p decreases, i.e the probability of customers' not showing up increases, the impact of the no-shows increases. If a large portion of customers do not show up, the expected waiting time of the customers who do show up, is much higher than the expected waiting time they would have had if the scheduled was originally designed only for them.

Measuring the differences between the expected waiting time measures of the models with no-shows to those of models where all customers show up, we note that the influence of the no-shows phenomenon is greater for smaller relative server's availability cost. Meaning that the impact of no-shows is more severe when the importance given to customers waiting times, in order to minimize them, is higher.

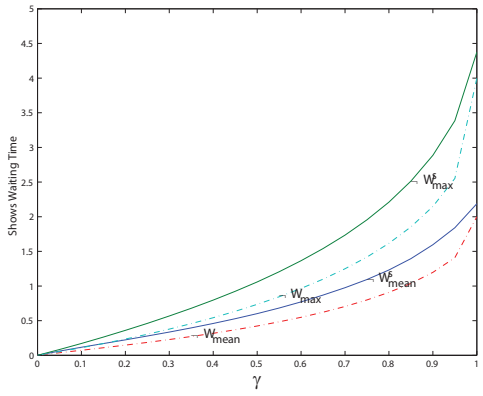

Figure 18: Waiting $S(8, 0.625)/M/1$

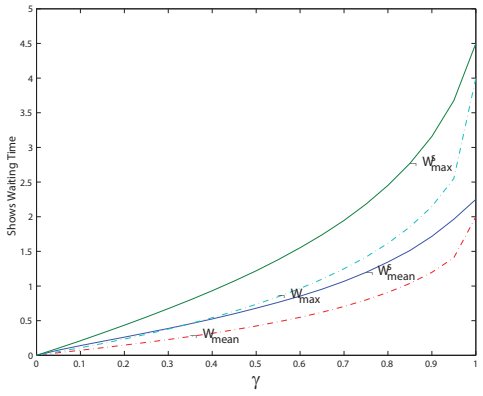

Figure 19: Waiting $S(10, 0.50)/M/1$

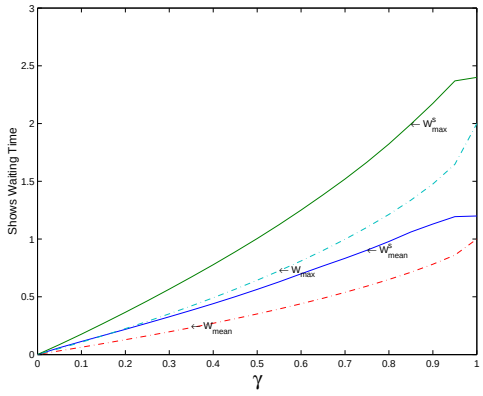

Figure 20: Waiting $S(5, 0.60)/M/1$

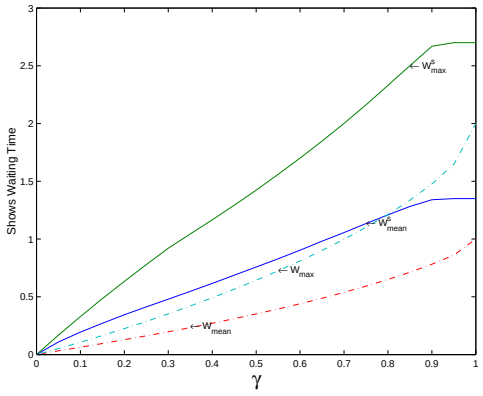

Figure 21: Waiting $S(10, 0.30)/M/1$

Table 1 presents by how many percents the average expected waiting time of customers who show up in a model with no-shows is greater than the same measure of an equivalent model where all customers show up. We see that for a system designed for ten customers when 20% of the customers are expected not to show up, i.e. $n' = np = 8$, the average expected waiting time for customers who show up increases by 15% already for relative customers' waiting cost of 0.05 ($\gamma = 0.95$) and goes up to 25% when $1-\gamma = 0.95$ ($\gamma = 0.05$). When the expected number of customers is three, for a system designed for five customers of which 60% show up, the impact on the average expected waiting time varies from 20% to 85%. When 70% of ten customers are expected not to show up the average expected waiting time doubles when the relative cost of customers' exceeds 0.35 and the impact increases as $1-\gamma$ increases (γ decreases). The impact of no-shows on the maximal expected waiting time measured in the same manner is nearly equivalent, thus not detailed.

6.2.4 Objective function's value and cost of waiting times

Studying the optimal values of the objective function, we note that for high γ, the minimum value of the objective function for $p = 1$, where all customers show up, is higher than for some other showing up probabilities. If $p < 1$ the expected number of customers is smaller, hence for some $p < 1$, the optimal schedule, determined according to the expected number of customers, results in a smaller operating cost than $p = 1$. This is not always true and for small values of γ the cost is higher for $p < 1$ than for $p = 1$. This is because

	$np = 3$			$np = 5$		$np = 8$
γ	$S(5, 0.6)/M/1$	$S(8, 0.375)/M/1$	$S(10, 0.3)/M/1$	$S(8, 0.625)/M/1$	$S(10, 0.5)/M/1$	$S(10, 0.8)/M/1$
0.05	84.69%	187.50%	241.25%	63.37%	100.80%	25.31%
0.10	77.55%	163.27%	204.40%	58.25%	90.72%	23.47%
0.15	73.28%	148.64%	182.78%	54.60%	83.59%	22.11%
0.20	69.91%	138.41%	167.96%	51.66%	78.14%	21.39%
0.25	67.55%	130.42%	154.31%	49.60%	74.13%	20.49%
0.30	65.50%	124.11%	143.11%	47.74%	70.59%	19.79%
0.35	63.87%	117.25%	134.29%	46.04%	67.53%	19.21%
0.40	62.50%	110.88%	127.25%	44.61%	64.98%	18.73%
0.45	61.32%	105.54%	121.35%	43.36%	62.69%	18.23%
0.50	60.40%	101.17%	115.68%	42.28%	60.72%	17.80%
0.55	59.59%	97.41%	110.14%	41.33%	58.15%	17.43%
0.60	58.96%	94.16%	105.39%	40.45%	55.65%	17.06%
0.65	57.37%	89.77%	101.27%	39.69%	53.45%	16.69%
0.70	54.80%	85.68%	96.27%	38.51%	51.46%	16.37%
0.75	52.52%	82.10%	91.51%	36.93%	49.79%	16.14%
0.80	50.55%	77.44%	86.05%	35.56%	48.22%	15.94%
0.85	48.86%	72.41%	79.79%	34.41%	45.65%	15.78%
0.90	44.56%	65.28%	71.44%	33.30%	43.38%	15.27%
0.95	38.62%	52.40%	56.76%	30.70%	39.42%	14.87%
1.00	20.00%	31.25%	35.00%	9.38%	12.50%	2.86%

Table 1: Increase in the average expected waiting time due to no-shows

the phenomenon of no-shows increases the uncertainty, so when the customers' relative cost is high the optimal schedule with no-shows has to leave enough space between scheduled arrivals to reduce waiting time if customers do show up, increasing the expected idle time of the server, if they end up not showing up. Thus from a practical point of view, assuming that the relative cost γ is given and that the system's operating costs are all represented in (4.12), the operating costs can be brought to the lowest expected value by manipulating a specific p portion of the customers to show up. Nevertheless this is not always possible nor desirable in systems serving people, especially people coming to see a doctor.

Definition 3 We define $\Omega^*(\bar{x})$ as the expected total cost of waiting in the system. It consists of the sum of the expected cost of customers waiting in the queue and the expected cost of idle time of the server.

$\Omega^*(\bar{x})$ differs from the original objective function $\Phi^1(\bar{x})$ only in the expected service cost, which is a constant that does not depend on the schedule. Hence $\Omega^*(\bar{x})$ can be obtained from the objective function (4.5) by:

$$\Omega^*(\bar{x}) = \Phi^1(\bar{x}) - E[\text{total service cost}] \tag{6.1}$$

where:

$$E[\text{total service cost}] = \frac{c_s p n}{\mu}. \tag{6.2}$$

Substituting (6.2) in (6.1) and rewriting and simplifying the equation as in Section 4.2.1 we obtain:

$$\Omega(\bar{x}) = \Phi(\bar{x}) - \frac{\tilde{\gamma} p (n-1)}{\mu}. \tag{6.3}$$

The motivation for defining Ω, which differs from the objective function only by a constant, is the following. We assume that the expected service time is inevitable, hence when analyzing the results we are interested in studying the specific effect of no-shows on controllable parameters, i.e. waiting cost of customers and idleness cost of the server.

Figures 22 to 25 present the optimal values of $\Phi(\bar{x})$ and $\Omega(\bar{x})$ as a function of the showing up probability, for five and ten customers for various values of γ. The figures reveal that for a specific relative cost, these graphs have a similar pattern, for different numbers of customers. Though not presented, this same pattern was noted also for all other sizes of systems studied. Also verified is the expected fact that for all studied n and p, for $\gamma = 1$, when only the server's cost is taken into account for determining the schedule, $\Omega(\bar{x}) = 0$. This is a result of all customers being scheduled to arrive together at $t_1 = t_n = 0$, so the server never waits idle for customers to show up, and there is no waiting cost of customers since the relative cost of customers is $1 - \gamma = 0$.

We note that for small values γ the maximal values of the cost functions are obtained for $p < 0.50$. Thus there is a "price" to no-shows. We study this issue in Section 6.2.6.

6.2.5 Operational costs per customer

We define and study now two other measures of the system, concerning operational costs per customer. These measures are defined in respect to $\Phi(\bar{x})$ and $\Omega(\bar{x})$ by dividing them by np.

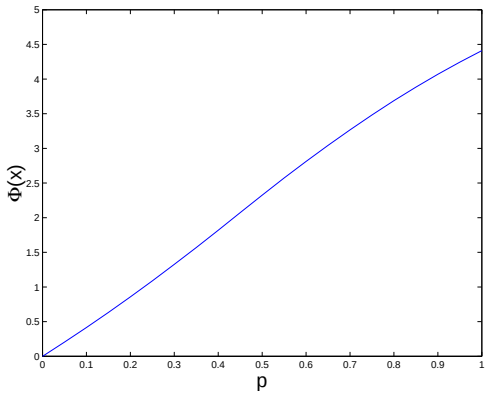

Figure 22: $S(5,p)/M/1, \quad \gamma = 0.80$

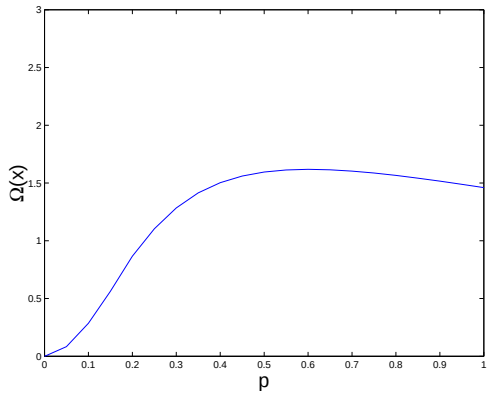

Figure 23: $S(5,p)/M/1, \quad \gamma = 0.20$

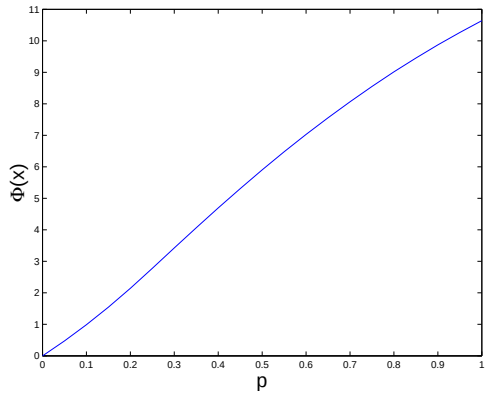

Figure 24: $S(10,p)/M/1, \quad \gamma = 0.80$

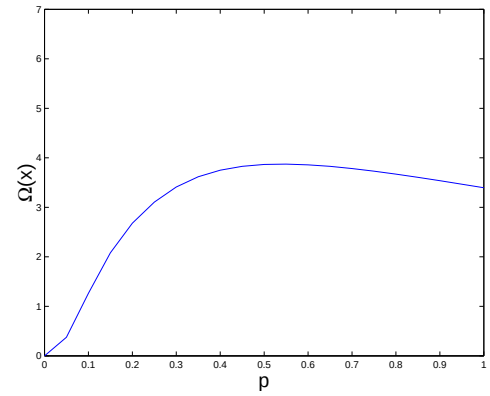

Figure 25: $S(10,p)/M/1, \quad \gamma = 0.20$

$\frac{\Omega(\bar{x})}{np}$ is a measure for system's waiting costs per customer, whereas $\frac{\Phi(\bar{x})}{np}$ takes into account also the service costs. Figures 26 to 29 present the values of these measures based on the optimal values of (4.12) and (6.3) as a function of the showing up probability, for five and ten customers for various values of γ.

We note that these measures behave like the expected cost function they originate from, $\Phi(\bar{x})$ and $\Omega(\bar{x})$ respectively, in the sense that the maximum value they obtain at optimum is not necessarily when all customers show up. The measures differ from their origin in the range of γ for which this phenomenon is noted; as mentioned before, for the original functions this is true for small values of γ, while for these operational measures this holds true for a much wider range, even up to about $\gamma = 0.80$.

Another difference noted when comparing the graphs of $\frac{\Phi(\bar{x})}{np}$ with those of $\Phi(\bar{x})$ and the graphs of $\frac{\Omega(\bar{x})}{np}$ with those of $\Omega(\bar{x})$ is that their picks are not obtained at the same showing up probability, and moreover the picks of the operational measures graphs are obtained for a lower p. This may imply that customers who show up "pay" more if many of the scheduled customers do not show up.

6.2.6 Cost of no-shows

In order to evaluate the cost of no-shows, we compare the costs of systems where all customers show up to systems with no-shows with the same expected number of customers who show

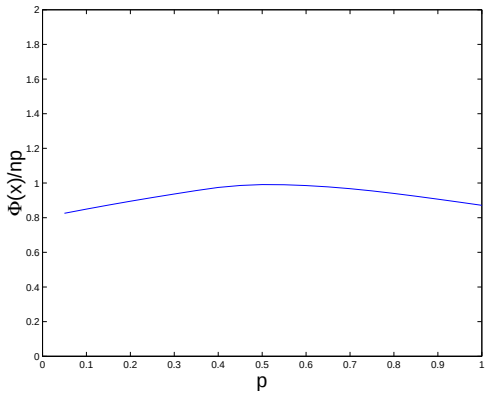

Figure 26: $S(5,p)/M/1, \quad \gamma = 0.70$

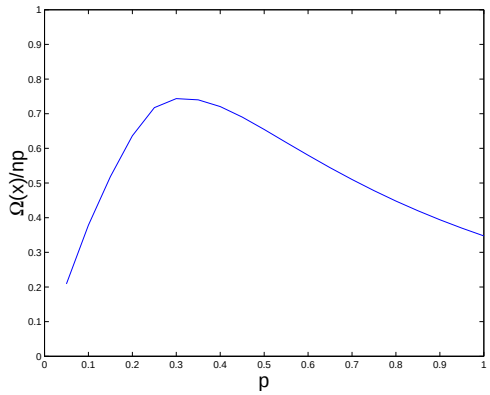

Figure 27: $S(5,p)/M/1, \quad \gamma = 0.30$

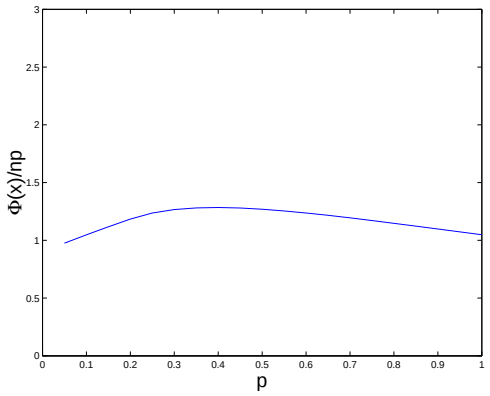

Figure 28: $S(10,p)/M/1, \quad \gamma = 0.70$

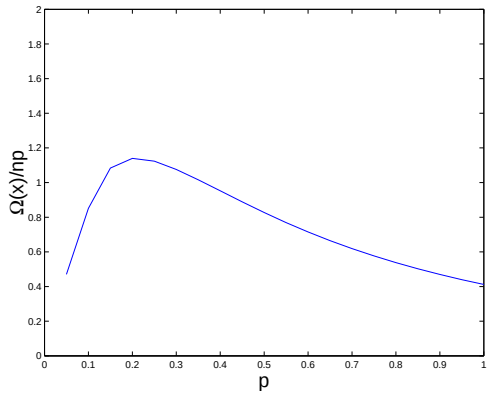

Figure 29: $S(10,p)/M/1, \quad \gamma = 0.30$

up. We compare the costs of $S(n, p < 1.0)/M/1$ models to the costs of $S(n', 1.0)/M/1$ models where $np = n'$.

Each of the Figures 30 to 33 presents a comparison between the costs of a $S(n, p < 1.00)/M/1$ model and the costs of the equivalent $S(n', 1.00)/M/1$ model where $np = n'$. In these graphs the relative server's availability cost is stated on the X-axis, noted by γ, and the costs are stated on the Y-axis. The costs of the models with no-shows are drawn in solid lines, whereas the costs of the models where all customers show up are drawn in a dashed lines.

We note that as p decreases, i.e., the probability of customers' not showing up increases, the cost of no-shows increases. We measure by how much the cost increases when there are no-shows, in comparison with the cost when all customer show up. We find that for $np = 3$, when the system is designed for five customers of which 40% do not show up, the phenomenon of no-shows increases the cost $\Phi(\bar{x})$ by $20\% - 150\%$ and the cost $\Omega(\bar{x})$ by $35\% - 170\%$. When 70% of ten customers do not show up, the phenomenon increases $\Phi(\bar{x})$ by $35\% - 572\%$ and $\Omega(\bar{x})$ by $53\% - 660\%$. The influence of the phenomenon on the costs is greater on smaller relative server's availability cost.

These findings conform with the findings detailed in Section 6.2.3 for the expected waiting times. Nevertheless the impact of no-shows on the expected waiting times, measured in percentages, is not so high. Thus the no-shows increase also the server's idle time.

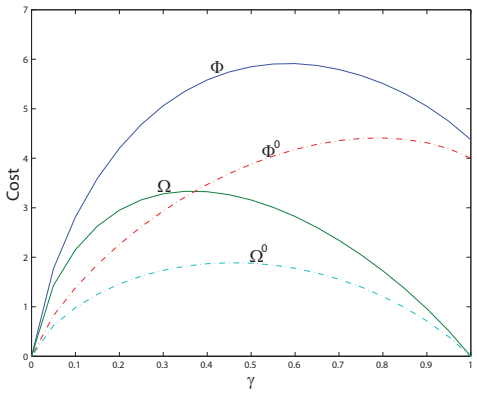

Figure 30: p Costs $S(8, 0.625)/M/1$

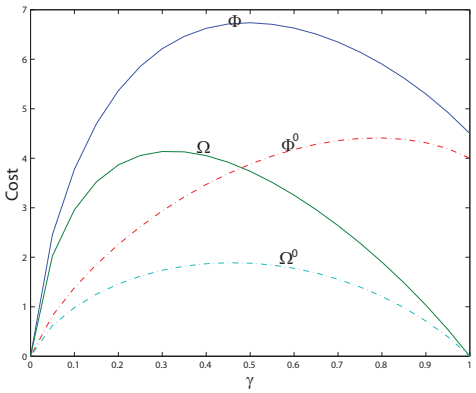

Figure 31: p Costs $S(10, 0.50)/M/1$

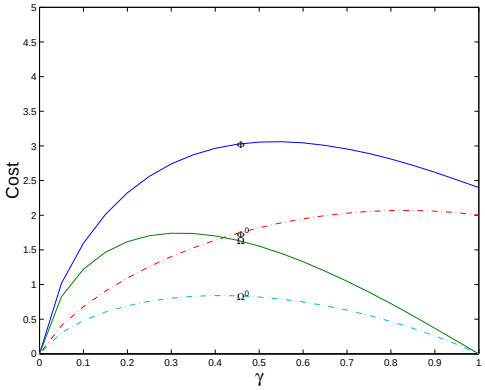

Figure 32: p Costs $S(5, 0.60)/M/1$

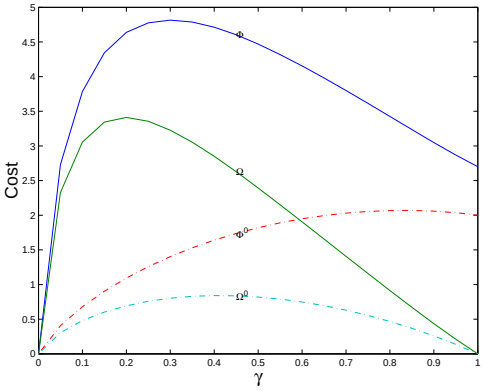

Figure 33: p Costs $S(10, 0.30)/M/1$

Also noted on these figures is the phenomenon mentioned in Section 6.2.4, that $\Omega(\bar{x}) = 0$ for $\gamma = 1$. Since all customers are scheduled to arrive together at $t_1 = t_n = 0$ so the server never waits idle for customers to show up, and there is no waiting cost of customers since the relative cost of customers is $1 - \gamma = 0$.

7 The equally spaced model

Following the work of Stein and Cote, we consider also the case where the customers are scheduled to arrive at the system at equally spaced times, i.e. $x_1 = x_2 = \cdots = x_{n-1}$. We add to their model the attribute that each customer shows up with a probability $p \in (0, 1]$.

As stated by Stein and Cote, the equally spaced model is of interest as it provides a realistic restriction to the scheduling problem, since in most appointment systems the appointments are scheduled using a fixed interval between scheduled arrivals.

7.1 General solution method

The equations representing the model remain the same, with the exception that $x_i = x_{eq} \quad \forall i$. Thus based on (4.12) our objective is to minimize:

$$\Phi_{eq}(x_{eq}) = (1 - \tilde{\gamma}) \sum_{i=2}^{n-1} w_i^s + \tilde{\gamma}(n-1)x_{eq} + w_n^s. \tag{7.1}$$

The solution methods are similar to those used in the unrestricted model. Even though we are now looking for the value of a single variable x_{eq} that optimizes the objective function (7.1) (in opposed to a vector of values) we tackle equations that do not have a closed form solution.

For instance, for three customers, after substituting $x_i = x_{eq} \quad \forall i$ in (5.10) we obtain:

$$\begin{aligned} \Phi_{eq}(x_{eq}) &= (1 - \tilde{\gamma})\frac{pe^{-\mu x_{eq}}}{\mu} + \tilde{\gamma} \cdot 2x_{eq} + \frac{pe^{-\mu x_{eq}}}{\mu}\left[1 + e^{-\mu x_{eq}} + p\mu x_{eq} e^{-\mu x_{eq}}\right] \\ &= 2\tilde{\gamma} \cdot x_{eq} + \frac{pe^{-\mu x_{eq}}}{\mu}\left[2 - \tilde{\gamma} + e^{-\mu x_{eq}}(1 + p\mu x_{eq})\right]. \end{aligned} \tag{7.2}$$

The value of x_{eq} that minimizes (7.2) must satisfy:

$$\begin{aligned} \frac{d\Phi_{eq}(x_{eq})}{dx_{eq}} &= 2\tilde{\gamma} - pe^{-\mu x_{eq}}\left[2 - \tilde{\gamma} + e^{-\mu x_{eq}}(1 + p\mu x_{eq})\right] \\ &+ \frac{pe^{-\mu x_{eq}}}{\mu}\left[-\mu e^{-\mu x_{eq}}(1 + p\mu x_{eq}) + e^{-\mu x_{eq}} p\mu\right] = 0 \\ &= 2\tilde{\gamma} - pe^{-\mu x_{eq}}\left[2 - \tilde{\gamma} + e^{-\mu x_{eq}}(2 + 2p\mu x_{eq} - p\mu)\right] = 0. \end{aligned} \tag{7.3}$$

For $n \geq 4$ we substitute $x_i = x_{eq} \quad \forall i$ in (4.15) and (4.16) and use the results to form w_i^s according to (4.13) and set it to the objective function (7.1). For the optimal solution the derivative of this function must equal zero. Yet the obtained equation has no close form. Hence, as in the unrestricted interval widths model, we obtain the solution numerically for the simplified case of $\mu = 1$. The solution can be obtained by modeling the Newton-Raphson approximation method in Matlab in a similar manner performed for the $S(3, p)/M/1$ unrestricted model as detailed in Appendix A.1, or by using approximation methods already embedded in Matlab optimization toolbox as details Appendix A.2. We obtain the optimal solution by the approximation methods embedded in Matlab optimization toolbox.

7.2 Analysis of the equally spaced optimal solution

We find that Stein and Cote's conclusions for the basic model, where all customers arrive on an equally spaced schedule, go also for the $S(n,p)/M/1$ model constrained to equally spaced arrivals. The affect of adding a constraint to force equally spaced intervals does not materially change the value of the objective function for any combination of p and γ and determines an optimal interval which is almost an average of the unrestricted interval widths.

7.2.1 Optimal equal interval between scheduled arrivals

As stated above the optimal equal interval between scheduled arrivals is approximately the average of the optimal unrestricted intervals.

A comparison between the optimal interval for this model and the optimal unrestricted interval widths, for three and five independent customers for various values of p and γ, is presented in Figures 34 to 37. Each of these graphs presents the optimal inter-arrival times for a unique combination of n, p and γ for the equally spaced and the unrestricted $S(n,p)/M/1$ models. In these graphs the relative server's availability cost is stated on the X-axis, denoted by γ, and the inter-arrival spacing is stated on the Y-axis, denoted by x_i. Graphs of systems designed for $n > 5$ customers are not presented due to the density of the resulting graphs which makes it difficult to note the details. Nevertheless the results follow the same pattern as in the presented graphs.

7.2.2 Relation between the equal spacing and the showing up probability

From a practical point of view it is of interest to study the relation between the optimal equal spacing and the showing up probability. Assuming a given server's availability relative cost, the optimal equal spacing depends upon the showing up probability. Figures 38 to 41 present for three, five, eight and ten customers, the optimal equal spacing solution, for given relative costs, as a function of the showing up probabilities. From these graphs we note that for a wide range of relative cost values there seems to be almost a linear relation between the optimal equal spacing and the showing up probability.

7.2.3 No-shows impact on the optimal equal spacing

Studying the impact of no-shows on the optimal equal spacing between scheduled arrivals, we compare the optimal equal spacing of systems where all customers show up to systems with no-shows with the same expected number of customers who show up. We compare the optimal spacing between scheduled arrivals of $S(n, p < 1.0)/M/1$ models to the ones of $S(n', 1.0)/M/1$ models where $np = n'$.

Each of the Figures 42 to 44 presents a comparison between the optimal equal spacing of $S(n, p < 1.00)/M/1$ models and the optimal equal spacing between scheduled arrivals of the equivalent $S(n', 1.00)/M/1$ model where $np = n'$. In these graphs the relative server's availability cost is stated on the X-axis, noted by γ, and the optimal spacings are stated on the Y-axis. The spacings of the models with no-shows are drawn in solid lines, whereas these of the models where all customers show up are drawn in a dashed lines.

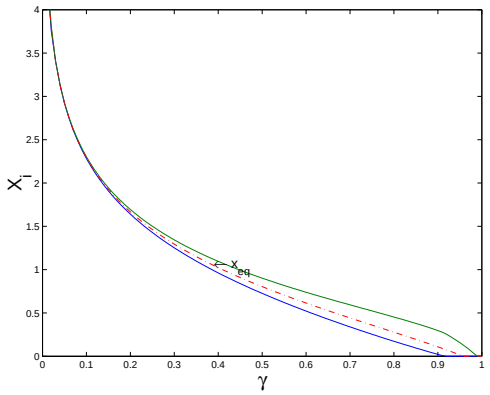

Figure 34: Spacing $S(3, 0.90)/M/1$

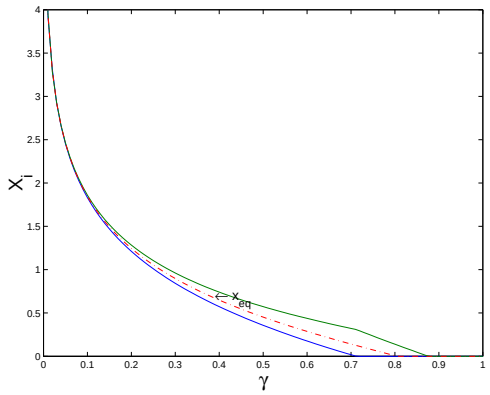

Figure 35: Spacing $S(3, 0.70)/M/1$

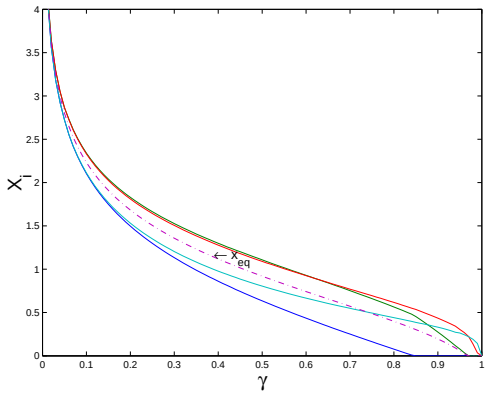

Figure 36: Spacing $S(5, 0.80)/M/1$

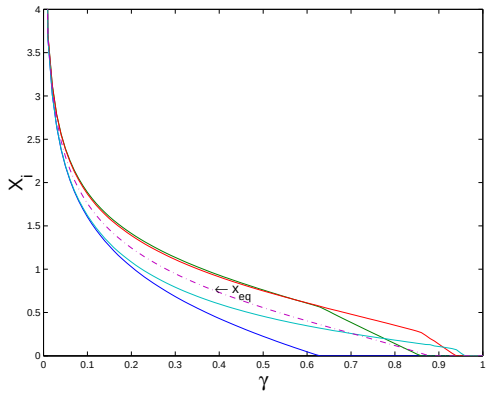

Figure 37: Spacing $S(5, 0.60)/M/1$

The comparison reviles that the optimal spacing for systems with no-shows is smaller than the spacing when all customers show up. Moreover, the optimal spacing decreases as the showing probability decreases. If the customers who do not show up would have notified in advanced that they are not going to show up, an optimal schedule could have been designed only for those who show up, resulting with a more spacious schedule. This aligns with the impact of no-shows on the expected waiting times and costs, as detailed in Sections 6.2.3, 6.2.6, 7.2.4 and 7.2.6. Due to expected no-show, the optimal schedule is more condensed, resulting in longer expected waiting times leading to higher costs than there would have been if the system was designed only for the customers who do show up eventually.

We also note that the impact of no-shows on the optimal spacing is smaller for extreme relative service availability costs. I.e., the impact when γ is very small or very large is not as notable as it is for intermediate values of γ. In these extreme cases the optimal schedule is highly influenced by the relative cost, hence the impact of no-shows is not as significant.

7.2.4 Resulting customers' expected waiting times

As in the unrestricted model, the resulting expected waiting times of customers who show up have an increasing pattern as the scheduled place in line increases, i.e. $w^s_{i+1} > w^s_i \quad \forall i = 1, ..., n-1$, and they are distributed with considerable variance.

Comparing the average and the maximal expected waiting times for a customer who

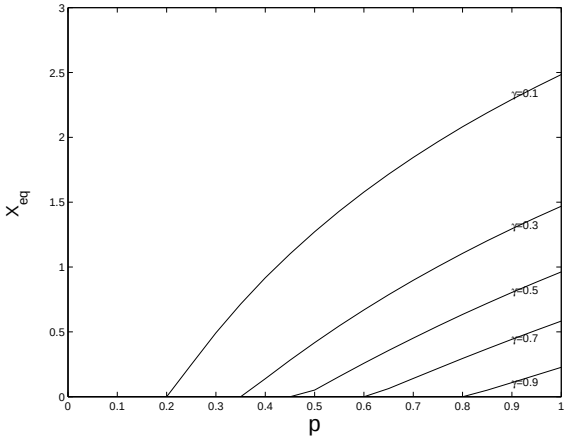

Figure 38: Equal $S(3,p)/M/1$

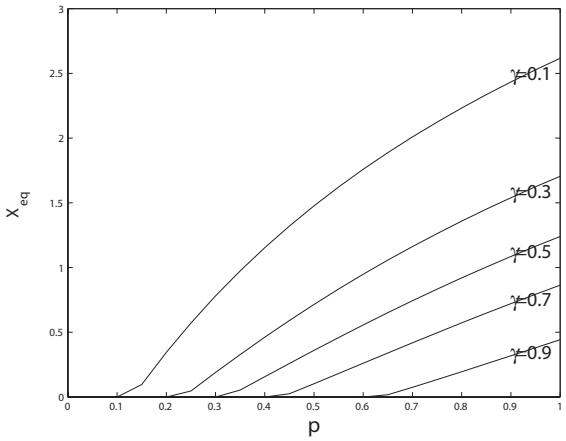

Figure 39: Equal $S(5,p)/M/1$

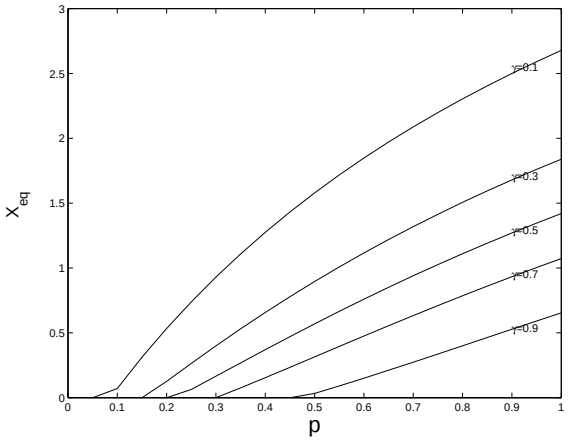

Figure 40: Equal $S(8,p)/M/1$

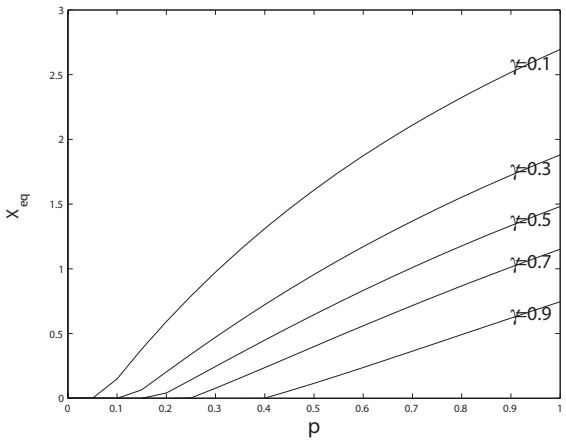

Figure 41: Equal $S(10,p)/M/1$

shows up in the equally spaced model and in the unrestricted model reveals that for a wide range of γ the average expected waiting time of a customer who shows up obtained by the equally spaced model more or less equals to the one obtained by the unrestricted model (the numeric solution hardly differs). Such comparisons are presented for five and ten customers for various values of p and γ in Figures 45 to 48. In these graphs the relative server's availability cost is denoted on the X-axis, noted by γ, and the expected waiting times of a customer who shows up for both models are denoted on the Y-axis. The waiting times resulting from the equally spaced inter-arrival model are drawn by dashed lines, while the results of the unrestricted model are drawn by solid lines.

Studying the impact of the no-shows phenomenon in the same manner performed for the unrestricted model, i.e., comparing the expected waiting times for customers who show up of $S(n,p<1.0)/M/1$ models to the ones of $S(n',1.0)/M/1$ models where $np=n'$, reveals similar results.

7.2.5 Minimal average expected waiting time

Another subject of interest concerning the average expected waiting time is related to Theorem 1.1 of Hajek [3]. According to this theorem the average waiting time of a customer, in a system with one exponential server, is at minimum for constant inter-arrival times. Compar-

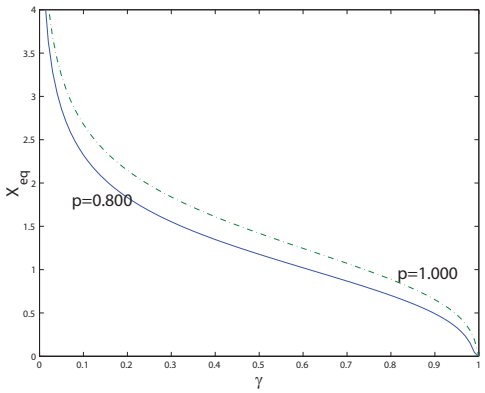

Figure 42: $x_{eq} \quad np = 8$

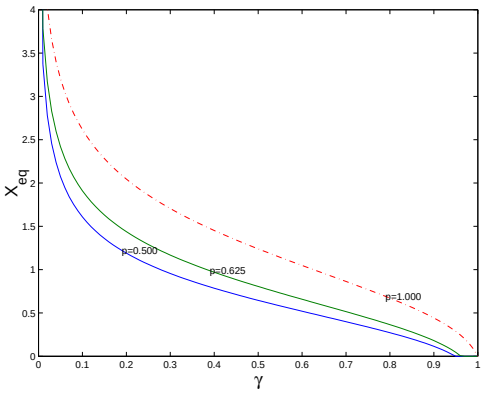

Figure 43: $x_{eq} \quad np = 5$

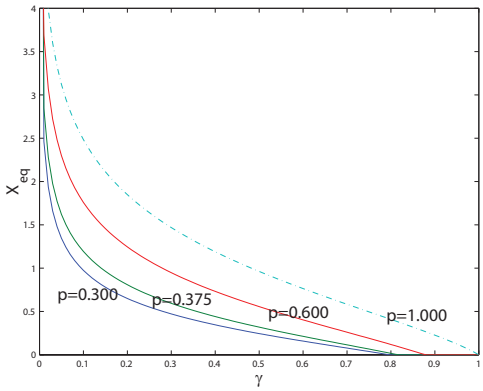

Figure 44: $x_{eq} \quad np = 3$

isons between the average expected waiting times for a customer who shows up in the equally spaced model and in the unrestricted model can be seen in Figures 45 to 48. As mentioned before we note that for a wide range of γ the average expected waiting times of customers who show up obtained by the equally spaced model hardly differs numerically from the one obtained by the unrestricted model. Furthermore, for high server's availability cost the average expected waiting time in the equally space model is higher than the one expected in the unrestricted model. The threshold value of γ from which the $S(n,p)/M/1$ model does not follow Hajek's Theorem decreases as the showing up probability decreases. The fact that the Hajek's Theorem does not always hold in a system with no-shows can be explained by the fact that the defined objective function Φ includes other factors that may contradict the minimization of the average customers expected waiting time. In a $S(n,p/M/1)$ model with the stated objective function Φ, reducing the customers' expected waiting times leads to increasing the server's idle time. The minimum of the objective function (4.12) is somewhat a balance between those other factors and the average expected waiting times. The range for which Hajek's Theorem does hold depend on the weight given in the objective function to the customers, i.e., the lower is the value of γ the higher is their importance. Moreover, as more customers arrive the range of γ in which the theorem holds is wider since their weight in the objective function is higher.

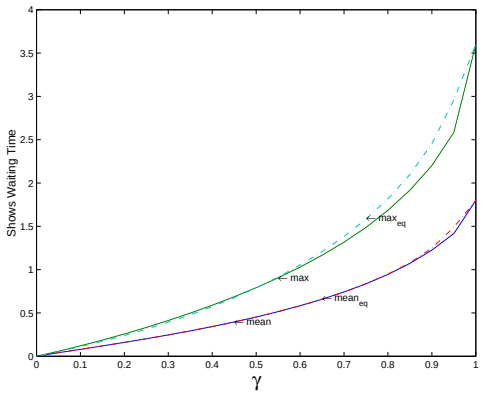

Figure 45: Waiting $S(5, 0.90)/M/1$

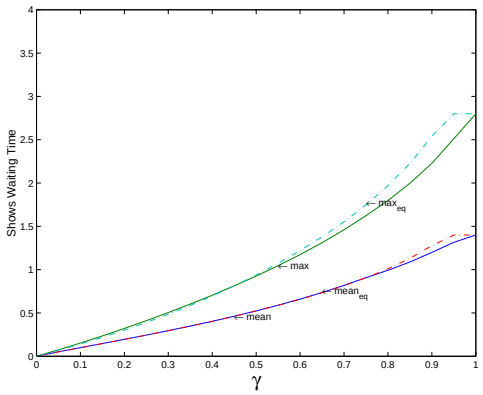

Figure 46: Waiting $S(5, 0.70)/M/1$

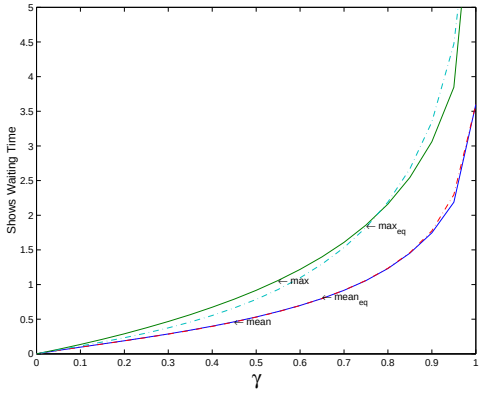

Figure 47: Waiting $S(10, 0.80)/M/1$

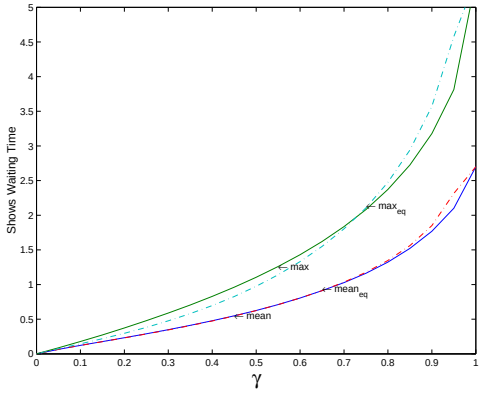

Figure 48: Waiting $S(10, 0.60)/M/1$

7.2.6 Objective function's value and cost of waiting times

As mentioned above, the constraint of equally spaced intervals does not have a significant affect on the objective function for any combination of p and γ. We also study the effect of equally spaced scheduled inter-arrival times on the value of the expected total cost of waiting in the system. Comparisons between the objective function values of the equally spaced model and the unrestricted model and between the cost of waiting times of the same models, are presented in Figures 49 to 52. In these graphs the relative server's availability cost is denoted on the X-axis and the cost functions' values are denoted on the Y-axis. These graphs demonstrate the minor differences between the objective functions' values of the equally spaced model and the unrestricted model. Also the affect on the expected cost of waiting is minor.

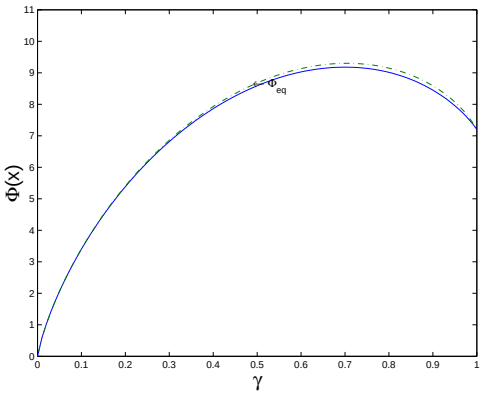

Figure 49: $\Phi(x)$ $S(10, 0.80)/M/1$

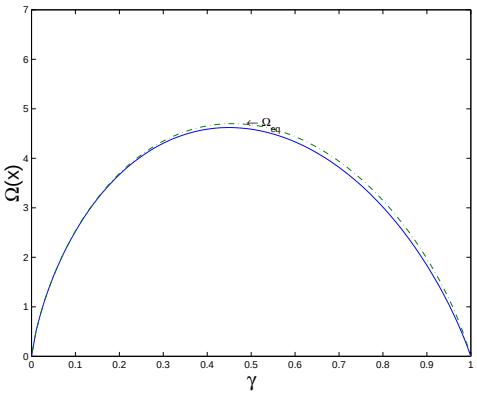

Figure 50: $\Omega(x)$ $S(10, 0.80)/M/1$

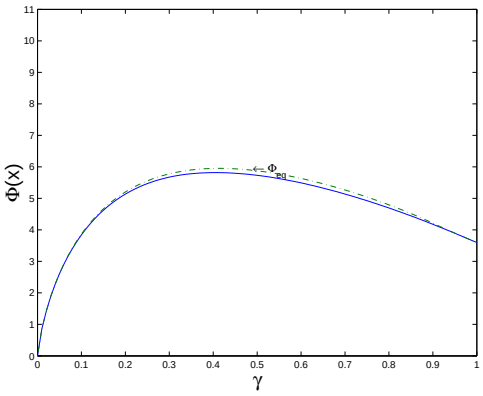

Figure 51: $\Phi(x)$ $S(10, 0.40)/M/1$

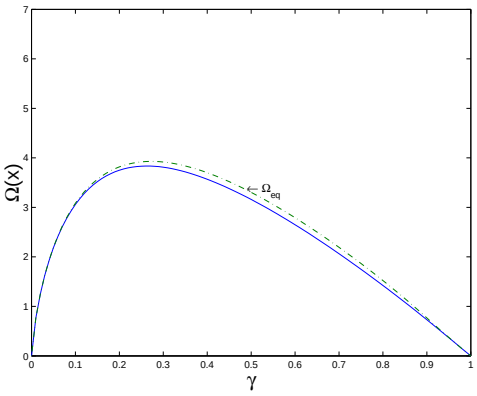

Figure 52: $\Omega(x)$ $S(10, 0.40)/M/1$

References

[1] Bailey N. (1952) "A study of queues and appointment systems in hospitals outpatients departments with special reference to waiting times." *Journal of the Royal Statistical Society*, 14: 185-199.

[2] Cayirli T. and E. Veral (2003) "Outpatient scheduling in health care: a review of literature." *Production and Operations Management*, Winter 2003, 12(4): 519-549.

[3] Hajek B. (1983) "The proof of a folk theorem on queueing delay with applications to routing networks." *Journal of the Association for Computing Machinery*, 30(4): 834-851.

[4] Kaandorp G. and G. Koole (2006) "Optimal outpatient appointment scheduling." *Working paper, Department of Mathematics, Vrije University, Amsterdam, The Netherlands.*

[5] Mercer A. (1960) "A queueing problem in which the arrival times of the customers are scheduled." *Journal of Royal Statistical Society, Series B*, 22(1): 108-113

[6] Mercer A. (1973) "Queues with scheduled arrivals: a correction, simplification and extension." *Journal of Royal Statistical Society, Series B*, 35(1): 104-116

[7] Mondschein S.V. and G.Y. Weintraub (2003) "Appointment policies in service operations service operations: a critical analysis of the economic framework." *Production and Operations Management*, Summer 2003, 12(2): 266-286.

[8] Pegden C.D. and M. Rosenshine (1990) "Scheduling Arrivals to Queues." *Computers Operations Research*, 17(4): 343-348.

[9] Preater J. (2001) "A bibliography of queues in health and medicine." *Keele Mathematics Research Report*, 2001-01.

[10] Stein W.E. and M.J. Cote (1994) "Scheduling Arrivals to Queues." *Computers Operations Research*, 21(6): 607-614.

A Appendices

A.1 Numerical solution for $S(3,p)/M/1$

As stated in Section 5.2, the optimal schedule for the $S(3,p)/M/1$ model, is obtained by using the Newton-Raphson method, which is known to be an efficient algorithm for finding approximations to the zeros of a real-valued function.

The idea of the method is as follows: one starts with a value which is reasonably close to the true zero, then replaces the function by its tangent and computes the zero of this tangent which will typically be a better approximation to the function's zero. The method can be iterated:
Suppose $f : [a,b] \to R$ is a differentiable function, we start with an arbitrary value $\hat{x}_0$, in the nth step of the algorithm $\hat{x}_n$ is defines as follows :

$$\hat{x}_n = \hat{x}_{n-1} - \frac{f(\hat{x}_{n-1})}{f'(\hat{x}_{n-1})}. \tag{A.1}$$

If f' is continuous, and if the unknown zero x is isolated, then there exists a neighborhood of x such that for all starting values $\hat{x}_0$ in that neighborhood, the sequence $(\hat{x}_n)$ converges towards x. Furthermore, if $f'(x) \neq 0$, then the convergence is quadratic (which means that the number of correct digits roughly doubles in every step).

We start the solution process by applying the Newton-Raphson approximation to find the zero value of $f(x_1)$ as defined in (5.18). Rewriting the function in terms of $x_1 = x$:

$$f(x) = pA(x)\left[1 + e^{-x_1}(1 - p - p\ln A(x))\right] - \tilde{\gamma}. \tag{A.2}$$

The derivative of this function (required for the Newton-Raphson algorithm) is:

$$\begin{aligned} f'(x) &= p\frac{\mathrm{d}A(x)}{\mathrm{d}x}\left[1 + e^{-x}(1 - p - p\ln A(x))\right] + \\ &+ pA(x)\left[-e^{-x}(1 - p - p\ln A(x)) + e^{-x}(-p)\frac{1}{A(x)}\frac{\mathrm{d}A}{\mathrm{d}x}\right] = \\ &= p\left[\frac{\mathrm{d}A}{\mathrm{d}x}(1 - pe^{-x}) + e^{-x}(1 - p - p\ln A)(\frac{\mathrm{d}A}{\mathrm{d}x} - A)\right]. \end{aligned} \tag{A.3}$$

Where $\frac{\mathrm{d}A}{\mathrm{d}x}$ is (for the definition of A see (5.15)):

$$\frac{\mathrm{d}A(x)}{\mathrm{d}x} = \frac{(\tilde{\gamma} - 1)e^{-x}}{(1 - pe^{-x})^2}.$$

We use the algorithm by implementing it in Matlab. The stopping condition for the implemented algorithm, is defined by the difference between two sequential steps, that is if $|\hat{x}_n - \hat{x}_{n-1}| \leq \epsilon$ the approximation for x_1 is: $x_1 = \hat{x}_n$. The obtained $\hat{x}_n$ may be negative, in which case the feasible solution $x_1 = 0$ is forced.

If $x_1 > 0$, x_2 is then obtained simply by substituting x_1 in (5.16). Otherwise, if $x_1 = 0$, we obtain x_2 by applying the algorithm to find the zero value function of (5.19) using Matlab. In any of these cases, if the obtained $x_2 < 0$ we force the feasible solution $x_2 = 0$.

Using the above method, we find the optimal inter-arrival times for three customers for various values of $p \in (0,1]$ and $\gamma \in (0,1]$.

A.2 Numerical solution for $S(n,p)/M/1 \quad n \geq 4$

As mentioned in Sections 6.1 and 7.1 we use Matlab optimization toolbox for computing the optimal solution. From this toolbox we use the function `fmincon` which implements a sequential quadratic programming (SQP) nonlinear programming method. SQP mimics Newton's method for constrained optimization just as is done for unconstrained optimization. At each major iteration, an approximation is made of the Hessian of the Lagrangian function using a quasi-Newton updating method. This is then used to generate a quadratic programming (QP) subproblem whose solution is used to form a search direction for a line search procedure. Full details about the algorithm can be found on the web at `http://www.mathworks.com/access/helpdesk/help/toolbox/optim/ug/f26684.html`.