MANAGING SUPPLY IN THE INDIAN CONTEXT: FLEXIBLE WORKERS, FULL-TIME EMPLOYEES AND FREELANCERS

A Project Report Submitted

for the Course

MA498 Project I and MA499 Project II

by

Manne Hema Priya

(Roll No. 170123032)

Tanvi Ohri

(Roll No. 170123051)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

April 2021

CERTIFICATE

This is to certify that the work contained in this report entitled **"Managing Supply in the Indian Context: Flexible Workers, Full-Time Employees And Freelancers"** submitted by **Manne Hema Priya** (**Roll No: 170123032**) and **Tanvi Ohri** (**Roll No: 170123051**) to Department of Mathematics, Indian Institute of Technology Guwahati towards partial requirement of Bachelor of Technology in Mathematics and Computing has been carried out by them under my supervision.

It is also certified that, along with literature survey, a few new results are established by the students under the project.

Turnitin Similarity: 18%

Guwahati - 781 039
April 2021

(Dr. N.Selvaraju)
Project Supervisor

ABSTRACT

The main aim of the project is to adapt the model proposed by Dong and Ibrahim (2020) [2] for the Indian market. There is an important extension we propose for this. With the increasing trend of startups in India, we feel we need to include another type of employee- the freelancer. The flexible employee is hired for a short interval, say when we typically expect higher traffic or a certain kind of work. The freelancers are hired for a particular task, that is, the profession's or startup's area of expertise. This is delegation of a task, it need not have a certain minimum duration. Freelancers have the freedom to decide their work schedules. There is a rise in the number of firms that are staffing a blended workforce. A blended workforce is a clever business strategy of employing some flexible workers with full-time employees, and now, freelancers as well. An optimal staffing strategy is to be devised to combat supply-side uncertainty and staffs freelancers, flexible workers and full-time employees. The strategy must take into account both quality and efficiency factors. In the first phase of the project, we surveyed some of the preliminary works in this area. We studied staffing under parameter uncertainty [1] and staffing with self-scheduling servers [6]. In the second phase of the project, we have adopted the model proposed by Dong and Ibrahim in [2] and added a new type of employee to the model- the freelancer.

Contents

List of Figures **viii**

List of Tables **ix**

1 Key Concepts **1**

1.1 Basics of Queuing Theory 1

1.1.1 Arrival Process 2

1.1.2 Service Mechanism 2

1.1.3 Queue Characteristics 3

1.1.4 Kendall's Notation 4

1.2 Performance Measures 4

1.3 Some results for M/M/1 system 5

1.4 The Fluid Approximation 6

1.4.1 Some General Relations 8

1.5 The Square Root Staffing Law 10

1.6 The Newsvendor Problem 10

1.6.1 Problem Formulation 11

1.6.2 Solution of the Newsvendor Problem 11

2 Managing Supply Under Demand-Side Parameter Uncertainty 13

2.1 Reviewing Existing Models 13

2.2 Proposing an Approximate Solution based on the Newsvendor Problem 15

2.2.1 Formulating the Problem 15

2.2.2 Proposing an Approximate Solution 17

2.3 Analysing the Performance of the Proposed Solution 19

2.3.1 Deterministic Arrival Rates 20

2.3.2 Uncertain Arrival Rates 23

2.4 Some Important Results 26

3 Optimizing the system management cost in the presence of self scheduling servers 28

3.1 Modelling framework : 29

3.2 System Manager's Problem 30

3.3 Benchmark Case : No Self Scheduling 31

3.4 The actual problem : Servers with Self- Scheduling capability 33

3.5 Exponential Abandonment and Non-equal Augmented Arrival Rates 36

4 Capacity sizing with only self scheduling serves under Demand-Side Parameter Uncertainty 38

4.1 Modelling Framework 39

4.1.1 A random number of servers 39

4.2 This papers model 40

4.3 The long term staffing problem 41

4.4 Fluid Approximation 42
4.4.1 The problem formulation 42
4.4.2 Asymptotic Accuracy 43
4.4.3 Optimal staffing policy 43
4.5 Stochastic-Fluid Approximation 43
4.5.1 The problem formulation 43
4.5.2 Asymptotic Accuracy 44
4.5.3 Optimal Staffing Policy 44

5 Capacity Sizing with a Blended Workforce 47
5.1 The Staffing Problem for Blended Workforce 48
5.2 The Fluid Approximation 49
5.2.1 Problem Formulation 49
5.2.2 Asymptotic Accuracy 49
5.2.3 Optimal Staffing Policy 49
5.3 The Stochastic-Fluid Approximation 50
5.3.1 Problem Formulation 50
5.3.2 Asymptotic Accuracy 50
5.3.3 Optimal Staffing Policy 51

6 Capacity Sizing with a Blended Workforce and Freelancers 53
6.1 The Need and Strategy for Hiring Freelancers 53
6.2 Optimal Staffing Strategy for Freelancers 55
6.3 Numerical Results 56
6.3.1 Firm Persepective : Cost Reduction 56
6.3.2 Customer Perspective : Quality of Service 57

Bibliography **60**

List of Figures

2.1 Performance of the Proposed Solution on Deterministic Arrival Rates 22
2.3 Performance of the Proposed Solution on Uncertain Arrival Rates 25

6.1 Total Staffing Costs 57
6.3 Expected number of customers that abandon the system . . . 58

List of Tables

2.1 Performance of the Proposed Solution on Deterministic Arrival Rates 22
2.2 Performance of the Proposed Solution on Uncertain Arrival Rates 24

Chapter 1

Key Concepts

This section has been added to cater to a reader who does not have prior knowledge about queuing theory .It includes some basic queuing models, approximations, laws and problems that have been used throughout this project.

1.1 Basics of Queuing Theory

The main reason for the usage of queues in modelling service systems is because the resources are limited. Hence it leads to the formation of queues.In these queuing systems, we often see a trade-off between the cost of maintaining the system and customer service, while modelling there is an attempt to optimize on both ends .To analyze any queuing system, we need to know the following parameters.

1.1.1 Arrival Process

We look at this in terms of

- Inter arrival times of different customers
- How the customers arrive (individually or in batches)
- Whether the number of potential customers (also known as calling population) is finite or infinite.

The arrival times could be random or scheduled, in random arrival times one of the important models is the exponential arrival process where the difference between two arrivals is distributed exponentially in a Poisson distribution

1.1.2 Service Mechanism

We look at this in terms of

- The distribution of service times
- No. of servers in the system and how that changes
- Relation between servers in different service centers and preemption in processes.

In general, we assume that the service times are independent of other factors and that they are exponential. We also assume that the number of servers is fixed.

1.1.3 Queue Characteristics

- **Queue Capacity**

  The queue capacity can be finite or infinite.

- **Queue Behaviour**

  The actions of customers when they are waiting in the queue for the service to begin .There are various types of behaviours like

  - **Balk**: A behaviour in which the customers won't join the system when the queue is too long
  - **Renege or Abandon**: A behaviour in which the customers are impatient and leave before getting served.
  - **Jockey**: A behaviour in which the customers move from one queue to a shorter queue

  Customer abandonment or impatient customers is a important factor which we look at when modelling queuing systems.The distribution of customer abandonment times is also taken into consideration when optimizing a system.

- **Queue Discipline**

  The algorithm which determines which customer gets served when a server gets free, there are various algorithms, with their own merits and demerits . For example,

  - Service according to priority (PR).
  - Shortest processing time first (SPT)

- Last in first out (LIFO)
- First in first out or First come first serve (FIFO)
- Customers are served in a random orders (SIRO)

1.1.4 Kendall's Notation

In this notation , any queuing system can be represented as $A/B/c/N/K/D$ where

D is the queue discipline

K is the size of potential number of customers in the system

N is the capacity of the queue

c represents the number of servers in the system

B represents the distribution of service process

A represents the distribution of arrival process

In general , when not specified the last three parameters N, K, D are taken as ∞, ∞, FIFO respectively

For exponential distribution, we use M and for the general distribution we use G, for the arrival and service time distribution.

1.2 Performance Measures

For analysing the performance of a queuing system , we need to know the following values

- Average waiting time
- Average occupancy of queue

- Server utilization
- Probability of waiting time exceeding a given time
- Probability of queue occupancy exceeding a given occupancy
- Service time of n^{th} customer
- Arrival time between n^{th} customer and $(n-1)^{th}$ customer
- Parameters of the system in the long run

The answers to the above questions can help us in designing a optimal system to minimize some cost in the system.

1.3 Some results for M/M/1 system

Consider a model with poisson distributed customer arrival (arrival rate: λ) and exponential service times (rate: μ).
The inter-arrival time distribution is $p_\lambda(t) = \lambda e^{-\lambda t}$
The service time distribution is $p_\mu(t < t_0) = 1 - e^{-\mu t_0}$
For a stable system , the rate of customer arrival should be less than or equal to the rate at which they are being processed $\lambda < \mu$
We define system utilization as $\rho = \frac{\lambda}{\mu} = P[system\,is\,busy]$
Let's define S_n as the state with n customers in the system ($n-1$ in queue , the n^{th} being served) and let $p_n = P[n\,customers\,in\,the\,system]$, for steady state. Then,
$rate_{in} = rate_{out} \rightarrow p_n\lambda = p_{n+1}\mu \rightarrow p_{n+1} = p_0\rho_{n+1}$

Since $p_0 = 1 - \rho$

$$P[n\,customers\,in\,the\,system] = p_n = \rho_n(1-\rho)$$

To find the average number of customers in the system,

$$\bar{N} = \sum k.(\rho_k(1-\rho)) = (1-\rho)\sum k\rho_k = \frac{\rho}{1-\rho} \quad [7] \tag{1.1}$$

Average system time is given by,

$$T = \frac{\bar{N}}{\lambda}\ (Little's\,Result) = \frac{1}{\mu - \lambda} \tag{1.2}$$

For more results and further explanation , refer *Fundamentals of Queuing Systems* [7].

1.4 The Fluid Approximation

In most practical scenarios, the value of number of arrivals is quite large with respect to the time period under consideration. For such systems, customer flow may be approximated as the flow of a continuous fluid. This means we don't consider customers as discrete entities.

Let's imagine the fluid queue's physical analogue to gain better understanding.

Arrivals to the queue (Arrival Rate: λ) $\leftrightarrow$ Water flowing out of a tap

Server $\leftrightarrow$ Drain

Queue $\leftrightarrow$ Water in the sink

Processing Capacity of one server (Service Rate: μ) $\leftrightarrow$ Maximum flow rate through one drain

What does this mean w.r.t. overcrowding in the queue ($\leftrightarrow$ accumulation of water in the sink)?

We are answering this based on single-server model. If we have multiple servers($\leftrightarrow$ drains) then the processing($\leftrightarrow$ drainage) capacity becomes $b * \mu$ where b is the number of servers/drains.

- If $\lambda < \mu$, water gets drained out faster than it flows out of the tap, no accumulation of water $\leftrightarrow$ queue size = 0.

- If $\lambda > \mu$, water gets drained out slower than it flows out of the tap, there is accumulation of water which increases up over time $\leftrightarrow$ queue size increases over time.

- If λ varies over time, the water level varies with time $\leftrightarrow$ queue size varies over time. This is one of the most important applications of the fluid approximation. It has the ability to handle uncertain arrivals which is absent in many other models.

1.4.1 Some General Relations

Let

$A(t)$: cumulative number of arrivals by time t.

$$A(t) = \int_0^t \lambda(u)du \quad [7]$$

$D(t)$: cumulative number of departures by time t.

Since the fluid is assumed to be infinitely divisible, the values of $A(t)$ and $D(t)$ are continuous. They are also non-decreasing and $A(t) \geq D(t)$ because water can not drain out before it comes out of the tap. Thus, queue length $= A(t) - D(t) \geq 0$. Also, if FCFS queue discipline is in place, the time n^{th} customer spends in the queue $= D^{-1}(n) - A^{-1}(n)$. This formula holds even if n is not a whole number since we consider customers to be infinitely divisible. So,

Time n^{th} customer spends in the queue $= D^{-1}(n) - A^{-1}(n)$ [7]

Total time spent in queue by all the customers $= \int_0^N [D^{-1}(n) - A^{-1}(n)]dn$ [7]

This can also be looked at in terms of the queue lengths.

Total time spent in queue by all the customers $= \int_0^T [A(t) - D(t)]dt$ [7]

Now,

$$\int_0^N [D^{-1}(n) - A^{-1}(n)]dn = \int_0^T [A(t) - D(t)]dt \quad [7]$$

$$=> \frac{N}{T}.\frac{1}{N}\int_0^N [D^{-1}(n) - A^{-1}(n)]dn = \frac{1}{T}\int_0^T [A(t) - D(t)]dt \quad [7]$$

$$=> \lambda.W = L \quad [7]$$

where L is the average queue length, $\lambda(=\frac{N}{T})$ is the average arrival rate over the time period under observation and W is the average delay per customer.

Remark 1.4.1. We have assumed that the system starts and ends in an empty state.

To know the length of the queue as a function time, we need to know the function $D(t)$

Remark 1.4.2. We know that the departures cannot be accumulated at a rate greater than μ, so

$$\frac{dD(t)}{dt} \leq \mu$$

Remark 1.4.3. When the queue is non-empty, we have $A(t) > D(t)$ and the server serves at maximum capacity , so

$$\frac{dD(t)}{dt} = \mu$$

Remark 1.4.4. When the queue is empty, we have $A(t) = D(t)$ and the rate of departure would not be greater the arrival rate $\lambda(t)$, so

$$\frac{dD(t)}{dt} = min(\lambda(t), \mu)$$

Accumulating all the above remarks , we have

$$\frac{dD(t)}{dt} = \begin{cases} \mu & A(t) > D(t) \\ min(\lambda(t), \mu) & A(t) = D(t) \end{cases} \qquad [7]$$

Using the above function , we can integrate the differential and find $D(t)$ us-

ing graph method and then find the length of queue as a function of time. For further results and explanation refer *Fundamentals of Queueing Systems*[7]

1.5 The Square Root Staffing Law

In queuing theory, the square root staffing law is a rule-of-thumb used to compute the capacity that would be required to meet an increased amount of traffic in the queue. The law is widely used to help in capacity planning in the QED (Quality-and-Efficiency-Driven) regime. More formally, the question we are dealing with is that if the current quality of service is deemed to be acceptable and necessary then how much must the capacity be increased to serve the increased demand. Rule of thumb answers this question. It says that to hold quality of service constant, you must have a variability hedge equal to the square root of the load increase. This law dates back to Erlang's work in 1917 [3].

1.6 The Newsvendor Problem

The newsvendor problem is a familiar problem in operational research that has application in determining the optimal level of inventory one should stock. Fixed prices and uncertain demand are typical characteristics of the newsvendor problem. This problem gets its name because it mirrors the situation a newspaper vendor faces while deciding the number of copies of the daily paper to buy. Since the demand is uncertain, unsold copies will be wasted. Order is placed before demand materializes and there is a cost

incurred for ordering too much as well as for ordering too few items. These costs are analogous to the cost for idle servers and cost for poor customer service in the staffing problem.

1.6.1 Problem Formulation

Number of units bought is denoted by Q. The per unit selling price is denoted by P, the per unit buying price is denoted by C and the per-unit amount the newsvendor can get for an unsold unit, also known as the salvage price, is denoted by S. F is the cumulative distribution function of demand. The per-unit cost for any items that cannot be sold is called the overage cost (C_o) and the per-unit cost for not meeting demand is called the underage cost (C_u).

1.6.2 Solution of the Newsvendor Problem

The optimal Q is given by:

$$F(Q^*) = \frac{C_u}{C_u + C_o}$$

This is called the Critical Fractile Formula. Now,

$$C_o = C - S \;\; and \;\; C_u = P - C$$

So,

$$Q^* = F^{-1}(\frac{C_u}{C_u + C_o}) = F^{-1}(\frac{P - C}{P - C + C - S}) = F^{-1}(\frac{P - C}{P - S})$$

The solution we see here was originally proposed in the works of Harrison and Zeevi (2005) [5] and Whitt (2006) [8].

Chapter 2

Managing Supply Under Demand-Side Parameter Uncertainty

First, we solve a simpler problem that involves parameter uncertainty only on the demand side. We would later extend these findings to supply side parameter uncertainty. This problem is more complicated than it sounds. In every day life, the arrival rates are uncertain. However, most of the traditional literature assumes that all model primitives are known with certainty, and "noise" is restricted to stochastic variability.

2.1 Reviewing Existing Models

There are mainly 2 key factors considered to make the choice of capacity level:

1. **Efficiency:** This concerns itself with basic operating costs, e.g, factors like cost of hiring play an important role.

2. **Quality:** This concerns itself with quality of customer service, e.g., factors like abandonment costs play an important role.

Capacity planning is a trade-off between these 2 factors. The best quality of service would require a high amount of spare capacity which could be wasteful and the best efficiency would require preventing occurrence of idle servers this can greatly degrade the quality of service.

If the arrivals follow a Poisson process with rate λ and services are exponentially distributed with rate μ.
Some common regimes that are in use give the capacity C as:

1. **Efficiency-Driven (ED) Regime:**

$$C = \frac{\lambda}{\mu} - \gamma\frac{\lambda}{\mu} \quad where \quad 0 < \gamma < 1$$

   This model is likely to be under-staffed.

2. **Quality-Driven (QD) Regime:**

$$C = \frac{\lambda}{\mu} + \delta\frac{\lambda}{\mu} \quad where \quad \delta > 0$$

   This model is likely to be over-staffed.

3. **Quality- and Efficiency-Driven (QED) Regime:**

$$C = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} \quad where \quad -\infty < \beta < \infty$$

This regime is also known as the Halfin-Whitt regime [4].

In all the cases above, the term $\frac{\lambda}{\mu}$ is a base capacity used to meet mean demands. The second term is the variability hedge. The QED regime is popular because it takes into account both the quality and efficiency factors. It takes the form dictated by the The Square Root Staffing Law that was covered in the Key Concepts chapter.

However, the QED regime is not very good for practical implementation. The square-root-law expects a fairly accurate estimate of the service and arrival rates. However, in a practical setting, such parameters are estimated and predicted on the basis of historical data, and hence can be quite "noisy." That means there is expected to be ambiguity in the parameters, such as, arrival rate itself. Hence, we try to find a better solution to the problem.

2.2 Proposing an Approximate Solution based on the Newsvendor Problem

2.2.1 Formulating the Problem

- **Servers:** b identical servers
- **Customer Arrivals:** Doubly stochastic Poisson process, that is, the

arrival rate Λ is also a random variable. It has distribution F and mean λ. The arrival process is a homogeneous Poisson process with this rate.

- **Service Requirements:** i.i.d. exponential random variables. They are independent of the arrival process and rate and have mean $\frac{1}{\mu}$
- **Queue Discipline:** FCFS (First come first served) is used.
- **Waiting Policy:** Servers do not remain idle unless queue is empty. There is an infinite capacity buffer. The per-customer per-unit-time cost of a customer waiting in queue is called the holding cost and it is h per customer per unit time.
- **Abandonment Policy:** Customers have the impatience random variable τ, exponentially distributed with mean $\frac{1}{\gamma}$. A customer may abandon the system after getting impatient from waiting. The customer will abandon the system when her total waiting time in queue reaches τ time units. There is a cost incurred at a rate p per customer, this is called the abandonment cost.
- **Staffing Cost:** c per unit time per server.
- **Queuing Model:** $M/M/b + M$ with a random arrival rate.
- N is the random variable that represents the number of customers in the system in steady state.
- Length of the planning horizon is normalized to 1.
- Integrality constraints are ignored and b is assumed to be real-valued.

Putting things together,

Expected Steady State Queue length $= \mathbb{E}[N - b]^+$.

Cost related to Quality, or the total expected customer cost in steady state $= (h + p\gamma)\mathbb{E}[N - b]^+$.
Cost related to Efficiency, or the total staffing cost $= cb$.

The Optimization Problem:

$$Minimize \;\; \Pi(b) = (h + p\gamma)\mathbb{E}[N - b]^+ + cb$$

$$subject \;\; to \;\; b \geq 0$$

Let b^* denote the minimizer and $\Pi^* := \Pi(b^*)$ the corresponding minimum cost.

2.2.2 Proposing an Approximate Solution

The optimization problem looks simple. However, it is not possible to get an exact solution to it. This is because the distribution of N itself depends on b.

Approach: We ignore the stochastic variability in customer arrivals and service requirements. We only focus our attention on the uncertainty in the arrival rate. This relaxation makes things easy because now customer arrivals form a fluid queue with the fluid flowing at the rate Λ per unit time. The processing capacity is is equal to μb and is fixed.

$$Rate \;\; of \;\; Abandonment \approx \mathbb{E}[\Lambda - \mu b]^+$$

Also,

$$Rate \;\; of \;\; Abandonment = \gamma \mathbb{E}[N - b]^+$$

So,

$$\mathbb{E}[N - b]^+ \approx \frac{1}{\gamma}\mathbb{E}[\Lambda - \mu b]^+ \qquad [1]$$

This helps us get rid of N which was causing some problems. Now,

The Optimization Problem:

$$Minimize \;\; \bar{\Pi}(b) = (p + \frac{h}{\gamma})\mathbb{E}[\Lambda - \mu b]^+ + cb$$

$$subject \;\; to \;\; b \geq 0$$

Here, $\bar{\Pi}(.)$ is the new objective function that is being proposed.
This is an instance of the newsvendor problem that was visited in the Key Concepts chapter where $C = \frac{c}{\mu} \;\; and \;\; P = p + \frac{h}{\gamma} \;\; and \;\; S = 0$. Also, the Number of Units available is analogous to the Processing Capacity available. So, $Q = \mu b$. If F is the cumulative distribution function of Λ, then, $\bar{F} = 1$ - F. So,

$$\bar{Q} = F^{-1}(\frac{P - C}{P - S})$$

$$\bar{Q} = F^{-1}(\frac{P - C}{P - 0}) = F^{-1}(\frac{P - C}{P})$$

$$\bar{Q} = \bar{F}^{-1}(\frac{C}{P})$$

$$\mu\bar{b} = \bar{F}^{-1}(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}})$$

$$\bar{b} = \frac{1}{\mu}\bar{F}^{-1}(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}})$$

Remark 2.2.1. Note that if the cost of capacity (per unit) is higher than the penalty charge (per unit), the optimal solution would not install any capacity. Our proposed solution would not install any capacity either. More formally, If,

$$\frac{c}{\mu} > p + \frac{h}{\gamma}$$

Then,

$$b^* = \bar{b} = 0$$

2.3 Analysing the Performance of the Proposed Solution

We will be using some numerical data to analyse the performance of the proposed solution relative to the optimal solution. For all the cases we consider $c = 1/3$, $p = 1$, $h = 1$, $\mu = 1$ and $\gamma = 1/3$. These values ensure that the condition of Remark 2.2.1 does not arise. In all the cases, the optimal solution has been picked from the internet and the proposed solution is computed using the formula derived in section 2.2.2. All values have been rounded to 1 decimal place.

After forming our intuitions based on the numerical data we mention some related mathematical results that justify our observations. These relations are derived in Bassamboo and Randhawa (2010) [1] and are stated here without proof.

Before moving further, a few terms need to be defined that will be used in this section.

Definition 2.3.1. The safety capacity (β) is defined as the difference between the optimal solution b^* and the proposed solution $\bar{b}$.

Definition 2.3.2. The accuracy gap ($\Delta(b)$) is defined as the difference between the value of the actual objective function at b, $\Pi(b)$, and the value of the proposed newsvendor objective function at b, $\bar{\Pi}(b)$.

Definition 2.3.3. The optimality gap is defined as the difference between the value of $\Pi(\bar{b})$ and the value of $\Pi(b^*)$.

2.3.1 Deterministic Arrival Rates

We first consider the simple case where there is no uncertainty in the arrival rate, that is, the case where arrival rate takes constant values.

Computing the Proposed Solution

$$\bar{b} = \frac{1}{\mu}\bar{F}^{-1}(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}})$$

$$\bar{b} = \frac{1}{1}\bar{F}^{-1}(\frac{\frac{1/3}{1}}{1 + \frac{1}{1/3}})$$

$$\bar{b} = \bar{F}^{-1}(\frac{1}{12})$$

$$\bar{b} = F^{-1}(\frac{11}{12})$$

Now, for the deterministic case,

$$F^{-1}(p) = \lambda \quad where \quad 0 < p < 1$$

So,

$$\bar{b} = \lambda$$

Comparing the Proposed Solution with the Optimal Solution using Numerical Data

Table 2.1 and Figure 2.1 show the performance of the proposed solution on deterministic arrival rates.

We see that the optimality gap increases as the value of the arrival rate increases. We see that the gap between b and b^* also increases with increasing arrival rate.

Remark 2.3.4. When the arrival rate increases by 4 times in the last 2 rows of Table 2.1, we see that the optimality gap approximately doubles. The value of the safety capacity is positive in all cases and it doubles in the aforementioned case. Looking carefully we see that β grows proportionally to $\sqrt{\lambda}$. This is very much in line with the square root staffing law, as one would expect in case of deterministic arrivals.

Related Mathematical Results

$$b^* \approx \bar{b} + C_0\sqrt{\frac{\lambda}{\mu}} \quad and \quad \Pi(b^*) \approx \Pi(\bar{b}) - C_1\sqrt{\frac{\lambda}{\mu}} \quad [1]$$

Arrival Rate	Optimal Solution		Proposed Solution		Comparison	
λ	b^*	$\Pi(b^*)$	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	Optimality Gap
37.5	42	15.9	37	17.7	5	1.7
75	83	29.6	75	31.6	8	2.0
300	316	109	300	112.4	16	3.4

Table 2.1: Performance of the Proposed Solution on Deterministic Arrival Rates

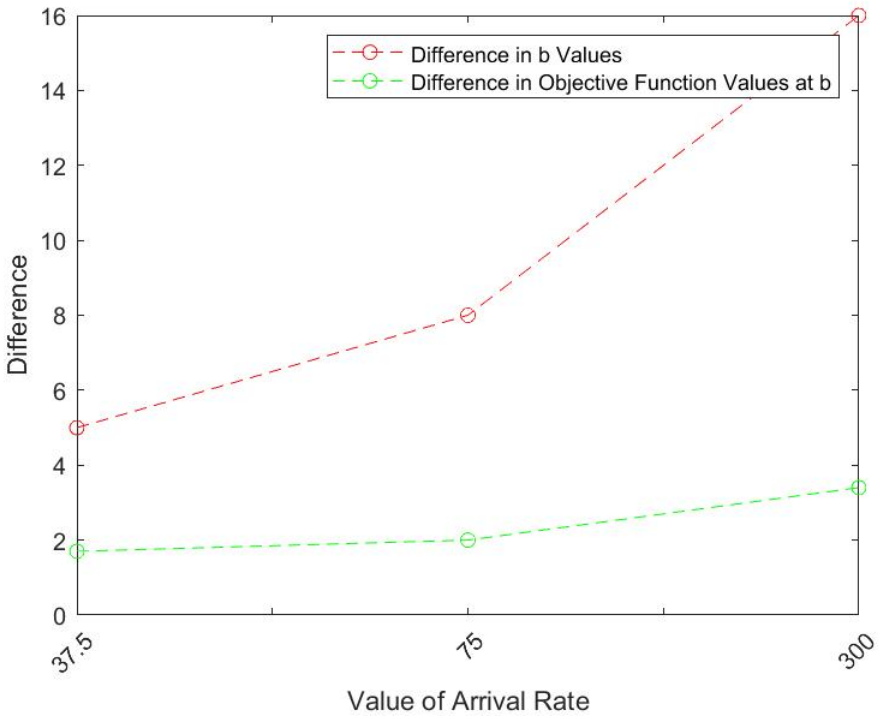

Figure 2.1: Performance of the Proposed Solution on Deterministic Arrival Rates

(a) This image shows the deviation of the proposed solution from the optimal one in the case of different arrival rate values.

(b) The red line shows the values of $|b^* - \bar{b}|$ and the green line shows the value of $\Pi(\bar{b}) - \Pi(b^*)$.

2.3.2 Uncertain Arrival Rates

We assume that the arrival rates follow a uniform distribution, U[a,b].

Computing the Proposed Solution

$$\bar{b} = \frac{1}{\mu}\bar{F}^{-1}(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}})$$

$$\bar{b} = \frac{1}{1}\bar{F}^{-1}(\frac{\frac{1/3}{1}}{1 + \frac{1}{1/3}})$$

$$\bar{b} = \bar{F}^{-1}(\frac{1}{12})$$

$$\bar{b} = F^{-1}(\frac{11}{12})$$

Now, for U[a,b],

$$F^{-1}(p) = a + p(b - a) \quad where \quad 0 < p < 1$$

So,

$$\bar{b} = a + \frac{11}{12}(b - a) = \frac{11}{12}b - \frac{1}{12}a$$

Comparing the Proposed Solution with the Optimal Solution using Numerical Data

Table 2.2 and Figure 2.3 show the performance of the proposed solution on uncertain arrival rates.

We observe that the performance is quite bad when the mean arrival rate is small. It improves as we move towards larger mean arrival rates. The drop in

the value of optimality gap is, infact, quite drastic in the beginning. We see good performance after U[10,20]. The poor performance in case of smaller mean arrival rates can be ignored in a practical scenario because mostly large traffic is seen.

Remark 2.3.5. Observe that arrivals rate values in Table 2.1 are nothing but the mean of the distributions in last 3 rows of Table 2.2. The difference in the value of optimality gap in the two tables, however, is quite high. This shows that the proposed solution works better when the arrival rate has uncertainty.

Arrival Rate	Optimal Solution		Proposed Solution		Comparison	
Distribution	b^*	$\Pi(b^*)$	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	Optimality Gap
U[1,2]	3	1.4	1	5.3	2	3.8
U[2,4]	5	2.2	3	3.3	2	1.1
U[5,10]	11	4.3	8	5.0	3	0.7
U[10,20]	20	7.7	17	8.0	3	0.3
U[15,30]	29	11.0	26	11.1	3	0.2
U[20,40]	37	14.2	35	14.3	2	0.1
U[25,50]	46	17.3	43	17.5	3	0.2
U[50,100]	89	33.1	87	33.2	2	0.1
U[200,400]	351	127.1	350	127.1	1	0.0

Table 2.2: Performance of the Proposed Solution on Uncertain Arrival Rates

Related Mathematical Results

$$\Delta(b) = \Pi(b) - \bar{\Pi}(b) \approx K\mathbb{E}\left[\sqrt{\frac{\Lambda}{\mu}} exp(-\frac{(\frac{\Lambda}{\mu} - b)^2}{2\frac{\Lambda}{\mu}})\right] \quad [1]$$

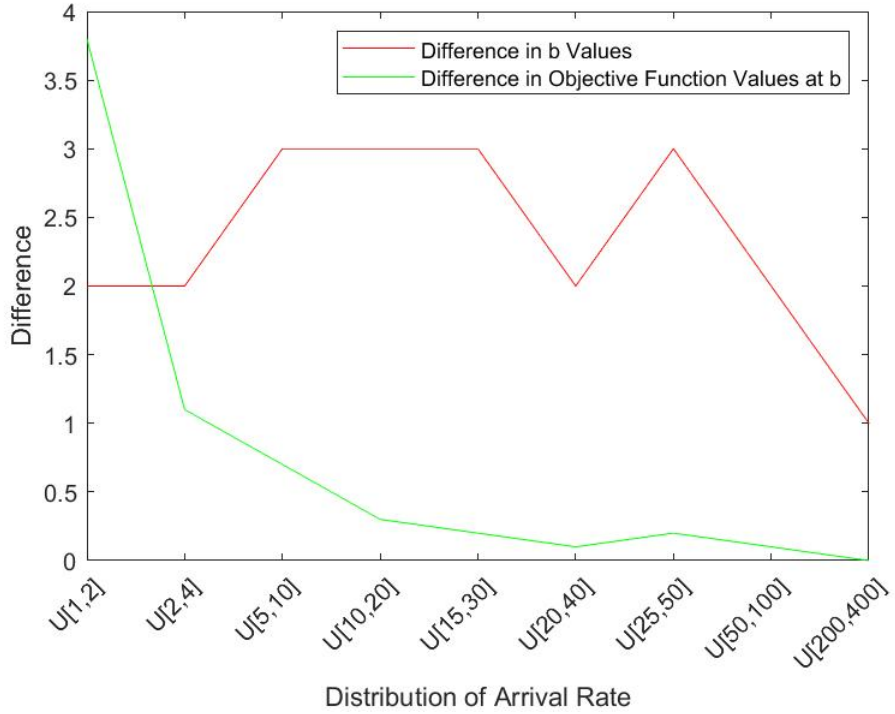

Figure 2.3: Performance of the Proposed Solution on Uncertain Arrival Rates

(a) This image shows the deviation of the proposed solution from the optimal one in the case of various arrival rate distributions.

(b) The red line shows the values of $|b^* - \bar{b}|$ and the green line shows the value of $\Pi(\bar{b}) - \Pi(b^*)$.

Remark 2.3.6. The exponential term dominates in RHS. This term is clearly small when there is some amount of uncertainty in the arrival rates. This means that minimising $\Pi(b)$ essentially means minimising $\bar{\Pi}(b)$. That is why, in case of uncertainty, the newsvendor solution works very well.

2.4 Some Important Results

This section mentions some important results derived in [1]. These are stated here without proof. In this section, the terms have been appended with an additional subscript λ where $\lambda = \mathbb{E}(\Lambda)$ is assumed to be large, as is the case in most practical scenarios. We assume that $\sigma_\lambda < \infty$. Some new terms must be defined before moving forward.

Definition 2.4.1. The coefficient of variation of the arrival rate (CV_λ) is defined as the ratio of σ_λ and λ.

Definition 2.4.2. The offered load (ε_λ) is defined as the ratio of λ and μ.

Theorem 2.4.3. *(Performance of the solution in different regimes)*
(a) (Uncertainty-dominated regime.) If $CV_\lambda \gg \frac{1}{\sqrt{\varepsilon_\lambda}}$, *then,*

$$\Pi_\lambda(\bar{b_\lambda}) = \Pi_\lambda(b^*_\lambda) + \mathcal{O}\left(\frac{1}{CV_\lambda}\right)$$

(b) (Variability-dominated regime.) If $CV_\lambda \ll \frac{1}{\sqrt{\varepsilon_\lambda}}$, *then,*

$$\Pi_\lambda(\bar{b_\lambda}) = \Pi_\lambda(b^*_\lambda) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

(c) If the variability and uncertainty are balanced, then,

$$\Pi_\lambda(\bar{b_\lambda}) = \Pi_\lambda(b^*_\lambda) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

(d) (Deterministic regime). If $CV_\lambda = 0$, *then,*

$$\Pi_\lambda(\bar{b_\lambda}) = \Pi_\lambda(b^*_\lambda) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

Theorem 2.4.4. *(A special case of $\mathcal{O}(1)$ optimality)*
If CV_λ is bounded away from zero, then,

$$\Pi_\lambda(\bar{b_\lambda}) = \Pi_\lambda(b_\lambda^*) + \mathcal{O}(1)$$

Chapter 3

Optimizing the system management cost in the presence of self scheduling servers

(This chapter has been inspired from ***Ibrahim(2018)*** **[6]** , for all further explanations and results refer [6])

We have seen how the parameter side uncertainty can affect a system in the previous chapter, in this chapter, we will take a look at the system where the number of servers is random, because the servers can self-schedule for different shifts.

In a real life scenario, for many working systems, uncertainty in the availability of servers is inevitable .For models like Call Centres, Uber, Lyft, where

the employee can self-schedule their working hours, it is important for the system manager to maintain a optimal number of employees to keep up with the demand and keep the customers happy while also minimising the cost of maintaining the system.

So let us consider a multi period queuing model where the servers can self schedule whether to serve in a particular shift and the customers are impatient as in any real life system.

3.1 Modelling framework :

Let us consider a system with k shifts and the queuing models for a individual shift to be $G/G/N_j^n + GI$ in steady state. Here , N_j^n is a random variable that indicates the number of servers in j^{th} shift which depends on the total number of servers n. We also assume that all the servers are identical and service are identically distributed independent random variables and they follow a general distribution with mean $\frac{1}{\mu}$, and W.L.O.G , $\mu = 1$. Also there is no restriction on any server in successive shifts or multiple shifts.

When the customer is waiting in the queue, after a random time, the customer leaves without being served, we call this time as patience time. Let us assume that the patience times are identically distributed independent random variables with a cdf F, complementary cdf ccdf $\bar{F}$, density function f, hazard-rate function h_a and mean $\frac{1}{\theta}$. We assume the arrival, service, abandonment processes are mutually independent and also independent of the number of servers in a shift N_j^n. Also the queue is $FIFO$ and the capacity is unlimited. The customer arrival rate is stationary process of λ_j for the j^{th}

shift.

3.2 System Manager's Problem

Being consistent with *Bassamboo and Randhawa*[1] , let's consider two costs in the system

- a delay cost h_j incurred per unit time a customer waits in the queue for being served in the j^{th} shift
- a penalty p_j incurred per every customer who abandons the system in the j^{th} shift

Let $Q_{N_j^n}$ be the queue length and $\alpha_{N_j^n}$ be the net abandonment. The system manager can decide on the staffing level n , the problem becomes

$$\min_{n\in\mathbb{N}} \Pi(n) = \sum_{1\leq j\leq k} (c_j\mathbb{E}[N_j^n] + p_j\mathbb{E}[\alpha_{N_j^n}] + h_j\mathbb{E}[Q_{N_j^n}]) \tag{3.1}$$

where c_j is the compensation for each server in j^{th} shift.
Since an exact analysis of (3.1) is difficult, let us consider a fluid approximation of the above problem, we will later show that the fluid approximation works good in case of a binomial distribution for the number of servers.
"For an $G/G/s + GI$ system, $\bar{q}_{\rho_s}$ and $\bar{\alpha}_{\rho_s}$ are the fluid approximations of queue length and net abandonment rates respectively, with traffic intensity $\rho_s \equiv \frac{\lambda}{\mu s}$. This result is obtained from *Whitt(2006)* [9]"

So the fluid approximation for (3.1) is

$$\min_{n\in\mathbb{N}} \Pi(n) = \sum_{1\leq j\leq k} nG(c_j)c_j + p_j.\bar{\alpha}_{\rho_s/G(c_j)} + h_j.\bar{q}_{\rho_s/G(c_j)}[6] \tag{3.2}$$

Here, $\mathbb{E}[N_j^n] = G(c_j)$, since $G(c_j)$ is a constant , let it be r_j , which we define as availability of a server in j^{th} shift. In the fluid approximation, we can see that the optimal staffing level is only dependent on the expectation of the number of servers in a particular shift , but in reality , the variance is a important factor as well which we will discuss later in the paper.

3.3 Benchmark Case : No Self Scheduling

Before we proceed to the problem, let us take a benchmark case where there is no self scheduling which means r_j=1 , for all the shifts. Clearly , in a system where self scheduling is absent the manager can choose to staff each shift optimally individually , where as in a system where the servers can self schedule they have to choose the pool of servers before hand and allow them to reschedule.

The queue length in j^{th} shift can be written as

$$q_j = \int_0^{w_j} D_j(u)\, du \tag{3.3}$$

where $D_j(u)$ is the density of fluid that has been waiting for u units of time in j^{th} shift and it can be written as

$$D_j(u) = \lambda_j \bar{F}(u) \tag{3.4}$$

Combining (3.3) and (3.4) , we have

$$q_j = \int_0^{w_j} \lambda_j \bar{F}(u)\, du \tag{3.5}$$

The net abandonment rate in j^{th} shift can be written as

$$\alpha_j = \lambda_j F(w_j) \tag{3.6}$$

where w_j denotes the waiting time in given shift j. Clearly, when there is no self-scheduling the individual optimal staffing in each shift

$$n_j^* = \lambda \bar{F}(w_j^*) \tag{3.7}$$

where w_j^* would be the optimal waiting time. Also clearly the individual staffing level for each shift is less than the λ_j for that shift since $\bar{F}(w_j) \leq 1$, so staffing more than λ_j in any shift would not be optimal.
With the results in (3.5) and (3.6) the system managers problem becomes

$$min_{w_j \geq 0} \left(\lambda_j c_j \bar{F}(w_j) - \lambda_j p_j \bar{F}(w_j) + \lambda_j h_j \int_0^{w_j} \lambda_j \bar{F}(u)\, du \right) \tag{3.8}$$

Lets look at some assumptions and propositions

Assumption 3.1. *For all* j , $c_j < \min(h_j/h_a(0) + p_j, h_j/\theta + p_j)$

The above assumption tells us that this source of servers are economical enough for the manager to staff from them and not avoid them altogether, this is consistent with the assumption in [1] , also if the abandonment process is exponential , we have $h_a(t) = \theta t$, so $h_a(0) = 0$, the first term becomes

inf , so the assumption becomes

$$c_j < h_j/\theta + p_j \tag{3.9}$$

Using this , we will try and show the next proposition (4.1.)

Proposition 4.1. *Under the above assumption , to optimise the costs in a system with no self scheduling , we should load all the shifts critically , i.e.* $n_j^* = \lambda_j$ for all j

Since we know that the staffing costs are reasonably inexpensive , it is in the interest of the system manager to match the demand as there will be no reneging in the system and it is optimal, where if there was no upper bound on the staffing cost , it would have been advisable to leave some customers unattended to optimize the overall problem.

3.4 The actual problem : Servers with Self-Scheduling capability

In this case , the r_j factor is no longer 1, so we define a new arrival rate called the *augmented arrival rate* Γ_j.

Definition 3.4.1. The augmented arrival rate is defined as the ratio of actual arrival rate and the showup probability $i.e\ \Gamma_j = \lambda_j/r_j$

Also W.L.O.G , we will assume that the shifts are ordered in the increasing order of Γ_j $i.e\ \Gamma_{j-1} \leq \Gamma_j$ for j=$1, 2, ..., k$, with this ordering , we can have the following:

Remark 3.4.2. If we take a staffing level n , such that $\Gamma_{j-1} < n < \Gamma_j$, then we can see that the shifts with $i \geq j$ are overloaded and the shifts with $i < j$ are under loaded.

Remark 3.4.3. If we take a staffing level n , such that $n = \Gamma_j$, then we can see that the shifts with $i > j$ are overloaded and the shifts with $i < j$ are under loaded , while the shift j is critically loaded.

Remark 3.4.4. In an overloaded shift, we have $\Gamma_i \bar{F}(w_i) = n$, *i.e* , $w_i = \bar{F}^{-1}(n/\Gamma_i)$.

As we observed in the benchmark case ,underloading all the shifts can result in a congested system , while on the other hand overloading might increase the cost of maintaining the system, so the system manager's problem in terms of the augmented arrival rate Γ_j becomes

$$min_{0 \leq n \leq \Gamma_k} C(n) = \left(\sum_{j=1}^{k} I(\Gamma_{j-1} \leq n < \Gamma_j) u_j(n) \right) \tag{3.10}$$

where $I(n \in S)$ is a indicator random variable that tells us whether the value n is in the set S and $u_j(n)$ is given by

$$u_j(n) = \sum_{i=1}^{k} c_i n r_i + \sum_{i=j}^{k} \left(p_i(\lambda_i - n r_i) + h_i \lambda_i \int_0^{\bar{F}^{-1}(n/\Gamma_i)} \bar{F}(u)\, du \right) \tag{3.11}$$

this u_j represents the cost of the system if n is chosen in the $[\Gamma_{j-1},\Gamma_j)$, also we can see that only one of the indicator variables is non-zero and the rest of them are zero.So if we choose it in the $[\Gamma_{j_0-1},\Gamma_{j_0})$ we have the cost of system is u_{j_0} and all the shifts below j_0 the system is under staffed and the manager faces customer related losses, where as for the shifts j_0 and above it is over

staffed and the staffing costs are high.

We will now derive a proposition which compares this system to the benchmark and tells us in which condition, maintaining a self scheduling system is not more expensive for the manager when compared to a system with no self scheduling

Proposition 3.2. *If the calculated Γ_j ("resulting augmented arrival rates") are equal across all shifts then and only then, the self scheduling system is not more expensive than a system with no self scheduling*

Proof. Let all the augmented arrival rates be equal $i.e\ \Gamma_j = \Gamma$, then assuming that $n^* = \Gamma$ gives, $n_i^* = n^* r_i = \lambda_i$ for all shifts, as we have seen in proposition (4.1), this is the optimal staffing cost in the case of the benchmark, so if all the augmented arrival rates are equal , we can eliminate the cost of self scheduling. □

The above proposition shows us the importance of having multiple shifts in the system. Because in the case of a single shift, the manager could just meet the demand in that shift by having a big enough staffing level like λ/r and have no cost of self scheduling, it is to accommodate multiple shifts, manager has to take in to consideration the cost of self scheduling.

In light of this proposition , let us consider a system with 2 shifts: shifts A and B, let us assume that the arrival rates in both the system is known to the manager and they are λ_A and λ_B and without loss of generality let us assume that $\lambda_A < \lambda_B$, then while selecting the staffing pool for this shifts, the manager could select a pool such that the probability of showing up is higher in shift A than B $i.e\ r_A < r_B$ and the ratios are equal and by doing the cost of self scheduling can be avoided. But as we know in a real life

scenario it may not be possible for the system manager to know these values beforehand. They can use the estimated historic values to make a decision.

Now let us dive into a case where the augmented arrival rates are not equal and the hazard rate for the abandonment process is increasing, specifically exponential.

3.5 Exponential Abandonment and Non-equal Augmented Arrival Rates

In this case, we actually see that it is efficient for the system to match the demand in one of the shifts with all of the shifts instead of matching it in every shift and leaving the remaining shifts underloaded or overloaded.

Proposition 3.3. *For exponential abandonment rates :*

1. *The optimization function in (3.10) is piece-wise linear.*
2. *There exists one shift s, such that by matching the demand in that shift we can optimize the system i.e. $n^* = \Gamma_s$.*
3. *i_0 satisfies the following condition:*

$$\sum_{j=1}^{k} c_j r_j - \sum_{j=i_0}^{k} (p_j + h_j/\theta)\, r_j < 0 \; and \sum_{j=1}^{k} c_j r_j - \sum_{j=i_0+1}^{k} (p_j + h_j/\theta)\, r_j > 0 \tag{3.12}$$

Proof. If the cumulative distribution function $F(x) = 1 - e^{-\frac{x}{\theta}}$ i.e the abandonment process is exponential, then for $\Gamma_{i-1} \leq n \leq \Gamma_i$

$$u_i(n) = \sum_{j=1}^{k} c_j r_j n + \sum_{j=i}^{k} (p_j + h_j/\theta)(\lambda_j - n r_j) \tag{3.13}$$

Clearly, under condition (3.12), C(n) is piece wise linear with

$$\frac{dC(n)}{dn} = \begin{cases} negative & n \leq \Gamma_{i_0} \\ positive & n > \Gamma_{i_0} \end{cases}$$

Hence,the minimum must occur at some Γ_s. □

Chapter 4

Capacity sizing with only self scheduling serves under Demand-Side Parameter Uncertainty

In this chapter we will look at a workforce consisting of only flexible i.e. self-scheduling servers and look at the optimal cost minimising model in this case while also considering the demand-side parameter uncertainty that we looked at chapter 2, hence combing the concepts of both chapter 2 and chapter 3 .It is essential that we look at a workforce with only flexible workers because two main reasons, one it is prevalent in a real life scenario and also it gives us clean insights into how a system would behave with service side uncertainty.

4.1 Modelling Framework

We will continue using the same modelling framework we will consider a single class $M/M/N+M$ model, but here N, the number of servers is random. The service times are i.i.d exponential with rate μ. As in any real life scenario the customers are impatient and their patience times are i.i.d with rate θ. The queue discipline is FIFO. We assume all the processes are mutually independent of each other and also independent of N. Also assume that the shift is long enough for abandonment to keep the system stable.We take that there are k shifts in the system and T_i is the length of period i.For the period i , the arrival rate of Poisson arrival process is λ_i. Let's assume a λ where $\lambda > 0$ and $\lambda_i = \lambda\xi_i$ where $\xi_i > 0$ for all i. We also arrange the shifts in the increasing order of λ_i i.e. $\lambda_i > \lambda_j$ for $i > j$.

4.1.1 A random number of servers

Let n_λ^i be the number of flexible servers decided beforehand for the shift i with arrival rate λ_i. Now let $N_{flex}(n_\lambda^i)$ be the number of servers that showup in the shift i, which we could clearly see is a function of n_λ^i.So,

$$N_{flex}(n_\lambda^i) = \eta_{n_\lambda^i} + \epsilon_{n_\lambda^i}$$

Here, $\eta_{n_\lambda^i} = E[n_\lambda^i]$ and $\epsilon_{n_\lambda^i}$ represents the randomness factor where $E[\epsilon_{n_\lambda^i}] = 0$ and $Var[\epsilon_{n_\lambda^i}] = \sigma^2_{n_\lambda^i}$. Our main task here is to find the optimal value of n_λ^i, for this we need to know the relation between the $N_flex(n_\lambda^i)$ and n_λ^i, this distribution mainly depends on the value of $\sigma^2_{n_\lambda^i}$. We need to find the variance $\sigma^2_{n_\lambda^i}$ as a function of n_λ^i.

For this, we will take the starting point as a Binomial models where servers are treated as independent of each other and with constant probability like in *Ibrahim*[6]. Considering this ,for a single period we get

$$N_{flex}(n_\lambda) = \sum_{j=1}^{n_\lambda} I_j$$

where , for $1 \leq j \leq n_\lambda$, $I_j = 1$ if the servers shows up else 0 and have a constant probability p of showing up. Since, the I_j are i.i.d, we have $\eta_{n_\lambda} = n_\lambda.p$ and $\sigma^2_{n_\lambda} = n_\lambda.p.(1-p)$. We can see here that we are getting that the variability is of order $\sqrt{n_\lambda}$.

In a real life scenario, the servers decisions arent independent as there might be some common factors that effect their showup probability. Like in case of uber drivers , weather conditions can be a common factor. We see in most cases, the variability observed is far greater order than $\sqrt{n_\lambda}$.This tells us that we need to go beyond the binomial approximation and consider other distributions for server show-up probability.See *Dong* [2] for numerical results of Uber driver data and extensions of binomial approximation.

4.2 This papers model

After considering the various models for server showup, we can see that the $\eta_{n^i_\lambda} = n^i_\lambda.p$ so by slight abuse of notation, from here on we consider n^i_λ as the expected number of servers that show-up in a shift. From this, we get the simplified model,

$$N_{flex}(n^i_\lambda) = n^i_\lambda + \sigma_{n^i_\lambda}\epsilon_i$$

where ϵ_i are i.i.d random variables $-1 \leq \epsilon_i \leq 1$ with $E[\epsilon_i] = 0$. We also assume that in this simplified model ϵ_i distribution is independent of $n_{\lambda i}$ and has a *p.d.f* of f_ϵ on $(-1, 1)$ and a *c.d.f* of F_ϵ which is invertible in the given domain.For the ease of explanation, we also assume the specific form $\sigma_n = an^q$, for some $a > 0$ and $0 < q \leq 1$. For $q = 1$, we take $a < 1$ so that $N_{flex}(n_\lambda) \geq 0$. Our main task now is to find the optimal n^i_λ.

4.3 The long term staffing problem

As we discussed earlier the decision for number of server is done before each shift by the system manager. So in practice, at time zero the manager decides a number of flexible servers n^i for shift i. At the start of shift , we get to know the number of servers who showed up $N_{flex}(n^i) = s^i$ for the remainder of the shift, the system operates like a Erlang A queue with s^i servers.
Consistent to the *bassamboo*[1] , we consider two costs in the system related to the customer , the delay cost h and the abandonment penalty r . If we denote the cost per flexible server as c_{flex}, we have from **Assumption 3.1**

$$c_{flex} < (\frac{h}{\theta} + r)\mu$$

This assumption tells that it's not infinitely expensive to employ flexible workers.
Let $Q^i_\lambda(n^i_\lambda)$ be the steady state queue length in shift i and $X^i_\lambda(n^i_\lambda)$ be the steady state number of customers in shift i, we have

$$Q^i_\lambda(n^i_\lambda) = (X^i_\lambda(n^i_\lambda) - N_{flex}(n^i_\lambda))^+$$

If $\xi_\lambda^i(n_\lambda^i)$ is the steady state rate of customer abandonment, with exponentially distributed patience times with rate θ, we have

$$\xi_\lambda^i(n_\lambda^i) = \mathbf{E}[Q_\lambda^i(n_\lambda^i)]$$

If $\mathbf{n}_\lambda = (n_\lambda^1, n_\lambda^2, \ldots, n_\lambda^k)$, the problem can now be written as

$$\min_{\mathbf{n}_\lambda} \Pi_\lambda(\mathbf{n}_\lambda) \equiv \sum_{i=1}^{k} T_i(c_{flex} n_\lambda^i + h.\mathbf{E}[Q_\lambda^i(n_\lambda^i)] + r.\xi_\lambda^i(n_\lambda^i)) = \sum_{i=1}^{k} T_i(c_{flex} n_\lambda^i + (h + r\theta)\mathbf{E}[Q_\lambda^i(n_\lambda^i)]) \tag{4.1}$$

We can further simplify this into a single shift problem and ignore the i and write it as

$$\min_{n_\lambda \geq 0} \Pi_\lambda^i(n_\lambda) \equiv c_{flex} n_\lambda + (h + r\theta)\mathbf{E}[Q^i(n_\lambda)] \tag{4.2}$$

From here on we denote the optimal solution using n_λ^*

4.4 Fluid Approximation

In this, we ignore both stochastic variation and parameter uncertainty.

4.4.1 The problem formulation

$$\min_{n} \bar{\Pi_\lambda}(n) \equiv c_{flex}.n + (\frac{h}{\theta} + r)\mu(\frac{\lambda}{\mu} - n)^+ \tag{4.3}$$

Let, $\beta = (\frac{h}{\theta} + r)$ and We denote $\bar{n_\lambda}$ as the solution to (4.3).

4.4.2 Asymptotic Accuracy

Theorem 4.4.1. *For large λ,*

$$\Pi_\lambda(\bar{n_\lambda}) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(max(\sigma_\lambda, \sqrt{\lambda}))$$

From this theorem, we can see that when the uncertainty in the number of servers in the system i.e σ_λ is small , the order of optimality gap for the fluid solution is less i.e. $\mathcal{O}(\sqrt{\lambda})$, where as when the uncertainty is high the gap is large and not optimal.

4.4.3 Optimal staffing policy

Since from **Assumption 3.1** W.K.T, $c_{flex} < \beta$, this problem is same as the problem in chapter 3 , we have the optimal solution $\bar{n_\lambda} = \frac{\lambda}{\mu}$, i.e. match the mean supply with mean demand.

4.5 Stochastic-Fluid Approximation

In this, we only ignore the stochastic variation.

4.5.1 The problem formulation

$$\min_n \tilde{\Pi_\lambda}(n) \equiv c_{flex}.n + \beta E[(\frac{\lambda}{\mu} - n_\lambda - \sigma_{n_\lambda}\epsilon)^+] \tag{4.4}$$

We denote the solution for (4.4) with $\tilde{n_\lambda}$.

4.5.2 Asymptotic Accuracy

Theorem 4.5.1. *For large λ,*

$$\Pi_\lambda(\tilde{n_\lambda}) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(min(\lambda/\sigma_\lambda, \sqrt{\lambda}))$$

As we can see from the theorem when the uncertainty in number of servers is large, the optimalty gap is very small, but when it is small, the gap is of the order that is same as the fluid approximation $\mathcal{O}(\sqrt{\lambda})$, so in case of small σ_λ, there is no advantage in using the stochastic approximation and we can use the fluid approximation.

4.5.3 Optimal Staffing Policy

As we saw , the stochastic fluid approximation gives very small gaps when the uncertainty is large but the solution is only numerically solvable. Here, we will also try to change the $\tilde{n_\lambda}$ to some other closed form equation with the same complexity. We will see that the structure of the solution $\tilde{n_\lambda}$ is closely dependent on the dependency of the uncertainty on the number of servers, i.e on the value of q.

To solve the equation (4.4), we need to essentially solve

$$\Pi'_\lambda(n_\lambda) \equiv c_{flex} - \beta F_\epsilon(\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}) - \beta\sigma'_{n_\lambda} \int_{-1}^{\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}} x f_\epsilon(x)dx = 0 \qquad (4.5)$$

Theorem 4.5.2. *We can divide the solution into mainly 4 regimes and the optimal staffing is as follows.*

(I)[Variability-dominated.] *If $0 \leq q \leq 1/2$, we have the same solution*

as fluid i.e. $n_\lambda = \frac{\lambda}{\mu}$ *and*

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\sqrt{\lambda})$$

(II)[Moderately uncertainty-dominated.] *If* $1/2 < q \leq 3/4$*, we have* $n_\lambda = \frac{\lambda}{\mu} - \gamma\sigma_{\frac{\lambda}{\mu}}$*, where* $\gamma = F_\epsilon^{-1}(c_{flex}/\beta)$ *and*

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda)$$

(III)[Strongly uncertainty-dominated.] *If* $3/4 < q < 1$*, we have* $n_\lambda = \tilde{n_\lambda}$ *and*

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda)$$

(IV)[Extremely uncertainty-dominated.] *If* $q = 1$ *and* $0 < a < 1$ *, then* $n_\lambda = \tilde{n_\lambda} = \frac{\lambda}{\mu}\eta$ *, where* η *is the solution of*

$$c_{flex} + \beta a \int_{-1}^{1/(a\eta)-1/a} F_\epsilon(u)du - \frac{\beta}{\eta}F_\epsilon(\frac{1}{a\eta} - \frac{1}{a}) = 0$$

In this case,

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(1)$$

Here, we should note that in I and II , the values are not optimal solution to (4.5) but rather solutions to a closed form solution which gives the same

complexity of optimalty gap. In the case of III and IV, those are the exact solution of (4.5).

For detailed explanation, refer *Dong and Ibrahim* [2].

Chapter 5

Capacity Sizing with a Blended Workforce

A blended workforce would be a mix of fixed as well as flexible servers. We assume that the number of fixed servers remains constant across the time horizon. The number of flexible servers varies for different periods. We let m_λ denote the number of fixed servers and $\mathbf{n}_\lambda = (n_\lambda^1, .., n_\lambda^k)$ denote the number of flexible servers. The total number of servers in period i is, thus,

$$N(m_\lambda, n_\lambda^i) = m_\lambda + N_{flex}(n_\lambda^i) = m_\lambda + n_\lambda^i + \sigma_{n_\lambda^i}\epsilon_i$$

The manager must decide on the number of fixed and flexible servers during the initial planning stage. Remember that the per unit time staffing cost of a fixed server is c_{fix}, and the per unit time staffing cost of a flexible server is c_{flex}.

Remark 5.0.1. Analogous to our assumption about flexible workers, we have

one for fixed workers to avoid no staffing from the resource. That is,

$$c_{fix} < (\frac{h}{\theta} + r)\mu$$

5.1 The Staffing Problem for Blended Workforce

We are now prepared to formulate the staffing problem for a blended workforce, which is similar to the one for only flexible employees:

$$\begin{aligned}\min_{m_\lambda, \mathbf{n}_\lambda} \Pi_\lambda(m_\lambda, \mathbf{n}_\lambda) &\equiv \sum_{i=1}^{k} T_i(c_{fix} m_\lambda + c_{flex} n_\lambda^i + h\mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] + r\xi(m_\lambda, n_\lambda^i)) \\ &= \sum_{i=1}^{k} T_i(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta)\mathbb{E}[Q^i(m_\lambda, n_\lambda^i)])\end{aligned} \tag{5.1}$$

Due to the fixed servers, we can no longer decompose the problem into k single-period problems.

As we did earlier, we consider the fluid approximation and the stochastic-fluid approximation to make the problem simpler.

5.2 The Fluid Approximation

5.2.1 Problem Formulation

$$\min_{m_\lambda, \mathbf{n}_\lambda} \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv \sum_{i=1}^{k} T_i(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta(\lambda_i/\mu - m_\lambda - n_\lambda^i)^+)$$

5.2.2 Asymptotic Accuracy

Let $n_\lambda^{(m)} \equiv \min \mathbf{n}_\lambda$ and $n_\lambda^{(M)} \equiv \max \mathbf{n}_\lambda$. Let m_λ^* and n_λ^* denote the optimal solution of (5.1) and let $\bar{m}_\lambda$ and $\bar{n}_\lambda$ be the optimal solution of the above equation.

Theorem 5.2.1. *For large λ,*

$$\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}(max\{\sigma_{\bar{n}_\lambda^{(M)}}, \sqrt{\lambda}\}) \qquad [2]$$

That is, if, $\bar{n}_\lambda^{(M)} = \Theta(\lambda)$, then, $\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}(max\{\sigma_\lambda, \sqrt{\lambda}\})$

5.2.3 Optimal Staffing Policy

The optimal staffing strategy captures the tradeoff between the costs of staffing and the flexibility of scaling the number of flexible servers to accommodate seasonality in demand since the fluid approximation ignores parameter uncertainty. We note, on the other hand, that if we plan to staff enough fixed servers to satisfy demand in a given period, h, we must also staff these servers for all other periods. Now, let c_{fix}^h denote the time "modified" cost and be defined as below

$$c_{fix}^h = c_{fix} \cdot \frac{\sum_{i=1}^{k} T_i}{\sum_{i=h}^{k} T_i} \text{ for h} \geq 1$$

Lemma 5.2.2. *The solution to the fluid approximation problem has been given in [2] as,*

$$\begin{cases} \bar{m}_\lambda = 0, & \text{for } k_0 = 0, \\ \bar{m}_\lambda = \frac{\lambda_{k_0}}{\mu}, & \text{for } k_0 > 0, \\ \bar{n}_\lambda^i = 0, & \text{for } 1 \leq i \leq k_0, \\ \bar{n}_\lambda^i = \frac{\lambda_i - \lambda_{k_0}}{\mu}, & \text{for } k_0 < i \leq k \end{cases}$$

where k_0 is given by:

$$k_0 = \begin{cases} 0, & \text{if } c_{flex} < c_{fix}, \\ max\{1 \leq h \leq k : c_{flex} \geq c_{fix}^h\}, & \text{otherwise} \end{cases}$$

5.3 The Stochastic-Fluid Approximation

5.3.1 Problem Formulation

$$\min_{m_\lambda, \mathbf{n}_\lambda} \tilde{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv \sum_{i=1}^{k} T_i(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta \mathbb{E}[(\lambda_i/\mu - N(m_\lambda, n_\lambda^i))^+])$$

5.3.2 Asymptotic Accuracy

Let $\tilde{m}_\lambda$ and $\tilde{n}_\lambda$ be the optimal solution of the above equation.

Theorem 5.3.1. *For large λ,*

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}(\sqrt{\lambda}) \qquad [2]$$

That is, if, $\tilde{n}_\lambda^{(m)} = \Theta(\lambda)$, then, $\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}(min\{\lambda/\sigma_\lambda, \sqrt{\lambda}\})$

5.3.3 Optimal Staffing Policy

The optimal solution to the fluid staffing problem may not be reliable when uncertainty in the number of available servers is large. As a consequence, a stochastic-fluid refinement must be considered. The stochastic-fluid approximation, unlike the fluid approximation, takes parameter uncertainty into account. As a result, we can capture the tradeoffs between three variables when solving the stochastic-fluid optimal staffing policy: the cost of staffing, the flexibility in scaling the workforce to meet seasonality in demand, and supply-side uncertainty.

We define:

$$g(c) \equiv c + caF_{\epsilon}^{-1}(c/\beta) - \beta a \int_{-1}^{F_{\epsilon}^{-1}(c/\beta)} F_{\epsilon}(u)du$$

Theorem 5.3.2. *As specified in [2], in the solution of the stochastic-fluid problem, there are 2 cases:*

- ***Case 1. Variability-dominated, moderately and strongly uncertainty-dominated regimes.*** *If $\sigma_n = an^q$ for $0 < q < 1$ and $a > 0$. Also, $\bar{m}_\lambda$ and k_0 are as given in Lemma 5.2.2, then, the solution is:*

  - $\tilde{m}_\lambda = \bar{m}_\lambda$
  - $\tilde{n}_\lambda^i = 0$ *if* $i \leq k_0$
  - *If* $i > k_0$, $\tilde{n}_\lambda^i$ *is the minimizer of the following problem:*

$$\min_{n_\lambda \geq 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon)^+],$$

  as given by $Theorem(4.5.2)$

- ***Case 2. Extremely uncertainty-dominated regimes.*** *If* $\sigma_n = an$ *for* $0 < a < 1$. *Also,* $g(.)$ *and* c^h_{fix} *are as defined earlier. A new variable is now defined as below:*

$$k_1 = \begin{cases} 0, & \text{if } c_{flex} < g(c_{fix}), \\ max\{1 \leq h \leq k : c_{flex} \geq g(c^h_{fix})\}, & \text{otherwise} \end{cases}$$

The solution is then given by:

  - $\tilde{m}_\lambda = \lambda_{k_1}/\mu$
  - $\tilde{n}^i_\lambda = 0$ *if* $i \leq k_1$
  - *If* $i > k_1$, $\tilde{n}^i_\lambda$ *is the minimizer of the following problem:*

$$\min_{n_\lambda \geq 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - a n_\lambda \epsilon)^+],$$

as given by $Theorem(4.5.2)$

Chapter 6

Capacity Sizing with a Blended Workforce and Freelancers

A freelancer is different from fixed as well as flexible workers. A freelancer is hired on a contract basis to serve one customer only. The contract is assumed to be binding and the freelancer immediately services the customer. Thus, the show-up probability of a freelancer is one.

6.1 The Need and Strategy for Hiring Freelancers

The question that arises now is if there is a need for hiring freelancers on top of a blended workforce. As stated in [2], the fixed resource is used to match all or part of the demand and the flexible workers are used to match the remaining demand and to hedge against variability in capacity. However,

the flexible workers might or might not show up when the shift actually starts. This variability is partially handled by employing an optimal number of flexible workers. The decision to hire flexible workers (the value of n_λ^i for a shift) is taken before the shift starts by considering various factors, including their show-up probabilities. However, there is a possibility that too many flexible workers do not actually show up once the shift starts, for example, the day of a festival. Here, we propose an additional hedging strategy for a large no-show of flexible workers. Once the shift starts, the actual number of flexible workers that show up (the value of $N_{flex}(n_\lambda^i)$) becomes known. At this point, the manager knows the expected arrival in this shift as well as the number of available servers. He can now decide if there is a need to hire freelancers to make up for the flexible workers that didn't show up. The manager takes the decision about freelancers immediately after the start of a shift to minimize the waiting cost incurred. The trade-off would be between the cost of losing the customers and the cost of hiring the freelancers.

Remark 6.1.1. Before mathematically formulating the capacity sizing problem for freelancers, some assumptions must be stated. These assumptions are in line with real-life scenarios but are stated here to avoid confusion at later stages.

- The customer's abandonment time would be significantly smaller than the shift length. That is, no customers roll over to the next shift.
- The hiring cost of freelancers (c_{free}) is greater than the hiring cost for flexible or fixed workers.

6.2 Optimal Staffing Strategy for Freelancers

Let the number of freelancers hired for a particular shift be o_λ^i and the cost of hiring a freelancer is c_{free}. The number of fixed workers is denoted by m_λ, the number of flexible workers that show up for the shift are denoted by $N_{flex}(n_\lambda^i)$ and the length of the shift is denoted by T_i. The arrival process is denoted by Λ and the arrival rate of the Poisson arrival process in period i is given by λ_i. Then,

Penalty cost per customer = Waiting time till abandonment + Abandonment cost

$$\text{Expected penalty cost per customer} = \mathbb{E}(h\tau + r)$$

$$\text{Expected penalty cost per customer} = \frac{h}{\theta} + r$$

$$\text{Number of customers served by a server in shift of period } T_i = T_i\mu$$

$$\text{Expected number of customers left unserviced} = \mathbb{E}((T_i\Lambda - (N_{flex}(n_\lambda^i) + m_\lambda)(T_i\mu))^+)$$

$$\text{where } (x)^+ = max(x, 0)$$

$$\text{Expected number of customers left unserviced} = (T_i\lambda_i - (N_{flex}(n_\lambda^i) + m_\lambda)(T_i\mu))^+$$

Hiring should happen so that cost of hiring does not exceed the cost of losing customers. Thus,

$$o_\lambda^i.c_{free} \leq (\frac{h}{\theta} + r).(T_i\lambda_i - (N_{flex}(n_\lambda^i) + m_\lambda)(T_i\mu))^+$$

$$o_\lambda^i = \left\lfloor \frac{(\frac{h}{\theta} + r).(T_i\lambda_i - (N_{flex}(n_\lambda^i) + m_\lambda)(T_i\mu))^+}{c_{free}} \right\rfloor$$

Remark 6.2.1. If enough servers are available to meet the expected arrivals once the shift starts, our solution will not install any capacity, as is expected. Mathematically,

$$T_i\lambda_i \leq (N_{flex}(n_\lambda^i)+m_\lambda)T_i\mu \implies (T_i\lambda_i-(N_{flex}(n_\lambda^i)+m_\lambda)T_i\mu)^+ = 0 \implies o_\lambda^i = 0$$

6.3 Numerical Results

For our numerical analysis, we are considering a 2-shift model with one shift being a low-demand period with $\lambda_1 = 25$ and other high demand period with $\lambda_2 = 50$. We take that shifts are of length $T_1 = 2$ and $T_2 = 1$, where T_i is the time period of shift i. We assume that the distribution for ϵ is uniform on $(-1, 1)$ and that of server showup is binomial with $p = 0.8$ and the value of $a = 1$. We take that $h = r = 1$ and $\mu = 1, \theta = 0.5$.The costs for employing servers are $c_{fix} = 2/9, c_{flex} = 1/3, c_{free} = 0.35$. We are taking q from 0.2 to 1 in increments of 0.01.

To observe the workings of our blended workforce model, we need to look at two perspectives , the firm perspective and the customer perspective

6.3.1 Firm Persepective : Cost Reduction

From the Figure 6.1, we can see that the cost is lowest for the blended force, which grows closer to fixed cost as q increases because we are employing more number of flexible workers to hedge the uncertainty.We can also see that the

freelancers are adding an extra cost which is expected as we are employing freelancers for Quality of service but not cost reduction.

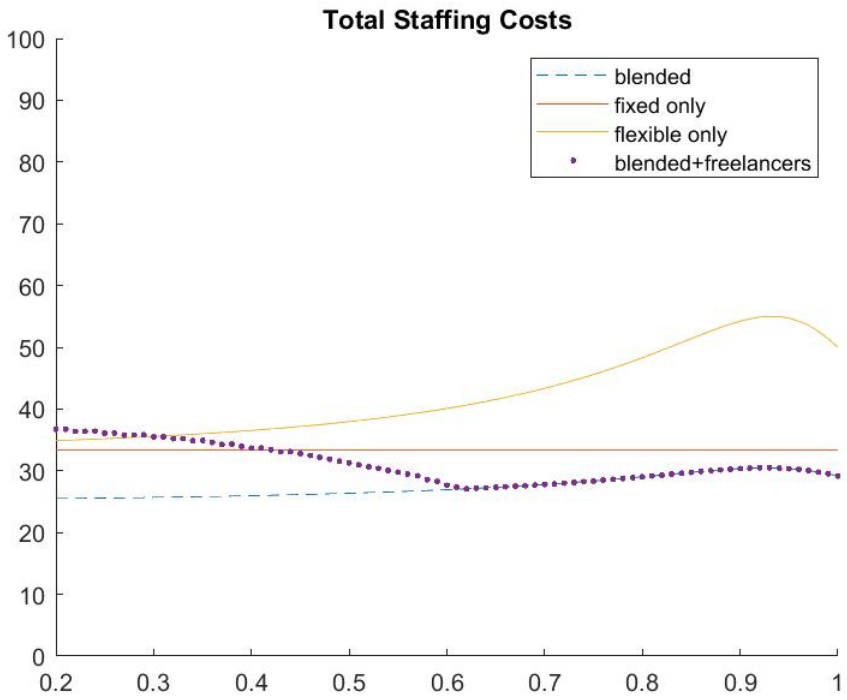

Figure 6.1: Total Staffing Costs

(a) This image shows the total staffing costs, that is for both the periods (high and low-demand) for different types of workforces.

6.3.2 Customer Perspective : Quality of Service

To observe the quality of service, we are looking at the expected number of customers being abandoned.

We can see that in the low demand period that blended and blended+freelancer

are the same as we are not employing any freelancers because the system is already overstaffed, and we can see that as the uncertainty increases we are overstaffing the system more and more hence increasing the quality but trading it off with the cost.

For the high demand period, employing freelancers is helping us in increasing in the quality of service when compared to blended or just flexible. We can also see that as q increases, we are overstaffing with flexible to hedge the uncertainty so we are not hiring any freelancers and the two plots merge.

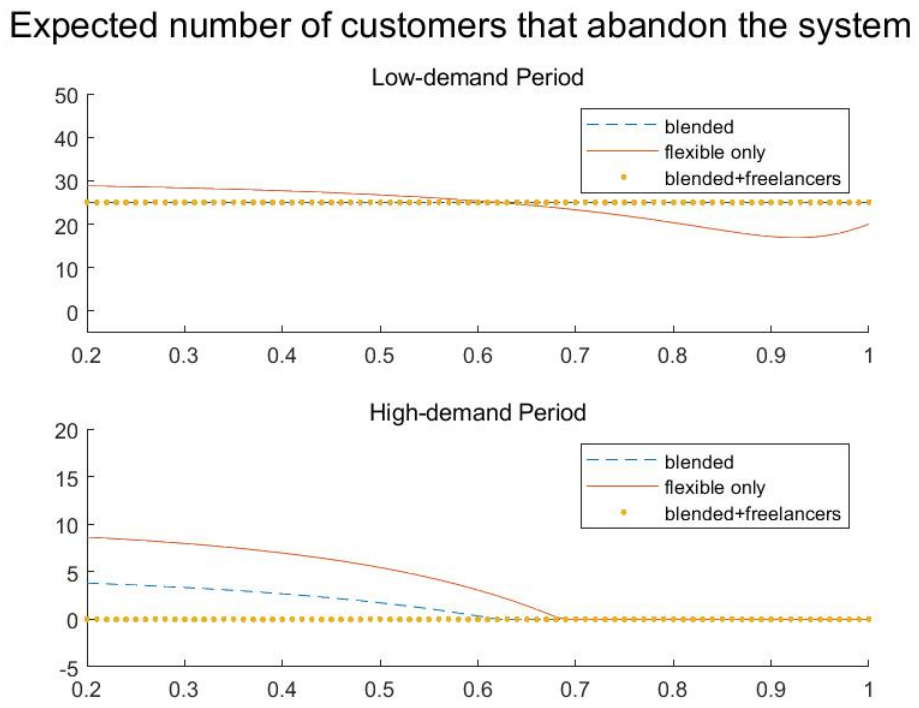

Figure 6.3: Expected number of customers that abandon the system

(a) This image shows the number of customers that are expected to leave the system in the 2 different periods.

CONCLUSION

We have started with a detailed survey to understand parameter uncertainty and self-scheduling servers. We extended those ideas to implement the blended workforce model proposed in *Dong and Ibrahim* [2].As we saw from the numerical results, this model was cost effective. We felt with the changing startup landscape in India and a increased emphasis on quality service,there is a need to extend this model to include freelancers.As shown in numerical results this increased the quality of service with a trade off of cost of the system.

Bibliography

[1] Achal Bassamboo, Ramandeep Randhawa, and Assaf Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56:1668–1686, 10 2010.

[2] Jing Dong and Rouba Ibrahim. Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Oper. Res.*, 68(4):1238–1264, July 2020.

[3] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197, 1917.

[4] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 06 1981.

[5] J. Michael Harrison and Assaf Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Operations Management*, 7(1):20–36, January 2005.

[6] Rouba Ibrahim. Managing queueing systems where capacity is random and customers are impatient. pages 9–16, 8 2017.

[7] John F. Shortle, James M. Thompson, Donald Gross, and Carl M. Harris. *Fundamentals of Queueing theory*, volume 5. Wiley Series In Probability And Statistics, 2018.

[8] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15:88–102, 2006.

[9] Ward Whitt. Fluid models for multiserver queues with abandonments. 1 2006.