Queues with a dynamic schedule

John Gilbertson

A thesis presented for the degree of
Master of Science (Mathematics and Statistics)

Supervised by Professor Peter Taylor
Department of Mathematics and Statistics
The University of Melbourne
October 2016

Declaration

This thesis is the sole work of the author whose name appears on the title page and it contains no material which the author has previously submitted for assessment at the University of Melbourne or elsewhere. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person, in the form of unacknowledged quotations or mathematical workings or in any other form, except where due reference is made. I declare that I have read, and in undertaking this research I have complied with, the University's Code of Conduct for Research. I also declare that I understand what is meant by plagiarism and that this is unacceptable.

Signed

John Gilbertson

Abstract

Abstract goes here.

Contents

Chapter 1

Introduction

Introduction goes here.

Chapter 2

Literature Review

Health care providers are under a great deal of pressure to improve service quality and efficiency (Goldsmith, 1989). There is a large body of literature studying the potential of appointment systems to reduce patient waiting times, and waiting room congestion. Fomundam and Herrmann (2007), and Cayirli and Veral (2003) provide comprehensive surveys of research on appointment scheduling. There is a fundamental trade-off in appointment policies. If patients are scheduled to arrive close together, they experience long waiting times. However, if appointment times are spread further apart, the doctor's idle time increases.

Most of the papers on scheduled arrivals in health care can be classed into two categories. Those that design algorithms to find good schedules, and those that evaluate schedules using simulation. While simulation studies can easily model complicated patient flows, queuing models often provide more generic results than simulation (Green, 2006).

The foundation paper on modeling queues with scheduled arrivals is Bailey (1952). Bailey proposes that customers' waiting times can be reduced without a significant increase in doctor's idle time. The Bailey rule, which is commonly referenced in literature, is that patients should be scheduled to arrive at fixed intervals with two patients scheduled to arrive at the start of service. Bailey found that a great deal of time wasted by patients could be reduced without a signficant increase in the doctor's idle time. Under the Bailey rule, patients with late appointments will wait longer than those with early appointments. This lack of uniformity might be undesirable due to issues of fairness.

Pegden and Rosenshine (1990) extend on Bailey's paper. They present an algorithm to iteratively determine the optimal arrival times for n patients that need to be scheduled. The optimal arrival times are those that minimise a weighted sum of the expected patients' waiting time and the expected doctor's idle time. Pegden and Rosenshine prove that their objective function is convex for $n \leq 4$, thus their algorithm finds the optimal schedule. While they conjecture that the objective function is convex for $n \geq 5$, it hasn't been proven.

Stein and Côté (1994) apply Pegden and Roshenshine's model to obtain numerical results for situations with more than three patients. The optimal times between successive patients become near constant as n grows. This is the often observed dome-shape. Optimal appointment intervals exhibit a common pattern where they initially increase towards the middle of a session, and then decrease. Stein and Côte simplify the model by requiring the intervals between arriving patients to be held constant. This realistic restriction (used commonly in the literature) makes the model more easily applicable in practice without significant altering the results.

Stein and Côte apply queuing theory results to solve the model for the optimal arrival interval assuming the queue reaches its steady state distribution. This assumption greatly reduces the computation required. However, in practice, it is common to find services that never reach steady state. Babes and Sarma (1991) attempted to apply steady state queuing theory, but found their results tended to be very different from those observed in real operation.

These key papers by Bailey, Pegden and Rosenshine, and Stein and Côte provide the basis for a more realistic exploration of health care systems. DeLaurentis et al. (2006) point out that patient no-shows can lead to a waste of resources. Mendel (2006) incorporates the probability of a patient not showing up into the model presented by Pegden and Rosenshine. Unsurprisingly, no-shows lead to lower expected waiting times for patients who do show up.

The presence of walk-ins (regular and emergency) can disrupt a schedule. Gupta, Zoreda, and Kramer (1971) propose a system where non-routine requests are superimposed on top of routine scheduled requests. Fiems, Koole, and Nain (2007) investigate the effect of emergency requests on the waiting times of scheduled patients. Fiem models a system with deterministic service times and discrete time. Despite this research, Cayirli and Veral suggest that walk-ins are neglected in most studies. Further research could investigate their effect on optimal arrival times.

Mondschein and Weintraub (2003) observe that the majority of the literature assumes that demand is exogenous and independent of patients' waiting times. These papers assume the total number of patients n is fixed and independent of waiting times. The vast majority of servers are now private (including medical servers), so face competitive environments. Mondschein and Weintraub thus present a model where demand depends on the patients' expected waiting time.

Simulation is a useful tool to analyse the effectiveness of appointment policies. Kao and Tung (1981) use simulation to compliment their results obtained from queuing theory. Ho and Lau (1992) study the performance of eight different appointment rules under different scenarios. They find that no rule will perform well under all circumstances.

Case studies can test the real world applications of an appointment system. While they lack generalisation, they are necessary to compliment the theoretical research. Rockart and Hofmann (1969) show individual block systems lead to more punctual doctors and patients, and less no-shows. Walter (1973) indicates that the simple grouping of inpatients and outpatients results in a substantial improvement in doctor's idle time.

Unfortunately, Cayirli and Veral (2003) lament that despite much published work, the impact of appointment systems on outpatient clinics has been limited. Doctors are often unwilling to change their old habits. O'Keefe (1985) had their proposed appointment system of classifying patients rejected by staff. Huarng and Hou Lee (1996) were unable to implement their system due to staff resistance. Bennett and Worthington (1998) found their recommendations weren't implemented successfully. Future research must attempt to develop models that will be accepted and implemented in real health care services.

Chapter 3

Static Schedule

3.1 Aim

The aim is to choose a schedule that minimises the expected cost of the system. The expected cost is a linear combination of the total customers' waiting time and the server's idle time. The schedule is fixed and chosen at the start of service.

3.2 Assumptions

To simplify this problem, need to make several assumptions:

- Service times are independent and identically distributed (iid)
- Each service time has an exponential distribution with mean service time μ
- There is a single server
- The queue operates on a first in, first out (FIFO) basis
- Customers are punctual and arrive at their scheduled time

3.3 List of Variables

List of variables goes here

Let w_i denote the expected waiting time of the i-th customer.

Let t_1 denote the scheduled arrival time of the first customer.

Let x_i denote the time interval between the scheduled arrival of the i-th and the $(i+1)$-th customer.

Let μ denote the mean service time of each customer.

3.4 Objective Function

The expected total customers' waiting time is the sum of the individual customer's expected waiting times:

$$\mathbb{E}\Big[\text{total customer's waiting time}\Big] = \sum_{i=1}^{n} w_i \tag{3.1}$$

The expected total server availability time is the sum of the n-th customer's scheduled arrival time, the expected waiting time of the n-th customer, and the expected service time of the n-th customer:

$$\mathbb{E}\Big[\text{total server availability time}\Big] = \left(t_1 + \sum_{i=1}^{n-1} x_i\right) + w_n + \mu \tag{3.2}$$

The objective function is a linear combination of the expected total customers' waiting time and the expected total server availability time:

$$\phi(\mathbf{x}) = c_W \sum_{i=1}^{n} w_i + c_S \left[t_1 + \sum_{i=1}^{n-1} x_i + w_n + \mu\right] \tag{3.3}$$

The first customer should obviously be scheduled for the start of service, so $t_1 = 0$, which implies the objective function is:

$$\phi(\mathbf{x}) = c_W \sum_{i=1}^{n} w_i + c_S \left[\sum_{i=1}^{n-1} x_i + w_n + \mu\right] \tag{3.4}$$

Scale $\phi(\mathbf{x})$ by dividing by $(c_W + c_S)$ and defining $\gamma = \frac{c_S}{c_W + c_S}$.

$$\phi(\mathbf{x}) = (1-\gamma) \sum_{i=1}^{n} w_i + \gamma \left[\sum_{i=1}^{n-1} x_i + w_n + \mu\right] \tag{3.5}$$

3.4.1 Expected Customer Waiting Times

Want to express the expected waiting time of customer i as a function of the scheduled arrival time intervals $\mathbf{x}$. If there are j customers in the system just prior to the arrival of customer i, then $w_i = j\mu$ by the memoryless property of the exponential distribution. Just prior to the arrival of customer i, the number of customers in the system must be $\in \{0, 1, \dots, i-1\}$. Thus, the expected waiting time of customer i is:

$$w_i = \sum_{j=0}^{i-1} (j\mu) \cdot \mathbb{P}\Big(N(t_i) = j\Big) = \begin{cases} 0 & \text{where } i = 1 \\ \sum_{j=1}^{i-1} (j\mu) \cdot \mathbb{P}\Big(N(t_i) = j\Big) & \text{where } i \geq 2 \end{cases} \tag{3.6}$$

As there are no customers in the system before the start of service:

$$\mathbb{P}\Big(N(t_1) = 0\Big) = 1 \tag{3.7}$$

For $i \geq 2$, can express the probability of a given number of customers in the system recursively. First, consider the base case of no customers in the system just prior to the arrival of customer i.

$$\begin{aligned}
&\mathbb{P}\Big(N(t_i) = 0\Big) \\
&= \sum_{k=1}^{i-1} \mathbb{P}\Big(N(t_{i-1}) = k-1\Big) \cdot \mathbb{P}\Big(N(t_i) = 0 \;\Big|\; N(t_{i-1}) = k-1\Big) \\
&= \sum_{k=1}^{i-1} \mathbb{P}\Big(N(t_{i-1}) = k-1\Big) \\
&\qquad \cdot \mathbb{P}\Big(\text{time between } (i-1)\text{-th and } i\text{-th arrival allows for} \geq k \text{ departures}\Big) \\
&= \sum_{k=1}^{i-1} \mathbb{P}\Big(N(t_{i-1}) = k-1\Big) \cdot \left[\sum_{l=k}^{\infty} \frac{x_{i-1}^{l}}{\mu^{l} \cdot l!} \exp\left(\frac{-x_{i-1}}{\mu}\right)\right] \\
&= \sum_{k=1}^{i-1} \mathbb{P}\Big(N(t_{i-1}) = k-1\Big) \cdot \left[1 - \sum_{l=0}^{k-1} \frac{x_{i-1}^{l}}{\mu^{l} \cdot l!} \exp\left(\frac{-x_{i-1}}{\mu}\right)\right]
\end{aligned}$$

Next, calculate the probability for $j \geq 1$:

$$\begin{aligned}
&\mathbb{P}\Big(N(t_i) = j\Big) \\
&= \sum_{k=0}^{i-j-1} \mathbb{P}\Big(N(t_{i-1}) = j+k-1\Big) \cdot \mathbb{P}\Big(N(t_i) = j \;\Big|\; N(t_{i-1}) = j+k-1\Big) \\
&= \sum_{k=0}^{i-j-1} \mathbb{P}\Big(N(t_{i-1}) = j+k-1\Big) \\
&\qquad \cdot \mathbb{P}\Big(k \text{ departures from queue between } (i-1)\text{-th and } i\text{-th arrival}\Big) \\
&= \sum_{k=0}^{i-j-1} \mathbb{P}\Big(N(t_{i-1}) = j+k-1\Big) \cdot \frac{x_{i-1}^{k}}{\mu^{k} \cdot k!} \exp\left(\frac{-x_{i-1}}{\mu}\right)
\end{aligned}$$

Therefore, the full recursive expression for the probability of j customers in the system immediately before the arrival of customer i is given by:

$$\mathbb{P}\Big(N(t_i) = j\Big)$$
$$= \begin{cases} 1 & \text{where } i = 1,\ j = 0 \\ \sum_{k=1}^{i-1} \mathbb{P}\Big(N(t_{i-1}) = k-1\Big) \cdot \left[1 - \sum_{l=0}^{k-1} \frac{x_{i-1}^l}{\mu^l \cdot l!} \exp\left(\frac{-x_{i-1}}{\mu}\right)\right] & \text{where } i \geq 2,\ j = 0 \\ \sum_{k=0}^{i-j-1} \mathbb{P}\Big(N(t_{i-1}) = j+k-1\Big) \cdot \frac{x_{i-1}^k}{\mu^k \cdot k!} \exp\left(\frac{-x_{i-1}}{\mu}\right) & \text{where } i \geq 2,\ j \geq 1 \end{cases} \tag{3.8}$$

3.4.2 Computing the Objective Function

Pegden and Rosenshine (1990) suggest a similar algorithm to Algorithm 1 for computing the value of the objective function given by Equation 3.5 for a given vector $\mathbf{x}$ and parameters γ and μ.

Algorithm 1 Return $\phi(\mathbf{x})$ for a given vector $\mathbf{x}$, γ and μ

```
function ObjectiveFunction(x,γ,μ)
    for i = 1, 2, ..., n do
        for j = 0, 1, ..., (i − 1) do
            compute P(N(t_i) = j) by Equation 3.8
    for i = 1, 2, ..., n do
        compute w_i by Equation 3.6
    return φ(x) computed by Equation 3.5
```

3.5 Example Models

3.5.1 Model for Two Customers

We consider the simplest case of this model where there are two customers to be scheduled (i.e., $n = 2$). As the first customer is scheduled to arrive at the start of service, the only unknown variable is the time interval between the first and second customers' arrivals (i.e., x_1).

By Equation 3.6, the expected waiting times of the two customers are:

$$w_1 = 0 \tag{3.9}$$

$$w_2 = \mu \exp\left(\frac{-x_1}{\mu}\right) \tag{3.10}$$

By Equation 3.5, the objective function to be minimised is:

$$\phi(x_1) = \mu\left[\gamma + \exp\left(\frac{-x_1}{\mu}\right)\right] + \gamma x_1 \tag{3.11}$$

This objective function is convex as:

$$\forall x_1 \quad \phi''(x_1) = \frac{1}{\mu}\exp\left(\frac{-x_1}{\mu}\right) > 0 \tag{3.12}$$

Due to the convexity of the objective function, the optimal policy that minmises $\phi(x_1)$ can be found by solving:

$$\phi'(x_1) = 0 \implies -\exp\left(\frac{-x_1}{\mu}\right) + \gamma = 0 \tag{3.13}$$

Thus, the optimal policy is:

$$x_1^* = \arg\min_{x_1} \phi(x_1) = -\mu \ln \gamma \tag{3.14}$$

Therefore, the optimal arrival times of the two customers are:

$$\mathbf{t}^* = (t_1^*, t_2^*) = \Big(0, -\mu \ln \gamma\Big) \tag{3.15}$$

As the server availability cost increases relative to the customer waiting cost (i.e., γ increases), the second customer is scheduled to arrive earlier (i.e., t_2^* decreases).

3.5.2 Model for Three Customers

The next simplest case is where there are three customers to be scheduled (i.e., $n = 3$). In this case, there are two unknown variables x_1 and x_2. Without loss of generalisation, can let $\mu = 1$.

By Equation 3.6, the expected waiting times of the three customers are:

$$w_1 = 0 \tag{3.16}$$

$$w_2 = \exp(-x_1) \tag{3.17}$$

$$w_3 = \exp\Big(-(x_1+x_2)\Big)\Big[1 + x_2 + \exp(x_1)\Big] \tag{3.18}$$

By Equation 3.5, the objective function to be minimised is:

$$\begin{aligned}\phi(x_1, x_2) = \exp\Big(-(x_1+x_2)\Big)\Bigg[&1 + x_2 + \exp(x_1) \\ &+ \exp(x_2)\Big(1 - \gamma + \gamma(1 + x_1 + x_2)\exp(x_1)\Big)\Bigg]\end{aligned} \tag{3.19}$$

This objective function is convex for $x_1, x_2 \geq 0$ and $0 \leq \gamma \leq 1$ as:

$$\frac{\partial^2 \phi}{\partial x_1^2} = \exp\Big(- (x_1 + x_2)\Big)\Big(1 + x_2 + \exp(x_2)(1-\gamma)\Big) > 0 \tag{3.20}$$

$$\frac{\partial^2 \phi}{\partial x_2^2} = \exp\Big(- (x_1 + x_2)\Big)\Big(x_2 + \big[\exp(x_1) - 1\big]\Big) \geq 0 \tag{3.21}$$

$$\frac{\partial^2 \phi}{\partial x_1 \partial x_2} = x_2 \exp\Big(- (x_1 + x_2)\Big) \geq 0 \tag{3.22}$$

In a similar way to Section 3.5.1, due to the convexity of the objective function, the optimal policy that minimises $\phi(x_1, x_2)$ can be found by jointly solving:

$$\frac{\partial \phi}{\partial x_1} = 0 \implies \exp(x_2)\Big(\gamma - 1 + \gamma \exp(x_2)\Big) = x_2 + 1 \tag{3.23}$$

$$\frac{\partial \phi}{\partial x_2} = 0 \implies \exp(x_1)\Big(\gamma \exp(x_2) - 1\Big) = x_2 \tag{3.24}$$

Unfortunately, as Pegden and Rosenshine (1990) found, no algebraic solution for (x_1, x_2) exists to these equations, so they need to be solved numerically.

γ	x_1	x_2	$\phi(\mathbf{x})$
1.00	0.00	0.00	3.00
0.95	0.07	0.30	2.99
0.90	0.14	0.41	2.96
0.85	0.22	0.50	2.92
0.80	0.30	0.58	2.87
0.75	0.39	0.65	2.80
0.70	0.48	0.73	2.73
0.65	0.58	0.80	2.65
0.60	0.68	0.88	2.55
0.55	0.78	0.96	2.44
0.50	0.89	1.05	2.32
0.45	1.01	1.15	2.19
0.40	1.13	1.26	2.04
0.35	1.27	1.38	1.88
0.30	1.43	1.51	1.70
0.25	1.61	1.68	1.51
0.20	1.83	1.87	1.29
0.15	2.10	2.13	1.05
0.10	2.48	2.49	0.78
0.05	3.12	3.13	0.46

Table 3.1: Optimal solution for $n = 2$ customers with mean service time $\mu = 1$ over various values of γ found using `scipy.optimize.minimize` in Python. For $\mu \neq 1$, these solutions are the optimal values of μx_1 and μx_2.

Chapter 4

Dynamic Schedule

4.1 Aim

The aim is to choose a schedule of customer arrivals times that minimises the expected cost of the system. The expected cost is a linear combination of the total customers' waiting time and the server's idle time. Instead of choosing a fixed schedule at the start of service as is common in literature, the schedule will be chosen progressively. Immediately after a customer arrives and begins waiting for service, the scheduler chooses the arrival time of the next customer.

4.2 Assumptions

To simplify this problem, need to make several assumptions:

- Service times are independent and identically distributed (iid)
- Each service time has an exponential distribution with mean service time μ
- There is a single server
- The queue operates on a first in, first out (FIFO) basis
- Customers can be scheduled to arrive at any future (or present) time
- Customers are punctual and arrive at their scheduled time

4.3 List of Variables

μ	:	mean service time of each customer
c_W	:	cost of customer's waiting time per unit time
c_I	:	cost of server's idle time per unit time
k	:	current number of customers waiting
j	:	number of customers waiting immediately after the next customer's arrival
n	:	number of customers remaining to be scheduled
a	:	time next customer is scheduled to arrive (relative to current time)
$C_n^*(k)$	:	the expected cost of having k customers waiting and n customers remaining to be scheduled
$C_n(a,k)$	:	the expected cost of having k customers waiting, n customers remaining to be scheduled and the next customer scheduled to arrive after a time units
$p_a(k,j)$	:	the probability of transitioning from k customers waiting to j customers waiting after a time units if the next customer is scheduled to arrive after a time units
$R_a(k,j)$	:	the expected cost of transitioning from k customers waiting to j customers waiting after a time units if the next customer is scheduled to arrive after a time units

4.4 Objective Function

The state (n,k) refers to n customers remaining to be scheduled and k customers in the queue waiting for service. The time the next customer is scheduled to arrive is a, and the number of customers waiting for service immediately after that customer's arrival is j. The expected cost at the current state is a function of the expected cost involved in transitioning to the next state, the expected cost at the next state and the probability of transitioning to the next state over all possible next states.

The expected cost of the state (n,k) where $n \geq 1$ is given by the following form of Bellman's equation:

$$C_n^*(k) = \min_{a\geq 0} C_n(a,k) = \min_{a\geq 0}\left[\sum_{j=1}^{k+1} p_a(k,j)\Big(R_a(k,j) + C_{n-1}^*(j)\Big)\right] \tag{4.1}$$

Equation 4.1 is a recursive equation involving C^*. The optimal solution is found by solving for each customer's arrival time a iteratively. The optimal policy a^* is the customer's arrival time that attains the minimum cost whereby

$$C_n^*(k) = C_n(a^*,k) = \min_{a\geq 0} C_n(a,k) \tag{4.2}$$

It is reasonably intuitive that the minimum cost cannot occur at $a = \infty$. As $a \to \infty$, the probability that the server becomes idle converges to 1. In addition, the expected idle time of the server converges to ∞. As the cost of the server's idle time c_I is strictly positive, the overall expected cost must also converge to ∞ as $a \to \infty$. Thus, $\lim\limits_{a\to\infty} C_n(a,k) = \infty$.

Consider the set of possible policies $\mathcal{A}$ given by

$$\mathcal{A} = \{0\} \bigcup \left\{ a > 0 : \frac{\partial}{\partial a} C_n(a,k) = 0 \right\} \tag{4.3}$$

Solving Equation 4.2 involves solving a nonlinear optimisation problem over a left-closed interval. This solution is equivalent to the solution found by checking the left end point (where $a = 0$) and all points where $\frac{\partial}{\partial a} C_n(a,k) = 0$. Thus, the the optimal policy can be found by solving

$$C_n^*(k) = \min_{a \in \mathcal{A}} C_n(a,k) \tag{4.4}$$

As will be explained later, it is not possible to find a 'nice' closed form for $\frac{\partial}{\partial a} C_n(a,k)$ for general n and k. However, for given values of n and k, it is reasonably efficient to solve $\frac{\partial}{\partial a} C_n(a,k) = 0$. Thus, the expected cost can be found by computing $\frac{\partial}{\partial a} C_n(a,k)$ for each state (n,k). Of course, this method becomes more computationally inefficient, the larger the number of states.

4.4.1 Base Case

Finding the solution iteratively requires a solution for the base case where $n = 0$. If $n = 0$, there are no customers remaining to be scheduled, which implies the server will not be idle for the remaining of service. The cost of state $(0,k)$ (i.e., the base case) is thus the summation of the waiting cost of the k customers in the queue.

Let w_i be the expected waiting time of the customer that is currently in position i in the queue, and c_W be the cost of the customers' waiting time per unit time. The cost of the base case is thus given by

$$\begin{aligned} C_0^*(k) &= c_W \sum_{i=2}^{k} w_i + c_S \sum_{i=1}^{k} \mu \\ &= c_W \sum_{i=2}^{k} \mu(i-1) + c_S k\mu \\ &= \frac{c_W \mu k(k-1)}{2} + c_S k\mu \end{aligned}$$

Scale $C_0^*(k)$ by dividing by $(c_S + c_W)$ and substituting $\gamma = \frac{c_S}{c_S + c_W}$:

$$C_0^*(k) = (1-\gamma) \cdot \frac{\mu k(k-1)}{2} + \gamma k \mu \tag{4.5}$$

4.4.2 Transition Probability

Let S_i be the service time of the customer that is currently in position i in the queue. The service times $S_1, \ldots, S_n$ are iid (independent and identically distributed) exponential random variables with mean μ.

For $r \geq 1$, the waiting time of the customer in position $(r+1)$ in the queue is:

$$X = \sum_{i=1}^{r} S_i \sim \text{Erlang}(j, \mu) \tag{4.6}$$

which has the pdf:

$$f(x; r) = \frac{1}{\mu \cdot (r-1)!} \left(\frac{x}{\mu}\right)^{r-1} \exp\left(\frac{-x}{\mu}\right) \tag{4.7}$$

Let W_t be a Poisson Point Process with $W_t \sim$ Poisson $\left(\frac{t}{\mu}\right)$. For $r \geq 1$, the probability that the customer currently in position $(r+1)$ in the queue waits longer than a time units before service is equal to the probability that W_a is smaller than r, such that

$$\mathbb{P}(X > a) = \mathbb{P}(W_a < r) \tag{4.8}$$

The transition probability $p_a(k, j)$ is the probability that the queue length changes from k customers initially to j customers on the arrival of the next customer after a time units. In other words, it is the probability that there are $k - (j-1)$ departures from the queue over a time interval of length a. Computing this probability requires the cdf of the Erlang distribution, which is calculated (for $a > 0$) as follows:

$$\begin{aligned}
F(a; r) &= \mathbb{P}(X \leq a) \\
&= 1 - \mathbb{P}(X > a) \\
&= 1 - \mathbb{P}(W_a < r) \\
&= 1 - \sum_{i=0}^{r-1} \mathbb{P}(W_a = i) \\
&= 1 - \sum_{i=0}^{r-1} \frac{1}{i!} \left(\frac{a}{\mu}\right)^i \exp\left(\frac{-a}{\mu}\right)
\end{aligned}$$

Moreover, $F(0;r) = \mathbb{P}(X = 0) = 0$ by definition. Therefore, the cdf of the Erlang distribution is defined as

$$F(a;r) = \begin{cases} 1 - \sum_{i=0}^{r-1} \frac{1}{i!} \left(\frac{a}{\mu}\right)^i \exp\left(\frac{-a}{\mu}\right) & \text{where } a > 0 \\ 0 & \text{where } a = 0 \end{cases} \tag{4.9}$$

Can now compute the transition probability on a case by case basis. The full derivation is included in the appendix and only the final equation is presented here.

$$p_a(k,j) = \begin{cases} \mathbb{1}(j = 1) & \text{where } k = 0 \\ F(a;k) & \text{where } k \geq 1, j = 1 \\ F(a;k-j+1) - F(a;k-j+2) & \text{where } k \geq 1, 2 \leq j \leq k \\ 1 - F(a;1) & \text{where } k \geq 1, j = (k+1) \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

4.4.3 Expected Transition Cost

The cost involved in transitioning from state (n,k) to state $(n-1,j)$ in a time units is a linear combination of the expected total waiting time of the customers during the transition and the expected total server availability time during the transition as in Chapter 3. During the transition, the server is available for the entire time interval, so the total server availability time during the transition is always a.

The expected total waiting time of the customers depends on the conditional expectation of the Erlang distribution, which is derived here (for $a > 0$).

$$\begin{aligned} G(a;k) &= \mathbb{E}[X|X \leq a] \\ &= \int x \cdot \mathbb{P}(X \in dx | X \leq a) \\ &= \int x \cdot \frac{\mathbb{P}(X \in dx, X \leq a)}{\mathbb{P}(X \leq a)} \\ &= \frac{1}{F(a;r)} \int_0^a x \cdot f(x;r) dx \\ &= \frac{1}{F(a;r)} \int_0^a x \cdot \frac{1}{\mu \cdot (r-1)!} \left(\frac{x}{\mu}\right)^{r-1} \exp\left(\frac{-x}{\mu}\right) \\ &= \frac{\mu r}{F(a;r)} \int_0^a \frac{1}{\mu \cdot r!} \left(\frac{x}{\mu}\right)^{r} \exp\left(\frac{-x}{\mu}\right) \\ &= \mu r \cdot \frac{F(a;r+1)}{F(a;r)} \end{aligned}$$

In addition, $G(0;r) = \mathbb{E}[X|X \leq 0] = 0$ by definition. Therefore, the conditional expectation of the Erlang distribution is defined as

$$G(a;r) = \begin{cases} \mu r \cdot \frac{F(a;r+1)}{F(a;r)} & \text{where } a > 0 \\ 0 & \text{where } a = 0 \end{cases} \tag{4.11}$$

This expression for the conditional expectation makes intutive sense. The mean of the Erlang distribution is $\mathbb{E}[X] = \mu r$. In addition, $\forall a > 0$, $\frac{F(a;r+1)}{F(a;r)} < 1$. Thus, $\forall a > 0$, $\mathbb{E}[X|X \leq a] < \mathbb{E}[X]$ as expected. Moreover,

$$\lim_{a\to\infty} \frac{F(a;r+1)}{F(a;r)} = \frac{\lim\limits_{a\to\infty} F(a;r+1)}{\lim\limits_{a\to\infty} F(a;r)} = \frac{1}{1} = 1 \tag{4.12}$$

Thus, $\lim\limits_{a\to\infty} \mathbb{E}[X|X \leq a] = \mathbb{E}[X]$ as expected.

In a similar way to the transition probability, the expected transition cost is derived on a case by case basis. The per unit time cost of the customers' waiting time and the server's idle time are c_W and c_I respectively. The full derivation is included in the appendix and the final equation is presented here.

$$\begin{aligned} &R_a(k,j) \\ &= \begin{cases} \gamma a & \text{where } k \in \{0,1\} \\ (1-\gamma) \cdot \frac{G(a;k)(k-1)}{2} + \gamma a & \text{where } k \geq 2, j = 1 \\ (1-\gamma) \left[a(j-2) + \frac{G(a;k-(j-1))[k-(j-2)]}{2}\right] + \gamma a & \text{where } k \geq 2, 2 \leq j \leq k \\ (1-\gamma) \cdot a(j-2) + \gamma a & \text{where } k \geq 2, j = (k+1) \end{cases} \end{aligned} \tag{4.13}$$

4.5 Example Models

4.5.1 State (3, 2)

Assume that the mean service time μ is 1. In addition, assume that the per unit time costs of the customers' waiting time and server's idle time (c_W and c_I respectively) are both 1. Under these assumptions, the cost of state $(3,2)$ for various values of a is given by the following expression, which is continuous for $a \geq 0$.

$$C_3(a,2) = \begin{cases} 9.88121 & \text{where } a = 0 \\ \frac{\mathrm{e}^{-a}\left(\mathrm{e}^{a} a(2.29362-0.5a)-(0.5a+3.29362)a-1.15888\mathrm{e}^{a}+(a+5.52005)\mathrm{e}^{2a}-4.36117\right)}{\mathrm{e}^{a}-1} & \text{where } a > 0 \end{cases}$$

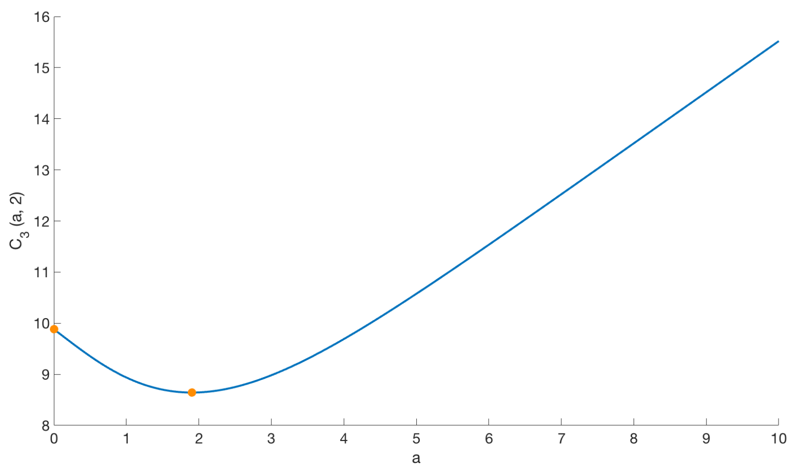

Figure 4.1: Cost of state with n = 3, k = 2 for various values of a

This equation is plotted for $a \in [0, 10]$ to visualise the relationship between the cost and a. The resulting figure is Figure 4.1.

The only value of a where $\frac{\partial}{\partial a} C_3(a, 2) = 0$ is 1.90481. Thus, the set of possible policies $\mathcal{A}$ is $\{0, 1.90481\}$ whereby there are two possible policies labelled as orange dots on Figure 4.1. Moreover, note that as a increases beyond 2, the cost increases approximately linearly as the expected waiting time of the next customer and the expected idle time of the server both increase linearly.

The optimal policy a^* that minimises $C_3(a, 2)$ is 1.90481 where the cost is 8.64338. If on arrival of a customer, there are 2 customers in the queue and 3 customers remaining to be scheduled, then the next customer should be scheduled to arrive in 1.90481 time units. This is slightly below the expected service time of the two customers in the queue to account for the idle cost of the server.

4.5.2 Model for Six Customers

Assume there are 6 customers that need to be scheduled for service who all have a mean service time μ of 1. The initial state is $(6, 0)$, and the possible states during service are all states in the set

$$\left\{(n, k) \in \{0, 1, \ldots, 6\}^2 : n + k \leq 6\right\} \tag{4.14}$$

Table 4.1 displays the calculated expected costs for each possible state if there are 6 total customers (assuming $c_W = c_I = 1$). The cost of the initial state is 10.0101, thus the expected cost of servicing six customers with an alterable schedule is 10.0101.

customers to be scheduled (n)	queue length (k) 0	1	2	3	4	5	6
0	0	1	3	6	10	15	21
1	1	2.69315	4.97937	8.04949	11.9605	16.7418	
2	2.69315	4.52005	6.81367	9.88121	13.7875		
3	4.52005	6.35018	8.64338	11.7105			
4	6.35018	8.18013	10.4733				
5	8.18013	10.0101					
6	10.0101						

Table 4.1: Cost of possible states for 6 total customers

Note that in this table, for $n \geq 1$, $C^*_n(0) = C^*_{n-1}(1)$. This is due to the fact that if there are no customers currently waiting (e.g., initially), then it is always optimal to schedule the next arrival immediately. The expected cost thus doesn't change as the next arrival occurs immediately.

The worst state on this table is the state $(0, 6)$, which is all six customers waiting for serviced. This can occur if all six customers are scheduled to arrive immediately (to ensure no idle time), or the first customer has an extremely long service time.

customers to be scheduled (n)	queue length (k) 0	1	2	3	4	5
1	0	0.693147	1.75711	2.82577	3.85505	4.8492
2	0	0.826902	1.90223	2.95223	3.95872	
3	0	0.83013	1.90481	2.95403		
4	0	0.82995	1.90455			
5	0	0.829925				
6	0					

Table 4.2: Optimal policy for each possible states for 6 total customers

Table 4.2 displays the corresponding arrival times for the costs in Table 4.1. This table does not include any optimal times for $n = 0$, as there is no next customer to schedule in those states.

The first pattern to notice is (as discussed earlier) if there are no customers currently waiting (i.e., $k = 0$), then the optimal solution is to schedule the next arrival immediately. This makes intutive sense as scheduling the next arrival immediately minimises the future idle time and ensures that the waiting time of the next customer is simpler their service time.

As k increases for constant n, the optimal schedule time increases approxiimately proportionately. The optimal a^* increases by approximately μ for each extra k to minimise the waiting time of the next customer.

Finally, for $n \geq 1$, the value of a^* such that $C_n^*(k) = C_n(a^*, k)$ is always less than μk (i.e., the expected time for the queue to be empty). This is due to the cost of the idle time of the customer leading to the desire for the next customer to arrive just before the queue becomes empty.

Chapter 5

Value Dynamic Schedule

Value dynamic schedule goes here.

Chapter 6

Simulation Studies

Simulation studies goes here.

Chapter 7

Conclusion

Conclusion goes here.

Appendix A

Dynamic Cost Derivation

A.1 Transition Probability

A.1.1 Case 1 $k = 0$

$$p_a(k, j) = \mathbb{1}(j = 1)$$

A.1.2 Case 2 $k \geq 1, j = 1$

$$p_a(k, j) = \mathbb{P}\left(\sum_{i=1}^{k} S_i \leq a\right) = F(a; k)$$

A.1.3 Case 3 $k \geq 1, 2 \leq j \leq k$

$$
\begin{aligned}
& p_a(k,j) \\
&= \mathbb{P}\left(\sum_{i=1}^{k-(j-1)} S_i \leq a, \sum_{i=1}^{k-(j-1)+1} S_i > a\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{k-(j-1)} S_i \leq a, \sum_{i=1}^{k-(j-1)} S_i + S_{k-(j-1)+1} > a\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{k-(j-1)} S_i \leq a, S_{k-(j-1)+1} > a - \sum_{i=1}^{k-(j-1)} S_i\right) \\
&= \int \mathbb{P}\left(\sum_{i=1}^{k-(j-1)} S_i \leq a, S_{k-(j-1)+1} > a - \sum_{i=1}^{k-(j-1)} S_i \,\middle|\, \sum_{i=1}^{k-(j-1)} S_i = z\right) \mathbb{P}\left(\sum_{i=1}^{k-(j-1)} S_i \in dz\right) \\
&= \int_0^\infty \mathbb{P}(z \leq a, S_{k-(j-1)+1} > a - z) f\big(z; k-(j-1)\big)\, dz \\
&= \int_0^a \mathbb{P}(S_{k-(j-1)+1} > a - z) f\big(z; k-(j-1)\big)\, dz \\
&= \int_0^a f\big(z; k-(j-1)\big)\big(1 - \mathbb{P}(S_{k-(j-1)+1} \leq a - z)\big)\, dz \\
&= \int_0^a f\big(z; k-(j-1)\big)\big(1 - F(a-z;1)\big)\, dz \\
&= \int_0^a \frac{1}{\mu \cdot \big(k-(j-1)-1\big)!}\left(\frac{z}{\mu}\right)^{k-(j-1)-1} \exp\left(\frac{-z}{\mu}\right) \cdot \exp\left(\frac{-(a-z)}{\mu}\right)\, dz \\
&= \frac{1}{\big(k-(j-1)-1\big)!}\left(\frac{1}{\mu}\right)^{k-(j-1)} \exp\left(\frac{-a}{\mu}\right) \int_0^a z^{k-(j-1)-1}\, dz \\
&= \frac{1}{\big(k-(j-1)-1\big)!}\left(\frac{1}{\mu}\right)^{k-(j-1)} \exp\left(\frac{-a}{\mu}\right) \cdot \frac{a^{k-(j-1)}}{k-(j-1)} \\
&= \frac{1}{\big(k-(j-1)\big)!}\left(\frac{a}{\mu}\right)^{k-(j-1)} \exp\left(\frac{-a}{\mu}\right) \\
&= \left[1 - \sum_{i=0}^{k-(j-1)-1} \frac{1}{i!}\left(\frac{a}{\mu}\right)^i \exp\left(\frac{-a}{\mu}\right)\right] - \left[1 - \sum_{i=0}^{k-(j-1)} \frac{1}{i!}\left(\frac{a}{\mu}\right)^i \exp\left(\frac{-a}{\mu}\right)\right] \\
&= F\big(a; k-(j-1)\big) - F\big(a; k-(j-1)+1\big) \\
&= F(a; k-j+1) - F(a; k-j+2)
\end{aligned}
$$

A.1.4 Case 4 $k \geq 1, j = (k+1)$

$$p_a(k,j) = \mathbb{P}(S_1 > a) = 1 - \mathbb{P}(S_1 \leq a) = 1 - F(a;1)$$

A.1.5 All Other Cases

$$p_a(k,j) = 0$$

A.1.6 Summary

These results can be summarised as:

$$p_a(k,j) = \begin{cases} \mathbb{1}(j=1) & \text{where } k = 0 \\ F(a;k) & \text{where } k \geq 1, j = 1 \\ F(a;k-j+1) - F(a;k-j+2) & \text{where } k \geq 1, 2 \leq j \leq k \\ 1 - F(a;1) & \text{where } k \geq 1, j = (k+1) \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

A.2 Expected Transition Cost

A.2.1 Case 1 $k \in \{0,1\}$

$$R_a(k,j) = c_S a = \gamma a$$

A.2.2 Case 2 $k \geq 2, j = 1$

$$
\begin{aligned}
R_a(k,j) &= c_W \sum_{i=2}^{k} \mathbb{E}\left[\sum_{l=1}^{i-1} S_l \,\middle|\, \sum_{n=1}^{k} S_n \leq a\right] + c_S a \\
&= c_W \sum_{i=2}^{k} \sum_{l=1}^{i-1} \mathbb{E}\left[S_l \,\middle|\, \sum_{n=1}^{k} S_n \leq a\right] + c_S a \\
&= c_W \mathbb{E}\left[S_1 \,\middle|\, \sum_{n=1}^{k} S_n \leq a\right] \sum_{i=2}^{k} (i-1) + c_S a \\
&= \frac{c_W k(k-1)}{2} \mathbb{E}\left[S_1 \,\middle|\, \sum_{n=1}^{k} S_n \leq a\right] + c_S a \\
&= \frac{c_W (k-1)}{2} \mathbb{E}\left[\sum_{n=1}^{k} S_n \,\middle|\, \sum_{n=1}^{k} S_n \leq a\right] + c_S a \\
&= \frac{c_W G(a;k)(k-1)}{2} + c_S a \\
&= (1-\gamma) \cdot \frac{G(a;k)(k-1)}{2} + \gamma a
\end{aligned}
$$

A.2.3 Case 3 $k \geq 2, 2 \leq j \leq k$

$$
\begin{aligned}
&R_a(k,j)\\
&= c_W \sum_{i=1}^{j-2} a + c_W \sum_{i=2}^{k-(j-2)} \mathbb{E}\left[\sum_{l=1}^{i-1} S_l \,\middle|\, \sum_{n=1}^{k-(j-1)} S_n \leq a\right] + c_S a\\
&= c_W a(j-2) + c_W \sum_{i=2}^{k-(j-2)} \sum_{l=1}^{i-1} \mathbb{E}\left[S_l \,\middle|\, \sum_{n=1}^{k-(j-1)} S_n \leq a\right] + c_S a\\
&= c_W a(j-2) + c_W \mathbb{E}\left[S_1 \,\middle|\, \sum_{n=1}^{k-(j-1)} S_n \leq a\right] \sum_{i=2}^{k-(j-2)} (i-1) + c_S a\\
&= c_W a(j-2) + \frac{c_W\big[k-(j-1)\big]\big[k-(j-2)\big]}{2} \mathbb{E}\left[S_1 \,\middle|\, \sum_{n=1}^{k-(j-1)} S_n \leq a\right] + c_S a\\
&= c_W a(j-2) + \frac{c_W\big[k-(j-2)\big]}{2} \mathbb{E}\left[\sum_{n=1}^{k-(j-1)} S_n \,\middle|\, \sum_{n=1}^{k-(j-1)} S_n \leq a\right] + c_S a\\
&= c_W \left[a(j-2) + \frac{G\big(a; k-(j-1)\big)\big[k-(j-2)\big]}{2}\right] + c_S a\\
&= (1-\gamma)\left[a(j-2) + \frac{G\big(a; k-(j-1)\big)\big[k-(j-2)\big]}{2}\right] + \gamma a
\end{aligned}
$$

A.2.4 Case 4 $k \geq 2, j = (k+1)$

$$
R_a(k,j) = c_W \sum_{i=1}^{j-2} a + c_S a = c_W a(j-2) + c_S a = (1-\gamma)\cdot a(j-2) + \gamma a
$$

A.2.5 Summary

These results can be summarised as:

$$R_a(k,j) = \begin{cases} \gamma a & \text{where } k \in \{0,1\} \\ (1-\gamma)\cdot \frac{G(a;k)(k-1)}{2} + \gamma a & \text{where } k \geq 2, j = 1 \\ (1-\gamma)\left[a(j-2) + \frac{G(a;k-(j-1))[k-(j-2)]}{2}\right] + \gamma a & \text{where } k \geq 2, 2 \leq j \leq k \\ (1-\gamma)\cdot a(j-2) + \gamma a & \text{where } k \geq 2, j = (k+1) \end{cases} \tag{A.2}$$

Bibliography

Bailey, Norman TJ (1952). “A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 185–199.

Rockart, John F and Paul B Hofmann (1969). “Physician and patient behavior under different scheduling systems in a hospital outpatient department”. In: *Medical Care* 7.6, pp. 463–470.

Gupta, Ishwar, Juan Zoreda, and Nathan Kramer (1971). “Hospital manpower planning by use of queueing theory”. In: *Health Services Research* 6.1, pp. 76–82.

Walter, SD (1973). “A comparison of appointment schedules in a hospital radiology department”. In: *British Journal of Preventive & Social Medicine* 27.3, pp. 160–167.

Kao, Edward PC and Grace G Tung (1981). “Bed allocation in a public health care delivery system”. In: *Management Science* 27.5, pp. 507–520.

O’Keefe, Robert M (1985). “Investigating outpatient departments: Implementable policies and qualitative approaches”. In: *Journal of the Operational Research Society* 36.8, pp. 705–712.

Goldsmith, Jeff (1989). “A radical prescription for hospitals”. In: *Harvard Business Review*.

Pegden, Claude Dennis and Matthew Rosenshine (1990). “Scheduling arrivals to queues”. In: *Computers & Operations Research* 17.4, pp. 343–348.

Babes, Malika and GV Sarma (1991). “Out-patient queues at the Ibn-Rochd health centre”. In: *Journal of the Operational Research Society* 42.10, pp. 845–855.

Ho, Chrwan-Jyh and Hon-Shiang Lau (1992). “Minimizing total cost in scheduling outpatient appointments”. In: *Management Science* 38.12, pp. 1750–1764.

Stein, William E and Murray J Côté (1994). “Scheduling arrivals to a queue”. In: *Computers & Operations Research* 21.6, pp. 607–614.

Huarng, Fenghueih and Mong Hou Lee (1996). "Using simulation in out-patient queues: A case study". In: *International Journal of Health Care Quality Assurance* 9.6, pp. 21–25.

Bennett, Joanne C and DJ Worthington (1998). "An example of a good but partially successful OR engagement: Improving outpatient clinic operations". In: *Interfaces* 28.5, pp. 56–69.

Cayirli, Tugba and Emre Veral (2003). "Outpatient scheduling in health care: A review of literature". In: *Production and Operations Management* 12.4, pp. 519–549.

Mondschein, Susana V and Gabriel Y Weintraub (2003). "Appointment policies in service operations: A critical analysis of the economic framework". In: *Production and Operations Management* 12.2, pp. 266–286.

DeLaurentis, Po-Ching et al. (2006). "Open access appointment scheduling - An experience at a community clinic". In: *IIE Annual Conference*. Institute of Industrial Engineers.

Green, Linda (2006). "Queueing analysis in healthcare". In: *Patient flow: reducing delay in healthcare delivery*. Springer, pp. 281–307.

Mendel, Sharon (2006). "Scheduling arrivals to queues: A model with no-shows". MA thesis. Tel-Aviv University.

Fiems, Dieter, Ger Koole, and Philippe Nain (2007). "Waiting times of scheduled patients in the presence of emergency requests". In: *Technisch Rapport*.

Fomundam, Samuel and Jeffrey W Herrmann (2007). *A survey of queuing theory applications in healthcare*. University of Maryland, The Insitute for Systems Research.