MANAGING SUPPLY IN THE INDIAN CONTEXT:

FLEXIBLE WORKERS, FULL-TIME EMPLOYEES AND FREELANCERS

Tanvi Ohri and Manne Hema Priya

TABLE OF CONTENTS

Background and Importance of the Problem

The main aim of the project is to adapt the model proposed by Dong and Ibrahim (2020) for the Indian market. 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 this phase of the project, we survey some of the preliminary works in this area.

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.

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 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. We consider the fluid, and hence, the customers to be infinitely divisible.

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

If $A(t)$ is the cumulative number of arrivals by time t and $D(t)$ is the cumulative number of departures by time t.
The queue length = $A(t) - D(t) \geq 0$ because water can not drain out before it comes out of the tap.

The Fluid Approximation

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

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

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

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

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

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 statements , we have

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

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.

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.

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.

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).

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,

$$C_o = C - S \text{ and } C_u = P - C$$

So,

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

Demand-side Parameter Uncertainty: Introduction

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

1. **Efficiency:** This concerns itself with basic operating costs.
2. **Quality:** This concerns itself with quality of customer service.

Capacity planning is a trade-off between these 2 factors.

Traditional Models
Most of the traditional literature assumes that all model primitives are known with certainty. However, in a practical setting, such parameters are estimated and predicted on the basis of historical data, and hence can be quite “noisy”.

Demand-side Parameter Uncertainty: Introduction

Arrivals: Poisson process, rate λ
Services: exponentially distributed, rate μ.
Some common regimes that are in use give the capacity C as:

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

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

   This model is likely to be under-staffed.
2. **Quality-Driven (QD) Regime:**

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

   This model is likely to be over-staffed.
3. **Quality- and Efficiency-Driven (QED) Regime:**

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

   This takes the form dictated by the The Square Root Staffing Law.

The term $\frac{\lambda}{\mu}$ is a base capacity used to meet mean demands. The second term is the variability hedge.

Proposing an Approximate Solution

Model Parameters

- **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.
- **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.
- **Waiting Policy:** Servers do not remain idle unless queue is empty. There is an infinite capacity buffer. The cost of a customer waiting in queue (the holding cost) is h per customer per unit time.
- **Abandonment Policy:** Customers have the impatience random variable τ, exponentially distributed with mean $\frac{1}{\gamma}$. 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.
- *N* is the random variable that represents the number of customers in the system in steady state.

Proposing an Approximate Solution

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:

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

$$\textit{subject to}\ \ b \geq 0$$

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

It is not possible to get an exact solution to this optimization problem. This is because the distribution of N itself depends on b.

Proposing an Approximate Solution

We model the customer arrivals as a fluid, that is, the 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,

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

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.

Proposing an Approximate Solution

This is an instance of the newsvendor problem 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,

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

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

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

Note: 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.

Analysing the Performance of the Proposed Solution

We consider $c = 1/3, p = 1, h = 1, \mu = 1$ and $\gamma = 1/3$.

Case I: Deterministic Arrival Rates
We consider constant Arrival Rates.

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

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: 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. This value actually grows proportionally to $\sqrt{\lambda}$. This is basically giving us the square-root law.

Analysing the Performance of the Proposed Solution

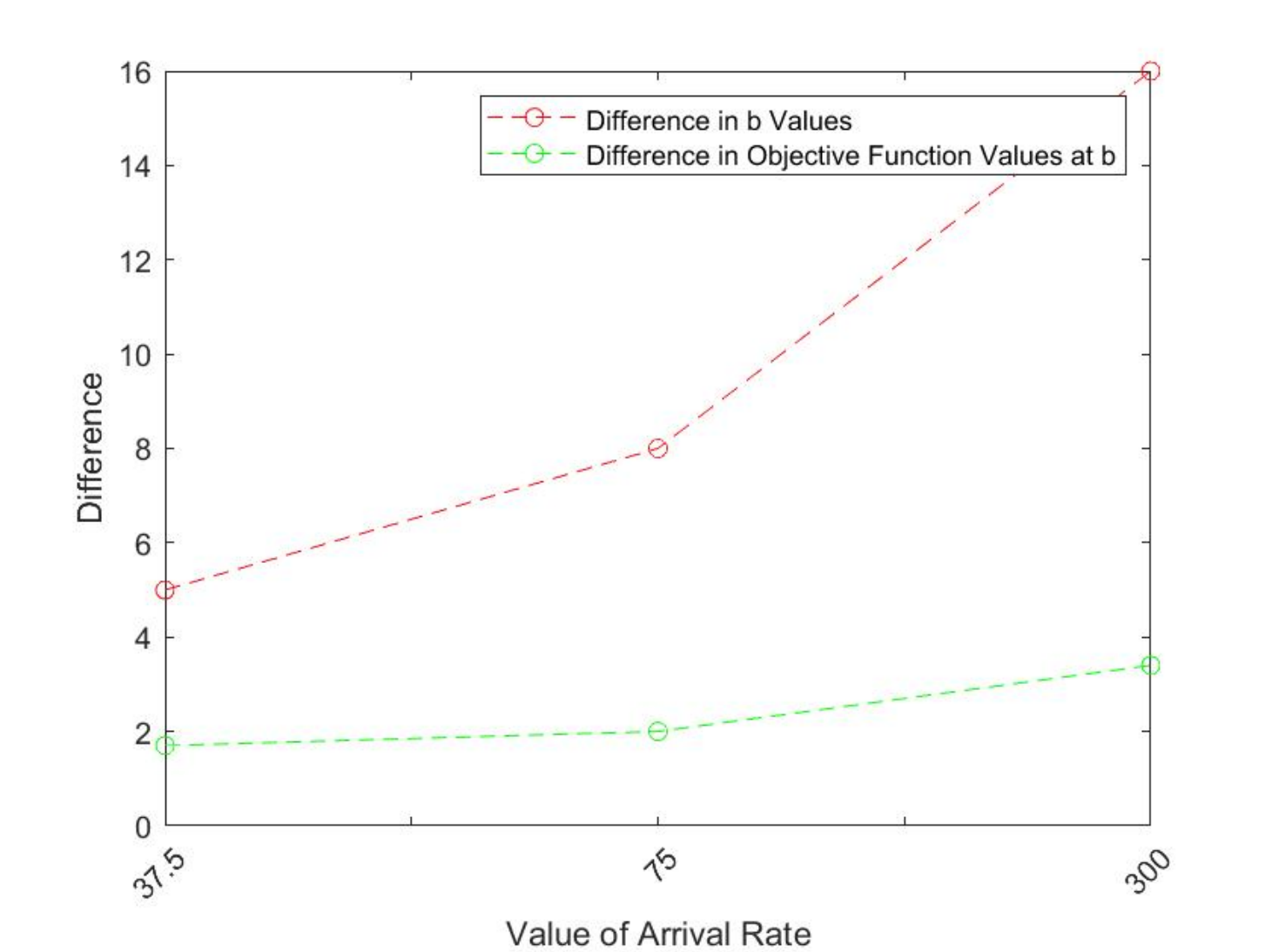

Figure: Performance of the Proposed Solution on Deterministic Arrival Rates

Analysing the Performance of the Proposed Solution

Case II: Uncertain Arrival Rates

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

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

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: Performance of the Proposed Solution on Uncertain Arrival Rates

The difference in the value of optimality gap in the two tables is quite high even though mean of arrival rate in the last 3 rows are same. This shows that the proposed solution works better when the arrival rate has uncertainty.

Analysing the Performance of the Proposed Solution

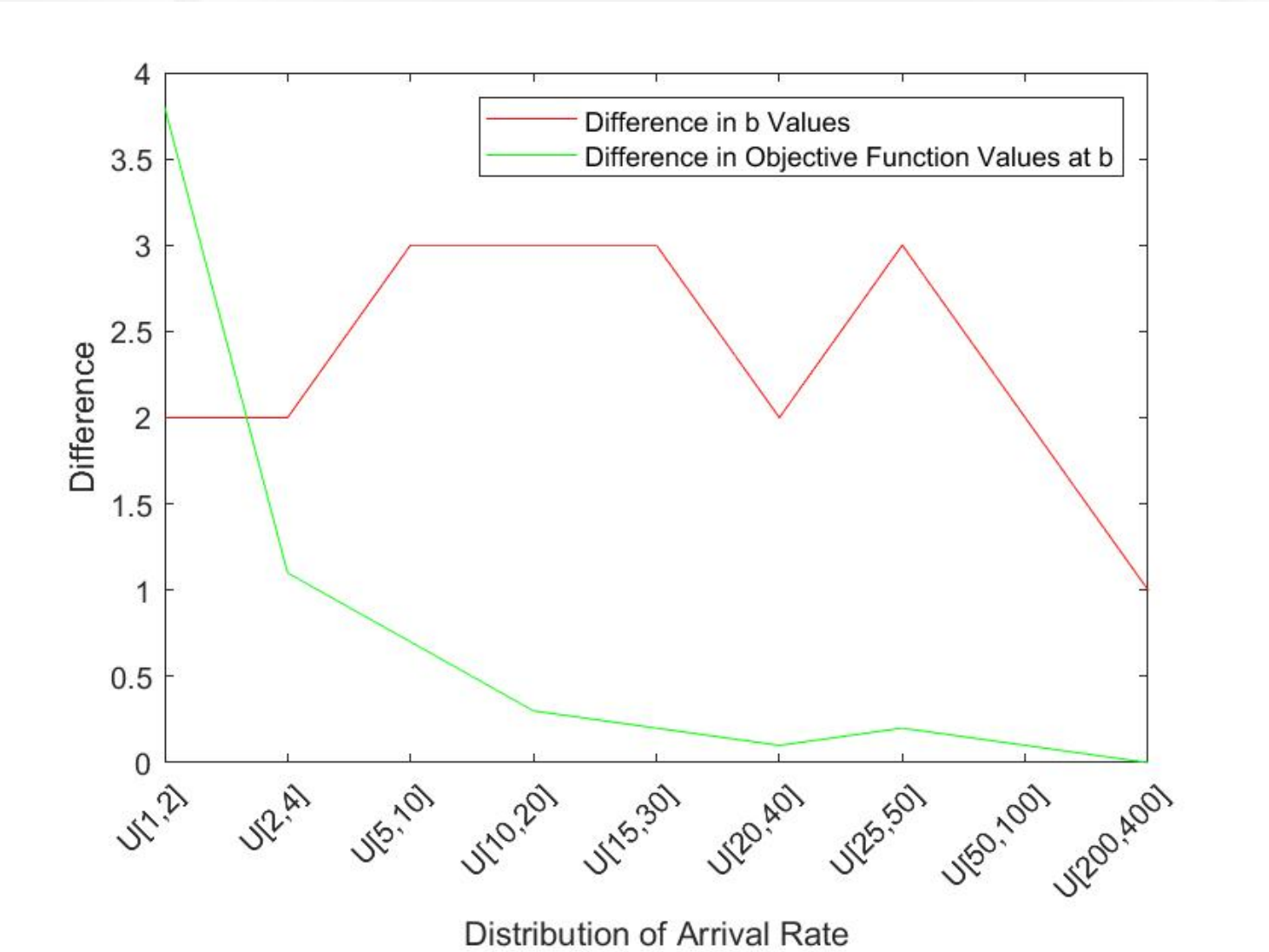

Figure: Performance of the Proposed Solution on Uncertain Arrival Rates

Self scheduling Servers Background

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.

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.

Modelling Framework

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.

System Manager's Problem

Being consistent with *Bassamboo and Randhawa* , 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

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

where c_j is the compensation for each server in j^{th} shift.

Fluid Approximation of System Manager's Problem

Since an exact analysis of (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)*"
So the fluid approximation for (1) is

$$\min_{n \in \mathbb{N}} \Pi(n) = \sum_{1 \le j \le k} nG(c_j)c_j + p_j.\bar{\alpha}_{\rho_s / G(c_j)} + h_j.\bar{q}_{\rho_s / G(c_j)} \tag{2}$$

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.

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

$$q_j = \int_0^{w_j} \lambda_j \bar{F}(u)\, du \tag{3}$$

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

$$\alpha_j = \lambda_j F(w_j) \tag{4}$$

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

$$n_j^* = \lambda \bar{F}(w_j^*) \tag{5}$$

where w_j^* would be the optimal waiting time.
With the results in (5) and (4) the system managers problem becomes

$$min_{w_j \geq 0} \left(\lambda_j c_j \bar{F}(w_j) - \lambda_j p_j \bar{F}(w_j) + \lambda_j h_j \int_0^{w_j} \lambda_j \bar{F}(u)\, du \right) \tag{6}$$

Some results for the bench marking case

Assumption *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 Bassamboo , 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

$$c_j < h_j/\theta + p_j \tag{7}$$

Using this , we will try and show the next proposition
Proposition *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*

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. 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
so the system manager's problem in terms of the augmented arrival rate Γ_j becomes

$$min_{0 \leq n \leq \Gamma_k} C(n) = \left(\sum_{j=1}^{k} I(\Gamma_{j-1} \leq n < \Gamma_j) u_j(n) \right) \tag{8}$$

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

$$u_j(n) = \sum_{i=1}^{k} c_i n r_i + \sum_{i=j}^{k} \left(p_i(\lambda_i - n r_i) + h_i \lambda_i \int_0^{\bar{F}^{-1}(n/\Gamma_i)} \bar{F}(u)\, du \right) \tag{9}$$

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.

Some results for self scheduling case

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 int the benchmark case, 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. □

Some results for self scheduling case

In case of Exponential Abandonment and Non-equal Augmented Arrival Rates:

1. *The optimization function in (8) 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:*

$$\sum_{j=1}^{k} c_j r_j - \sum_{j=i_0}^{k} (p_j + h_j/\theta)\, r_j < 0 \text{ and } \sum_{j=1}^{k} c_j r_j - \sum_{j=i_0+1}^{k} (p_j + h_j/\theta)\, r_j > 0 \qquad (10)$$

Future Steps ...

In the coming semester , we would try to combine the both parameter uncertainty and self scheduler servers and arrive at a optimal staffing for the queuing system taking the *Dong and Ibrahim* paper, also try to introduce a new worker *freelancer* as an extension to that model and look at the fluid approximations considered in this paper, and see if we can forgo them.