




# Optimization of a stochastic model having erratic server with immediate or delayed repair

Radhika Agarwal<sup>1</sup> · Divya Agarwal<sup>1</sup> · Shweta Upadhyaya<sup>1</sup> · Izhar Ahmad<sup>2</sup> 

Accepted: 23 May 2022 / Published online: 5 September 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

The thought to put forward a queuing model proposed in this work was its pertinence in everyday life wherever we can see the uses of computing and networking systems. Industrial software developers and system managers can consider the results of the model to evolve their system for better results. Here we present a novel queueing model having erratic server with delayed repair and balking. Two distinct breakdowns i.e. active and passive breakdown for the system are also considered with their respective amendments. This model is closely related with the smooth functioning of the system during some internal faults (virus attack, electricity failures etc.). The performance indicators which are utilized in enhancing the service standards are obtained using supplementary variable technique. Using ANFIS soft computing technique we have compared the analytical results with those of neuro fuzzy results. Furthermore single and bi-objective minimization problems are considered and minima is obtained using particle swarm optimization and multi objective genetic algorithm respectively. Also, the minimization problems are shown as a convex programming problem to ensure the global optimality of the result. The proposed approach makes it conceivable to accomplish a relevant harmony between operational expenses and administration quality.

**Keywords** Convex programming problem · Particle swarm optimization · Bi-objective optimization · Active–passive breakdown · Delayed repair · Balking

---

✉ Izhar Ahmad  
drizhar@kfupm.edu.sa

Radhika Agarwal  
radhika6696@gmail.com

Divya Agarwal  
dagarwal1@amity.edu

Shweta Upadhyaya  
supadhyay@amity.edu

<sup>1</sup> Amity Institute of Applied Sciences, Amity University, Sector 125, Noida, UP, India

<sup>2</sup> Department of Mathematics, Interdisciplinary Research Center for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

# 1 Introduction

In queueing frameworks, that system is considered as efficient which limits the long-run cost. Now a days due to the developing anxiety of clients for not waiting in queues during the briefest span, estimation of minimum waiting time is also done by few authors to develop a robust system. However in optimization, concept of convexity plays a vital role since it can guarantee the presence of global optimal solution. A local minima of a convex function on a convex feasible region is a global minima. In many papers related to optimization of cost or waiting time of queueing models, the optimal solution is obtained using optimization tools but the sufficient optimality conditions are not verified. In consequence we have tried to design a model which closely relates to the system developed for facilitating mankind and optimized not only the cost of the system but awaiting time as well. Also, we have shown the optimization problems as a convex programming problem by discussing the convexity of the objective functions using Hessian matrix test and tried to obtain the global minima of the minimization problem.

Queueing models with retrials find their application in most modern computer networks that include Wide Area Network (WAN) protocol, Frame Relay, Local Area Network (LAN) protocol, TCP/IP and X.25 as well. Mohammadi et al. (2014) have designed a reliable health-care network where the service provider is erratic with queues of finite capacity. Gao et al. (2020) inspected a queue with retrials along with distinct failures in the model. Also if the server is unoccupied and faces breakdown (passive breakdown), it gets a delayed repair due to unidentified fault whereas if the server is occupied and faces breakdown (active breakdown) then the repair process starts immediately. They have given a promising application of the model in packet-switching network. Before sending forward, the messages are slotted into IP packets which are then sent from a source to destination through the router. Here source host and the destination host are arrivals and servers, respectively. More applications of retrial queueing models with breakdown can be seen in Upadhyaya (2014), Kim and Kim (2016), Taleb and Aissani (2016), Lan and Tang (2020), Lee et al. (2020), Dragieva and Phung-Duc (2020).

## 1.1 Literature review

Study of stochastic models is done by many researchers. Choudhury and Tadj (2009) extended the classical  $M/G/1$  model with secondary service, breakdowns and delayed repair where erratic server randomly breaks-down while serving the clients and further evaluated the busy and awaiting time along with reliability indices. On the same lines, Taleb and Aissani (2016) also incorporated the impatient behaviour of the clients with secondary services while Saggou et al. (2017) focused mainly on recurrent clients.

Now we discuss a well known characteristic of clients called balking (resisting) where the arriving batch may get annoyed and resist to join the queue if the worker is occupied or on holiday. A queue having bulk arrival with retrials, balking together with altered vacations was studied by Upadhyaya (2014). In addition, the upshot of some basic parameters on long run probabilities and mean orbit size is also studied. Seeing the recent developments, Upadhyaya and Kushwaha (2020) elaborated  $M^X/G/1$  queueing model with retrials incorporating the impatience of clients with their feedback and delayed repair. They considered alternate vacation scheme and validated the analytical results using a soft computing technique known as ANFIS (Adaptive Neuro-Fuzzy Interface System) and then built a cost effective system.

Multi-objective optimization is of interest of researchers working in this field. In multi-objective optimization more than one objective function is to be optimized simultaneously so that the Pareto optimal solution can be obtained for the problem. Verma (1986) first introduced the concept of multiobjective optimization in queueing systems. Kahag et al. (2019) did a bi-objective optimization of  $M/M/m$  model for a hub allocation problem in traffic systems using a multi objective invasion weed optimization (MOIWO). Most recently, Wu and Yang (2021) considered a two phase heterogeneous service model. They first did a single objective optimization using *Canonical PSO* and then formulated a bi-objective cost optimization model for the system and awaiting time simultaneously. More studies on multi-objective optimization in queueing models can be read in the work of Mohammadi et al. (2014), Tavakkoli-Moghaddam et al. (2017) and Pourmohammadi et al. (2021).

Although lot of work is done in obtaining the optimal cost of the system for queueing models but only a handful of authors have discussed the concept of convexity. Zhang et al. (1997) studied  $M/G/1$  continuous model with two distinct vacations and developed the average cost function with optimal threshold policies. Later, Zhang (2006) evaluated the cost function's convexity for the model studied in Zhang et al. (1997). Sherman et al. (2009) have examined an erratic queue with retrial where normal and orbit queue have infinite capacity and the clients join the orbit if the server fails. They allow both active as well as idle breakdowns in their system. They have obtained the optimal repair and retrial rates and have done the convexity analysis in detail.

## 1.2 Focus of our study

Motivated by the pertinence of such models,  $M/G/1$  queue with retrials under active and passive breakdown services, delayed repair for passive faults with balking is studied in this work. The main focus is to attain the global minima of the cost function and further getting the global Pareto optimal solution for the bi-objective problem. Application of the developed model can be seen in health care systems, supermarkets, call centres, mobile and computer network organisations. This model is closely related with the smooth functioning of the system during some internal faults (virus attack, electricity failures etc.). Our model firmly reflects the vaccination system developed where effective regular and delayed repair can assume an essential part in conveying quality help rapidly under decreased expense. The manufactures and decision-makers can efficiently apply the outcomes of this model in developing system management policies. In addition to its functional pertinence, the model presented interesting numerical properties that permits the examination by their own doing.

## 1.3 Layout of the paper

The remaining article is composed as follows: Sect. 2 includes the basic notations and preliminaries that are to be considered throughout. Section 3 depicts queueing model with a practical application. The steady state study of the model is referred in Sect. 4 where the governing and boundary equations, along with the probability generating function (pgf) of the system and performance indices are evaluated. Section 5 elaborates the effect of some important parameters on the queue length with ANFIS validation. The single and multi-objective optimization of the model is discussed in Sect. 6 where in sub-segment 6.1, the expected cost function is fabricated to accomplish the ideal estimations of certain parameters and get the optimal cost value using PSO. Later in sub-segment 6.2, the bi-objective optimization where

two objectives: the awaiting time in the system and the cost function are optimized and lastly Sect. 7 incorporate concluding remarks along with future scope of the paper.

## 2 Preliminaries and notations

### 2.1 Preliminaries

For  $\bar{\omega}, \bar{\Omega} \in R^n$ , following notations for inequalities will be followed as:

$$\begin{aligned} \bar{\omega} \leq \bar{\Omega} &\iff \bar{\omega}_j \leq \bar{\Omega}_j, & j = 1, \dots, n; \\ \bar{\omega} \leq \bar{\Omega} &\iff \bar{\omega} \leq \bar{\Omega}, & \bar{\omega} \neq \bar{\Omega}; \\ \bar{\omega} < \bar{\Omega} &\iff \bar{\omega}_j < \bar{\Omega}_j, & j = 1, \dots, n. \end{aligned}$$

**Definition 1** Any function  $T : R^n \rightarrow R$  is convex if for  $\tilde{x} \in R^n$  we have  $T[k\tilde{x} + (1 - k)\tilde{x}] \leq kT(\tilde{x}) + (1 - k)T(\tilde{x})$  true  $\forall k$  where  $0 \leq k \leq 1$ . The function  $T$  is strictly convex if the above inequality holds strictly for  $x \neq \tilde{x}$ , and  $0 < k < 1$ .

**Definition 2** If all partial derivatives of  $T$  exists and are continuous over the domain, then the Hessian matrix  $M$  is a square  $n \times n$  matrix arranged as

$$M = \begin{bmatrix} \frac{\partial^2 T}{\partial x_1^2} & \frac{\partial^2 T}{\partial x_1, \partial x_2} & \dots & \frac{\partial^2 T}{\partial x_1, \partial x_n} \\ \frac{\partial^2 T}{\partial x_2, \partial x_1} & \frac{\partial^2 T}{\partial x_2} & \dots & \frac{\partial^2 T}{\partial x_2, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 T}{\partial x_n, \partial x_1} & \frac{\partial^2 T}{\partial x_n, \partial x_2} & \dots & \frac{\partial^2 T}{\partial x_n} \end{bmatrix}.$$

Assume  $S$  to be a nonempty open convex set in  $R^n$ , and let  $T : R^n \rightarrow R$  be two times differentiable on  $S$ . Then  $T$  is said to be convex iff the matrix  $M$  is positive semidefinite at each point in  $S$ . For more details refer Bazaraa et al. (2005).

Consider an optimization problem

$$(MP) \quad \text{Minimize } T(x) = (T_1(x), T_2(x), \dots, T_k(x))$$

where  $x \in X$  is an open subset of  $R^n$  and  $T : X \rightarrow R^k$  are differentiable on  $X$ . If  $k = 1$ , problem is called a scalar otherwise multi objective optimization problem.

**Definition 3 (Global Optimal point)** If in MP  $k = 1$ , any point  $\tilde{x} \in R^n$  is an optimal point if  $\exists$  no  $x \in X$  s.t.  $T(x) < T(\tilde{x})$ .

**Definition 4 (Global Pareto Optimality)** A point  $\tilde{x}$  is a global efficient/Pareto optimal solution of MP if  $\exists$  no  $x \in X$  s.t.  $T(x) \leq T(\tilde{x})$ .

Obviously any global Pareto solution is local Pareto optimal but converse is not always true. The converse holds for a convex multi-objective programming problem. Also, the multi-objective problem of optimization is convex if all the objectives and feasible region is convex. For more details refer Miettinen (1998).

## 2.2 Basic notations

$\lambda$	: Poisson arrival rate of the clients.
$\alpha$	: Exponential retrial rate of the clients.
$b(\bar{b} = 1 - b)$ ,	: If the system is occupied then the client's joining (resisting) probability from the orbit.
$\delta, \theta$	: Exponential breakdowns, rate for passive and active breakdowns, respectively.
$B, R, G$	: Random variable denoting general distribution of the serving time, repair time of active and passive breakdowns, respectively.
$B(x_1), R(x_1), G(x_1)$	: Cumulative density function (cdf) of the serving time, repair time of active and passive breakdowns, respectively.
$b(x_1), r(x_1), g(x_1)$	: Probability generating function (pgf) of the serving time, repair time of active and passive breakdowns, respectively.
$\beta_i, v_i, \mu_i$	: $i^{th}$ moment of serving time, repair time of active and passive breakdowns, respectively.
$\beta(x_1), v(x_1), \mu(x_1)$	: Conditional complementary rates for serving time, repair time of active and passive breakdowns, respectively.

## 3 Model representation and Practical illustration

In this segment, the mathematical model with stability condition is included along with a practical application of the model in real life scenario. In sub-segment. 3.1, a detailed model description is given with a schematic diagram as in Fig. 1 for the readers. sub-segment 3.2 provides a practical application using example for better understanding of the model. Further in sub-segment. 3.3 Markov chain is defined and the system's condition to be stable is stated.

### 3.1 Model description

We are considering an erratic retrial line with two distinct active and passive breakdowns and delayed repair because of the latter. The basic assumptions for the model are:

- *Arrivals and balking*: The arrival rate of a client is  $\lambda$  following Poisson process. If the event is that the server is occupied or is uncertain, then, at that point the arriving client may get annoyed and resist joining the queue with  $b$  probability or may exit from the system with  $\bar{b}$  probability.
- *Service and retrial policy*: The entering client, on seeing the server unoccupied gets the service straight away. Else on seeing the server occupied or inoperative, the clients will have to retry later from a virtual orbit following FCFS principle. The client on the top of the virtual track (orbit) retries to get served with exponential rate  $\alpha$ , if the service provider is unoccupied.
- *Breakdowns and repairs*: The server may face active and passive breakdowns in occupied and unoccupied states, respectively. In the active breakdown, the server just instantly starts repairing. While in passive breakdown, the unnoticed server stays inactive until a client enters and then only the server enters delayed repair state. In both the cases the service of the client continues (starts) its service once the server is ready to serve.

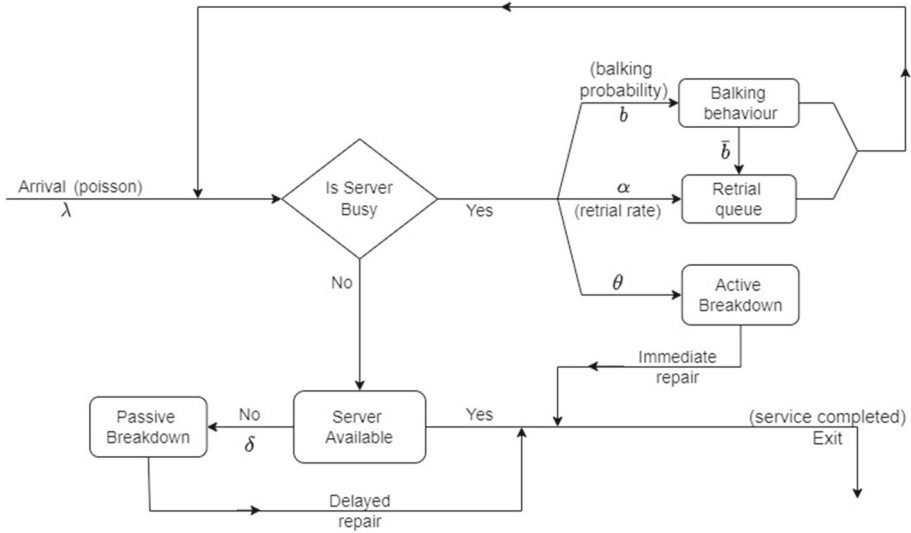


Fig. 1 Schematic diagram of the model

We assume the conditional complementary rates when the service provider is occupied, for repair time of active and passive failures, respectively, as

$$\beta(y) = \frac{b(y)}{\bar{B}(y)}, v(y) = \frac{r(y)}{\bar{R}(y)}, \mu(y) = \frac{g(y)}{\bar{G}(y)}.$$

For the remaining paper, if we consider CDF as  $F(y)$ , where  $y$  is any variable, then its complement is  $\bar{F}(y) = 1 - F(y)$ , its Laplace-Stieltjes transform (LST) is  $\tilde{F}(\kappa) = \int_0^\infty e^{-\kappa y} dF(y)$  and the Laplace transform (LT) of  $\bar{F}(y)$  is  $\bar{F}^*(\kappa) = \int_0^\infty e^{-\kappa y} (1 - F(y)) dy$  which gives  $\bar{F}^*(\kappa) = \frac{1 - \tilde{F}(\kappa)}{\kappa}$ .

### 3.2 Practical examples

A practical example for our retrial queue can be seen if a client goes to a bank to deposit cash or get the passbook updated. We assume that there is only one employ (service provider) in the bank who is assigned to do the job of cash deposit and passbook updates. The clients arriving the bank may get immediate service if the computer system (server) is idle. Else they will have to wait in a queue and try again after some time (retrial) for their service if the computer is already occupied with the earlier present client. Now the retrying clients may get restless and may want to exit from the bank without getting their work done (balk) if they don't have much time to wait. In that case the client may leave the bank (system) without service. We know that the bank's employ have all the data and information of all the clients in their computer system and being a machinery product it may face some faults (breakdown). On a new day initially the system is idle and if there is any fault in the computers at that point (passive breakdown) the bank's employ won't know it. As soon as the client enters the bank, the employ starts the computer to serve him and only then he will know about the fault in the computer and will send it to get repaired (delayed repair). The time interval between point when the passive fault occurred and the point at which the client entered is termed as

delayed period. The client that arrived during delayed period will be provided service once the repair work is done. Another possibility is that while the employer is serving the client the computer system may get a virus attack and hence may breakout (active breakdown). In this situation, it will have to be send for immediate repair. This bank scenario can be easily modelled with our retrial model having two faults and delayed repair.

The model closely fits to the recent Covid-19 scenario also where the main focus of all the countries and governments is to vaccinate maximum people as early as possible. It is not an easy task for a country like India with population of around 140 crores. Thus the whole vaccination process is divided into two stages. First stage includes the booking of the vaccination slots. Individuals have to book their respective slots to get vaccinated through an app like Cowin app in India. Our model fits into this situation where the individuals who are supposed to be vaccinated can be treated as patients and the website/app/computer system is the service provider where the slots are to be booked. The system may face some faults in the idle state (passive breakdown) whose repair will start once the patient starts the system for booking. Thus, there will be a delay in the repair for a passive breakdown (delayed repair). When the system is repaired and is ready to work upon, the patient will continuously try to book the vaccination slot. In India, the booking slots were restricted on daily basis. As lakhs of people try booking at the same time, the server could be busy at particular instances and so the patients have to retry for the booking after an interval of time (retrials). Some of the patients might get impatient and may want to resist from retrying, so they may leave the system without booking the slot (balking). Sometimes due to continuous retrials and busy server, the system may get heated up or may face a virus attack and may breakdown in the middle of the booking process (active breakdown). In this situation the system will be repaired immediately and then the service will be accomplished. Finally, when a slot would be booked for an individual, the service is completed. The second phase includes the vaccination center allotted while slot booking. Here the servers are the nurses assigned for vaccination purpose. The queue of patients wait outside the center for their turn. We aim to bring down the cost of the system along with the awaiting time of the patients. All the components of our model fits in this real-life situation and thus we try to get the optimal repair and retrial rates so that one can book the slot at the earliest and get vaccinated with minimum system cost and awaiting time.

### 3.3 Stability condition

We consider  $S_G$  to be the generalized time interval to serve the clients from starting to the end of the service, where CDF is  $S_G(y)$  and LST is  $\tilde{S}_G(y)$ . If we consider the possibility of active faults in the service, we get  $\tilde{S}_G(y) = \tilde{B}(y + \theta(1 - \tilde{R}(y)))$ , which gives  $E[S_G] = \beta_1(1 + \theta v_1) = \beta_1^*$  and  $E[S_G^2] = \beta_2(1 + \theta v_1)^2 + \theta\beta_1 v_2 = \beta_2^*$ .

Next we define

$$a_k = \int_0^\infty \frac{(\lambda bt)^k}{k!} e^{-\lambda bt} dG(t), \quad k = 0, 1, 2, \dots,$$

$$h_k = \int_0^\infty \frac{(\lambda bt)^k}{k!} e^{-\lambda bt} dS_G(t), \quad k = 0, 1, 2, \dots,$$

where  $a_k$  denotes the likelihood of  $k$  clients entering the orbit if the service provider is under repair of passive fault,  $h_k$  denotes the likelihood of  $k$  clients joining the virtual track during generalized service time. Defining  $c_k = \sum_{i=0}^k a_i h_{k-i}$ ,  $k \geq 0$  where  $c_k$  denotes the likelihood

of  $k$  clients entering in orbit when the service provider is in generalized service and under the passive repair time.

Let  $\Delta(x_3) = \lambda b(1 - x_3)$ . Then we define

$$\begin{aligned}
 H(x_3) &= \sum_{k=0}^{\infty} x_3^k h_k = \tilde{B}(\Delta(x_3) + \theta(1 - \tilde{R}(\Delta(x_3))))), \\
 H'(1) &= \left. \frac{dH(x_3)}{dt} \right|_{x_3=1} = \lambda b \beta_1^* = \rho, \\
 H''(1) &= \left. \frac{d^2 H(x_3)}{dt^2} \right|_{x_3=1} = (\lambda b)^2 \beta_2^*, \\
 A(x_3) &= \sum_{k=0}^{\infty} x_3^k a_k = \tilde{G}(\Delta(x_3)), \\
 A'(1) &= \left. \frac{dA(x_3)}{dt} \right|_{x_3=1} = \lambda b \mu_1 = \rho_1, \\
 A''(1) &= \left. \frac{d^2 A(x_3)}{dt^2} \right|_{x_3=1} = (\lambda b)^2 \mu_2, \\
 C(x_3) &= \sum_{k=0}^{\infty} x_3^k c_k = A(x_3)H(x_3).
 \end{aligned}$$

Considering  $T_k$  ( $T_0 = 0$ ) as the time epoch when the  $k$ th client exits from the system,  $O_k = O(T_k)$  be the size of the virtual track when the  $K$ th exit takes place, thus  $O_k$  for  $k \geq 0$  is a Markov chain process with state space given by  $\mathbb{N}$ . The theorem stated below gives the necessary and sufficient condition for the stable system.

**Theorem 3.1** *The Markov chain  $\{O_k, K \geq 0\}$  is stable iff  $\rho + \frac{\delta \rho_1}{\lambda b + \alpha + \delta} < \frac{\alpha}{\lambda b + \alpha}$ .*

**Proof** One can see the proof of Theorem 3.1 from Gao et al. (2020) and can establish the prove for the above theorem on same lines. □

### 4 Scrutinizing the steady state

We use supplementary variable technique (SVT) to procure the steady state solution along with the pgf of system size in this sub-plot. At a particular time  $t$ , the Markov process defines the system’s state as  $\{N'(t), S(t), \tau_1(t), \tau_2(t), \tau_4(t), t \geq 0\}$ , where  $N'(t)$  is the patient’s number and  $S(t)$  shows the server’s state which is as follows:

$$S(t) = \begin{cases} 0, & \text{unoccupied server,} \\ 1, & \text{occupied server,} \\ 2, & \text{repair of server’s active fault,} \\ 3, & \text{server’s delayed repair period,} \\ 4, & \text{repair of passive breakdown of the server,} \end{cases}$$

where when  $S(t) = 1$ ,  $\tau_1(t)$  gives the passed service time;  $S(t) = 2$ ,  $\tau_2(t)$  gives the passed repair time of active breakdown;  $S(t) = 4$ ,  $\tau_4(t)$  gives the passed repair time of passive breakdown;



The steady state probabilities and their densities are defined as follows:

$$\begin{aligned}
 Q_{n,i} &= P(N' = n, S = i) \\
 &= \lim_{t \rightarrow \infty} Q_{n,i}(t), \quad n \geq 0, \quad i = 0, 3, \\
 Q_{n,i}(x_1)dx_1 &= P(N' = n, S = i, x_1 < \tau_i \leq x_1 + dx_1) \\
 &= \lim_{t \rightarrow \infty} Q_{n,i}(t, x_1)dx_1, \quad n \geq 0, \quad i = 1, 4, \quad x_1 \geq 0, \\
 Q_{n,2}(x_1, x_2)dx_1dx_2 &= P(N' = n, S = 2, x_1 < \tau_1 \leq x_1 + dx_1, x_2 < \tau_4 \leq x_2 + dx_2) \\
 &= \lim_{t \rightarrow \infty} Q_{n,i}(t, x_1, x_2)dx_1dx_2, \quad n \geq 0, \quad x_1, x_2 \geq 0,
 \end{aligned}$$

### 4.1 The governing and boundary equations

Using the SVT we proceed with the basic equations defining steady states as:

$$(\lambda + \delta)Q_{0,0} = \int_0^\infty Q_{0,1}(x_1)\beta(x_1)dx_1, \tag{1}$$

$$(\lambda + \delta + \alpha)Q_{n,0} = \int_0^\infty Q_{n,1}(x_1)\beta(x_1)dx_1, \quad n \geq 1, \tag{2}$$

$$\begin{aligned}
 \frac{d}{dx_1} Q_{n,1}(x_1) &= -(\lambda b + \theta + \beta(x_1))Q_{n,1}(x_1) + \lambda b Q_{n-1,1}(x_1) \\
 &+ \int_0^\infty Q_{n,2}(x_1, x_2)v(x_2)dx_2, \quad n \geq 0, \quad x_1 > 0,
 \end{aligned} \tag{3}$$

$$\begin{aligned}
 \frac{d}{dx_1} Q_{n,2}(x_1, x_2) &= -(\lambda b + v(x_2))Q_{n,2}(x_1, x_2) \\
 &+ \lambda b Q_{n-1,2}(x_1, x_2), \quad n \geq 0, \quad x_1, x_2 > 0,
 \end{aligned} \tag{4}$$

$$\lambda b Q_{0,3} = \delta Q_{0,0}, \tag{5}$$

$$(\lambda b + \alpha)Q_{n,3} = \delta Q_{n,0}, \quad n \geq 1, \tag{6}$$

$$\frac{d}{dx_1} Q_{n,4}(x_1) = -(\lambda b + \mu(x_1))Q_{n,4}(x_1) + \lambda b Q_{n-1,4}(x_1), \quad n \geq 0, \quad x_1 > 0. \tag{7}$$

The boundary conditions that are to be used in solving Eqs. (1)–(7) are:

$$Q_{n,1}(0) = \lambda b Q_{n,0} + \alpha Q_{n+1,0} + \int_0^\infty Q_{n,4}(x_1)\mu(x_1)dx_1, \quad n \geq 0, \tag{8}$$

$$Q_{n,2}(x_1, 0) = \theta Q_{n,1}, \quad n \geq 0, \quad x_1 > 0, \tag{9}$$

$$Q_{n,4}(0) = \lambda b Q_{n,3} + \alpha Q_{n+1,3}, \quad n \geq 0, \tag{10}$$

where the normalizing condition is

$$\begin{aligned}
 \sum_{n=0}^\infty (Q_{n,0} + Q_{n,3}) &+ \sum_{n=0}^\infty \int_0^\infty (Q_{n,1}(x_1) + Q_{n,4}(x_1))dx_1 \\
 &+ \sum_{n=0}^\infty \int_0^\infty \int_0^\infty Q_{n,2}(x_1, x_2)dx_1dx_2 = 1.
 \end{aligned} \tag{11}$$

### 4.2 Probability generating function for the system states

We next define the generating functions that are to be used during further calculations as:

$$\begin{aligned}
 Q_i(x_3) &= \sum_{n=0}^{\infty} Q_{k,i}x_3^n, \quad i = 0, 3, \\
 Q_i(x_1, x_3) &= \sum_{n=0}^{\infty} Q_{n,i}(x_1)x_3^n, \quad i = 1, 4, \\
 Q_2(x_1, x_2, x_3) &= \sum_{n=0}^{\infty} Q_{n,2}(x_1, x_2)x_3^n.
 \end{aligned}$$

**Theorem 4.1** *The generating functions for the joint stationary distribution of the states of the service provider are as follows:*

$$\begin{aligned}
 Q_0(x_3) &= \frac{\alpha(\lambda b + \alpha)(x_3 - H(x_3)) - \delta\alpha C(x_3)(1 - x_3)}{\bar{D}(x_3)} Q_{0,0}, \\
 Q_1(x_1, x_3) &= \frac{N_a(x_3)}{\bar{D}(x_3)} (1 - B(x_1))\alpha Q_{0,0} \times \exp\{-\Delta(x_3) + \theta(1 - \tilde{R}(\Delta(x_3)))x_1\}, \\
 Q_2(x_1, x_2, x_3) &= \theta Q_1(x_1, x_2)(1 - R(x_2))\exp\{-\Delta(x_3)x_2\}, \\
 Q_3(x_3) &= \frac{\delta Q_{0,0}}{\lambda b + \alpha} \times \left[ \frac{\alpha}{\lambda b} + \frac{\alpha(\lambda b + \alpha)(x_3 - H(x_3)) - \delta\alpha C(x_3)(1 - x_3)}{\bar{D}(x_3)} \right], \\
 Q_4(x_1, x_3) &= \frac{N_b(x_3)}{\bar{D}(x_3)} (1 - G(x_1))\exp\{-\Delta(x_3)x_1\}\delta\alpha Q_{0,0},
 \end{aligned}$$

where

$$\begin{aligned}
 Q_{0,0} &= \frac{\lambda b}{\alpha} \frac{(\alpha + \lambda b)(\alpha + \lambda b + \delta)}{(\delta + \lambda b)(\alpha + \lambda b + \delta) + \delta(\alpha + \lambda b)\rho_1} \left( \frac{\alpha}{\alpha + \lambda b} - \rho - \frac{\delta\rho_1}{\alpha + \lambda b + \delta} \right), \\
 N_a(x_3) &= \delta A(x_3)(\alpha + \lambda b x_3 + (\alpha + \lambda b + \delta)(x_3 - 1)) - (\alpha + \lambda b)(\delta + \Delta(x_3)), \\
 N_b(x_3) &= (\alpha + \lambda b x_3)(1 - H(x_3)) + (\alpha + \lambda b + \delta)(x_3 - 1), \\
 \bar{D}(x_3) &= (\alpha + \lambda b)[(\alpha + \lambda b + \delta)x_3 - (\alpha + \lambda b x_3)H(x_3)] - \delta(\alpha + \lambda b x_3)C(x_3).
 \end{aligned}$$

**Proof** Using Eqs. (1)–(10) and doing some algebraic manipulations, the proof of the theorem can be obtained easily. □

**Theorem 4.2** *If there is a service provider which is occupied, under repair of active or passive fault, then the marginal pgfs of the size of orbit are as follows:*

$$\begin{aligned}
 Q_1(x_3) &= \frac{N_a(x_3)}{\bar{D}(x_3)} \frac{1 - H(x_3)}{\Delta(x_3) + \theta(1 - \tilde{R}(\Delta(x_3)))} \alpha Q_{0,0}, \\
 Q_2(x_3) &= \frac{N_a(x_3)}{\bar{D}(x_3)} \frac{1 - H(x_3)}{\Delta(x_3) + \theta(1 - \tilde{R}(\Delta(x_3)))} \frac{1 - \tilde{R}(\Delta(x_3))}{\Delta(x_3)} \theta \alpha Q_{0,0}, \\
 Q_4(x_3) &= \frac{N_b(x_3)}{\bar{D}(x_3)} \frac{1 - A(x_3)}{\Delta(x_3)} \delta \alpha Q_{0,0}.
 \end{aligned}$$

**Proof** We already have marginal probabilities for unoccupied and delayed repair state from the above theorem. So, we now need to get the marginal probabilities for the

busy and both the repair states. Thus evaluate  $Q_1(x_3) = \int_0^\infty Q_1(x_1, x_3)dx_1$ ,  $Q_2(x_3) = \int_0^\infty \int_0^\infty Q_2(x_1, x_2, x_3)dx_1dx_2$  and  $Q_4(x_3) = \int_0^\infty Q_4(x_1, x_3)dx_1$  using Theorem 4.1 and hence the proof can be obtained.  $\square$

**Theorem 4.3** *The PGFs  $\phi(x_3)$  and  $\psi(x_3)$  of the client’s number in the orbit and system respectively are*

$$\phi(x_3) = \frac{(x_3 - 1)(\alpha + \delta + \lambda b)(\delta + \lambda b) + \delta(\alpha + \lambda b)[C(x_3) - H(x_3)]}{\bar{D}(x_3)} \frac{\alpha}{\lambda b} Q_{0,0} ,$$

$$\psi(x_3) = \frac{\delta C(x_3)[(\delta + \lambda b)(x_3 - 1) + (\alpha + \lambda b)x_3] - (\alpha + \lambda b)(\delta + \Delta(x_3))H(x_3)}{\bar{D}(x_3)} \frac{\alpha}{\lambda b} Q_{0,0} .$$

**Proof** Let us suppose that  $\phi(x_3) = E[x_3^{N_0}]$  and  $\psi(x_3) = E[x_3^{N_1}]$  where  $N_0$  and  $N_1$  denotes the customer’s number in the orbit and system respectively. Then using  $\phi(x_3) = \sum_{n=0}^\infty Q_j(x_3)$  and  $\psi(x_3) = Q_0(x_3) + x_3Q_1(x_3) + x_3Q_2(x_3) + Q_3(x_3) + x_3Q_4(x_3)$ , proof of the theorem can be obtained.  $\square$

### 4.3 Performance indices

The aim of this sub-segment is to provide the important performance indices of the queuing system using the results of sub-segment 4.2.

**Theorem 4.4** (A) *Taking into account the steady state conditions,:*

- $Q_a$ - probability of service provider being unoccupied is

$$Q_a = Q_0(1) = \frac{\lambda b[\delta + (\alpha + \lambda b)(1 - \rho)]}{(\delta + \lambda b)(\alpha + \lambda b + \delta) + \delta\rho_1(\alpha + \lambda b)} .$$

- $Q_b$ - probability that the service provider is occupied is

$$Q_b = Q_1(1) = \frac{\lambda b(\alpha + \lambda b) + \delta((\alpha + \lambda b)\rho_1 + \delta + \alpha + 2\lambda b)}{(\delta + \lambda b)(\alpha + \lambda b + \delta) + \delta\rho_1(\alpha + \lambda b)} \lambda b\beta_1 .$$

- $Q_c$ - probability of service provider being under repair of active fault is

$$Q_c = Q_2(1) = \theta v_1 Q_b .$$

- The probability  $Q_d$  that the service provider is under delayed repair is

$$Q_d = Q_3(1) = \frac{(\alpha + \lambda b + \delta)(1 - \rho) - \lambda b\rho - \delta\rho_1}{(\delta + \lambda b)(\alpha + \lambda b + \delta) + \delta\rho_1(\alpha + \lambda b)} \delta .$$

- The probability  $Q_e$  that the service provider is under repair of passive fault is

$$Q_e = Q_4(1) = \frac{\delta + (\alpha + \lambda b)(1 - \rho)}{(\delta + \lambda b)(\alpha + \lambda b + \delta) + \delta\rho_1(\alpha + \lambda b)} \delta\rho_1 .$$

(B) *The mean orbit ( $L_0$ ) and system size ( $L_1$ ) are*

$$L_0 = \frac{N''(1)}{2N'(1)} - \frac{\bar{D}''(1)}{2\bar{D}'(1)} = Q_f + Q_g, \tag{12}$$

$$L_1 = Q_b + Q_c + Q_e + L_0 .$$

if

$$\begin{aligned}
 N &= (x_3 - 1)(\alpha + \delta + \lambda b)(\delta + \lambda b) + \delta(\alpha + \lambda b)(C(x_3) - H(x_3)), \\
 N'(1) &= (\alpha + \delta + \lambda b)(\delta + \lambda b) + \delta(\alpha + \lambda b)\rho_1, \\
 N''(1) &= 2\delta\lambda b\rho_1 + \delta(\alpha + \lambda b)((\lambda b)^2\mu_2 + 2\rho\rho_1), \\
 \bar{D}'(1) &= (\alpha + \delta + \lambda b)(\alpha + \lambda b) \left[ \frac{\alpha}{\alpha + \lambda b} - \rho - \frac{\delta\rho_1}{\alpha + \lambda b + \delta} \right], \\
 \bar{D}''(1) &= -(\alpha + \lambda b) [2\lambda b\rho + (\alpha + \lambda b)((\lambda b)^2\beta_2^*)], \\
 &\quad - \delta [2\lambda b(\rho + \rho_1) + (\alpha + \lambda b)((\lambda b)^2\mu_2 + 2\rho\rho_1 + (\lambda b)^2\beta_2^*)].
 \end{aligned}$$

(C) The mean awaiting time in the orbit ( $W_0$ ) and system ( $W_1$ ) are

$$\begin{aligned}
 W_0 &= \frac{L_0}{\lambda}, \\
 W_1 &= \frac{L_1}{\lambda}.
 \end{aligned} \tag{13}$$

**Proof**(A) Substituting  $x_3 = 1$  in Theorem 4.1 and 4.2 and directly calculating using L'Hopital's differentiation rule, we can prove it.

(B) Using  $L_0 = E[N_0] = d\phi(x_3)t_{x_3=1}$  and  $L_1 = E[N_1] = d\psi(x_3)t_{x_3=1}$  the mean orbit and system size can be obtained.

(C) From Little's theorem, the results can be established easily. □

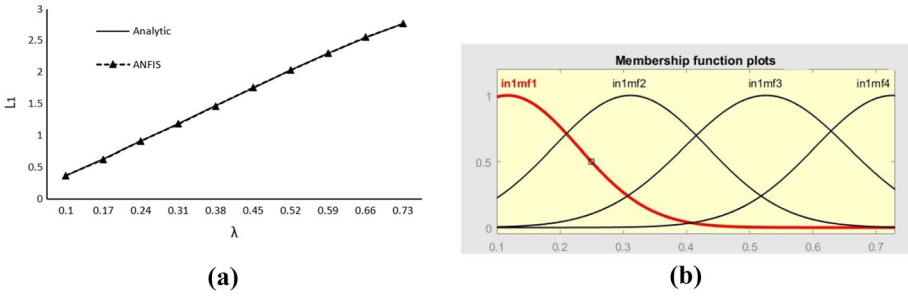
### 4.4 Special cases

Here we present some particular cases. It can be seen that under certain conditions, the system size of our model reduces to the expression obtained for previously studied models.

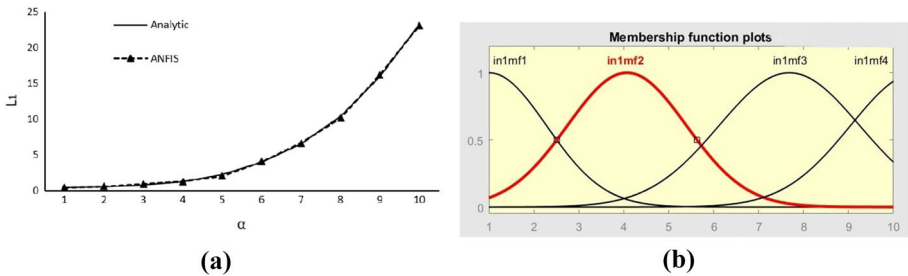
- (i) When  $b = 1$  (No balking), we observe that the model reduces to the model studied by Gao et al. (2020) and the results coexist.
- (ii) When  $\delta = 0$  (No passive breakdown), we observe the change as the model changes to continuous queue with constant retrial and active breakdown and the results coincide with that of Jin-ting (2006) on considering  $A(u) = 1 - e^{-au}; u > 0$ .
- (iii) When  $\delta = 0, \theta = 0, b = 1$  (No breakdowns, no repair and no balking), the results of our model coincide with the model studied by Gomes Corral (1999).
- (iv) When  $b = 1, \theta = 0$  (No balking and no repair), the results of our model coincide with those of Taleb and Aissani (2016) on considering no persistent, impatient clients and preventive maintenance in their model.

### 5 Numerical instance

For the numerical example, we assume that repair times and service times follow Erlangian distribution. The repair time for passive failure has mean  $\mu_1 = 5/\mu$  and variance  $\mu_2 = 5/\mu^2$ . The time to repair the active fault has mean  $v_1 = 2/v$  and variance  $v_2 = 2/v^2$ . The service time has mean  $\beta_1 = 2/s$  and variance  $\beta_2 = 2/s^2$ . The retrial time and passive and active breakdown values are given as:  $\lambda = 0.15, b = 0.25, \alpha = 2, \delta = 0.2, \theta = 0.1, \mu = 5, v = 5, s = 0.75$ .



**Fig. 2** a  $L_1$  vs  $\lambda$  for analytic and ANFIS results; b membership function for  $\lambda$



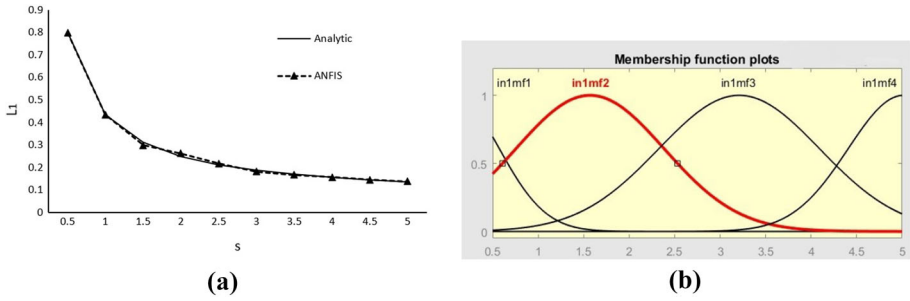
**Fig. 3** a  $L_1$  vs  $\alpha$  for analytic and ANFIS results; b membership function for  $\alpha$

A soft computing technique, ANFIS is a robust tool to discover significant outcomes that are applicable in everyday crowding situation. This technique helps in finding estimated solution for the measures whose definite outcomes are otherwise difficult to obtain. In ANFIS technique, we handle the parameters ( $\lambda$ ,  $\alpha$ ,  $s$  and  $\theta$ ) as linguistic variables which are executed for 4 epochs each. The linguistic values for all the parameters are defined as follows: less, moderate, high and extreme. We have used Gaussian membership function to present all the linguistic variables as shown in Figs. 2b, 3b, 4b and 5b for  $\lambda$ ,  $\alpha$ ,  $s$  and  $\theta$ , respectively. In Figs. 2a, 3a, 4a and 5a, the solid line shows the analytical results whereas dotted lines with solid symbol show the ANFIS results. It is clear from the figures that the analytical results for the model coincide with those of the neuro fuzzy results attained using ANFIS technique.

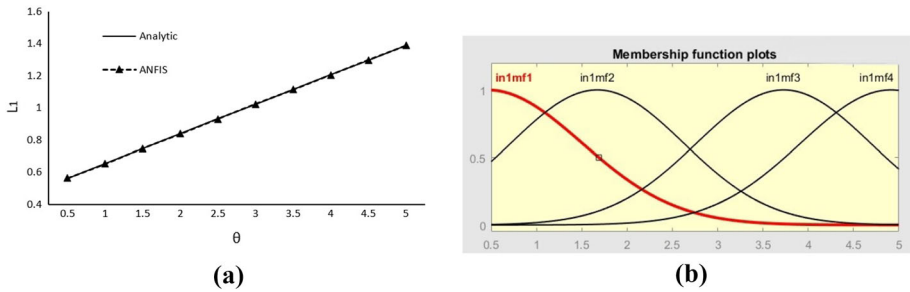
By applying this, we aim to present the effect on mean system size by changing  $\lambda$ ,  $\alpha$ ,  $s$  and  $\theta$ . It can be observed that  $L_1$  increases exponentially with rising retrials and decreases in logarithmic manner with increasing service rate. With the increase in  $\lambda$  and  $\theta$ ,  $L_1$  rises gradually in almost same manner. Thus the system designers and decision makers should make efforts and choose these parameters in such a way that their respective system becomes robust.

### 6 Cost optimization

Practically, the total operating cost of a system assumes a vital part in the investigation of many industrial systems. System architects and supervisors are typically keen on limiting the working expense per unit time to make the system more productive and profitable. More work for optimization of the cost function can be seen in Gao et al. (2020), Lan and Tang (2020), Upadhyaya and Kushwaha (2020). In this segment, the optimal outline of the retrial



**Fig. 4** **a**  $L_1$  vs  $s$  for analytic and ANFIS results; **b** membership function for  $s$



**Fig. 5** **a**  $L_1$  vs  $\theta$  for analytic and ANFIS results; **b** membership function for  $\theta$

queue including distinct failures and delayed repair is addressed. The analysis of a cycle for the system can be seen in detail in Gao et al. (2020) from where using the argument of an alternating renewal process, we get

$$E[\psi] = \frac{1}{(\lambda + \delta)Q_{0,0}}$$

where

$$Q_{0,0} = \frac{\lambda b}{\alpha} \frac{BC}{AB + \delta C \rho_1} \left( \frac{\alpha}{C} - \rho - \frac{\delta}{B} \rho_1 \right),$$

$A = \lambda b + \delta, \quad B = \alpha + \lambda b + \delta \quad \text{and} \quad C = \alpha + \lambda b.$

To exhibit the relevance of the outcomes acquired in the past conversation, we foster an expected operating cost function for the queueing model per unit time which is defined as

$$TC = C_h L_1 + C_u Q_a + C_b Q_b + C_a Q_c + C_d Q_d + C_p Q_e + C_s \frac{1}{E[\psi]}, \tag{14}$$

$$TC = C_h L_1 + C_u Q_a + C_b Q_b + C_a Q_c + C_d Q_d + C_p Q_e + C_s (\lambda + \delta) Q_{0,0}.$$

where the cost symbols corresponding to the cost function (per unit time) are as listed below:

- $C_h$  : Cost for holding each client in the system,
- $C_u$  : Cost of service provider being unoccupied,
- $C_b$  : Cost of service provider being occupied,
- $C_a$  : Cost of service provider being under repair for an active breakdown,
- $C_d$  : Cost of service provider being under delayed period,
- $C_p$  : Cost of service provider being under repair for a passive breakdown,
- $C_s$  : Set up cost for the busy cycle that's fixed.

### 6.1 Single objective optimization

Here we aim to minimize the system cost  $TC$  and try to find the optimal repair and retrial rates. Substituting the values of  $L_1, Q_a, Q_b, Q_c, Q_d, Q_e$  and  $Q_{0,0}$  in Eq. (14), we get

$$\begin{aligned}
 TC = TC(\alpha, v) = & C_h \left( \frac{2\delta\lambda b\rho_1 + \delta C((\lambda b)^2\mu_2 + 2\rho\rho_1)}{2[AB + \delta C\rho_1]} \right. \\
 & + \frac{C[2\lambda b\rho + C(\lambda b)^2\beta_2^*] + \delta[2\lambda b(\rho + \rho_1) + C((\lambda b)^2\mu_2 + 2\rho\rho_1 + (\lambda b)^2\beta_2^*)]}{2BC[\frac{\alpha}{C} - \rho - \frac{\delta}{B}\rho_1]} \\
 & + \frac{\lambda bC + \delta(C\rho_1 + \delta + \alpha + 2\lambda b)}{AB + \delta C\rho_1} \lambda b\beta_1 + \theta v_1 \frac{\lambda bC + \delta(C\rho_1 + \delta + \alpha + 2\lambda b)}{AB + \delta C\rho_1} \lambda b\beta_1 \\
 & \left. + \delta\rho_1 \frac{\delta + C(1 - \rho)}{AB + \delta C\rho_1} \right) + C_u \frac{\lambda b[\delta + C(1 - \rho)]}{AB + \delta C\rho_1} \\
 & + C_b \frac{\lambda bC + \delta(C\rho_1 + \delta + \alpha + 2\lambda b)}{AB + \delta C\rho_1} \lambda b\beta_1 + C_a\theta v_1 \frac{\lambda bC + \delta(C\rho_1 + \delta + \alpha + 2\lambda b)}{AB + \delta C\rho_1} \lambda b\beta_1 \\
 & + C_d\delta \frac{B(1 - \rho) - \lambda b\rho - \delta\rho_1}{AB + \delta C\rho_1} + C_p\delta\rho_1 \frac{\delta + C(1 - \rho)}{AB + \delta C\rho_1} \\
 & + C_s(\lambda + \delta) \frac{\lambda b}{\alpha} \frac{BC}{AB + \delta C\rho_1} \left( \frac{\alpha}{C} - \rho - \frac{\delta}{B}\rho_1 \right). \tag{15}
 \end{aligned}$$

It can be clearly seen that Eq. (15) is a nonlinear function of decision variables  $\alpha$  and  $v$  that are continuous. Thus, the minimization problem to get the minimum system cost can be described as follows:

$$(CP) \quad TC(\alpha^*, v^*) = \min TC(\alpha, v) \tag{16}$$

where  $S : \{(\alpha, v) : \alpha, v \in (0.2, 10)\} \subset R^2$ .

Clearly  $S$  is a convex feasible set. The certainty that the normal expense function is complex implies that we cannot use traditional slope based methodologies to minimize. Consequently, we utilize a heuristic calculation to manage the issue of advancement in Eq. (16). The Particle Swarm Optimization algorithm (PSO) presented by Kennedy and Eberhart (1995) is an adaptable optimization strategy that is performed to solve nonlinear objective functions. This algorithm works on the principle that the best positioned solution attract other possible values to get the optimal result in a particular search space.

For the optimization purpose the basic assumptions for critical parameters are:  $\lambda = 0.15, b = 0.25, s = 0.75, \delta = 0.2, \theta = 0.1, \mu = 5$ .

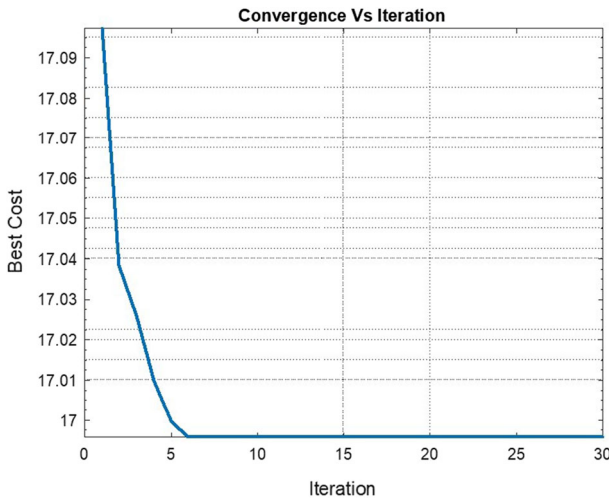
Furthermore we have considered four distinct cost sets for the evaluation of minimum cost which are given as:

- Cost Set-I:  $C_h = \$10, C_u = \$15, C_b = \$12, C_a = \$7, C_d = \$6, C_p = \$8, C_s = \$5$ .
- Cost Set-II:  $C_h = \$15, C_u = \$12, C_b = \$10, C_a = \$8, C_d = \$5, C_p = \$8, C_s = \$5$ .
- Cost Set-III:  $C_h = \$10, C_u = \$12, C_b = \$15, C_a = \$8, C_d = \$6, C_p = \$7, C_s = \$5$ .
- Cost Set-IV:  $C_h = \$15, C_u = \$10, C_b = \$12, C_a = \$10, C_d = \$5, C_p = \$10, C_s = \$5$ .

Under such conditions, we study the changing pattern of the cost function for changing values of  $\lambda$  and  $\theta$  in Table 1. Clearly the rise in the cost of the system with a rise in these parameters can be observed. Using the MATLAB code for PSO algorithm, the total optimal

**Table 1** Using PSO, the effect of  $\lambda$  and  $\theta$  is obtained on the system cost

Parameters	Cost set-I	Cost set-II	Cost set-III	Cost set-IV
$\lambda$	$TC$ for $(\alpha^*, v^*) = (0.9439, 9.9995)$			
1.5	17.0021	17.3785	17.8959	18.0127
2	20.0719	20.8804	21.2519	21.7131
2.5	22.9163	24.1443	24.3791	25.1709
3	25.5659	27.2018	27.3091	28.4183
3.5	28.0454	30.0778	30.0669	31.4810
$\theta$	$TC$ for $(\alpha^*, v^*) = (0.9439, 9.9995)$			
0.1	17.0068	17.5653	18.0795	18.2341
0.2	17.2964	17.9664	18.3972	18.6906
0.3	17.5859	18.3675	18.7148	19.1470
0.4	17.8753	18.7686	19.0325	19.6034
0.5	18.1648	19.1696	19.3501	20.0598



**Fig. 6** Convergence curve using PSO

cost of the system obtained for cost set-I is  $TC(\alpha, v) = \$17$  for  $(\alpha^*, v^*) = (0.9439, 9.9995)$ . The convergence of the objective function using PSO is shown in Fig. 6.

Now we discuss the concept of convexity of the cost function. Here using the Hessian matrix test we have shown that the objective function proposed in the problem ( $CP$ ) is convex and hence is a convex programming problem (CPP). Thus the optimal cost obtained above is a global minimum. Considering Eq. (14) as:

$$\begin{aligned}
 TC(\alpha, v) = & C_h(Q_b + Q_c + Q_e + Q_f + Q_g) + C_u Q_a + C_b Q_b \\
 & + C_a Q_c + C_d Q_d + C_p Q_e + C_s(\lambda + \delta) Q_{0,0}.
 \end{aligned}
 \tag{17}$$



The second order partial derivatives of the cost function Eq. (17) are

$$\begin{aligned} \frac{\partial^2 TC}{\partial \alpha^2} &= C_h \left( \frac{\partial^2 Q_b}{\partial \alpha^2} + \frac{\partial^2 Q_c}{\partial \alpha^2} + \frac{\partial^2 Q_e}{\partial \alpha^2} + \frac{\partial^2 Q_f}{\partial \alpha^2} + \frac{\partial^2 Q_g}{\partial \alpha^2} \right) + C_u \frac{\partial^2 Q_a}{\partial \alpha^2} \\ &\quad + C_b \frac{\partial^2 Q_b}{\partial \alpha^2} + C_a \frac{\partial^2 Q_c}{\partial \alpha^2} + C_d \frac{\partial^2 Q_d}{\partial \alpha^2} + C_p \frac{\partial^2 Q_e}{\partial \alpha^2} + C_s (\lambda + \delta) \frac{\partial^2 Q_{0,0}}{\partial \alpha^2}, \\ \frac{\partial^2 TC}{\partial \alpha \partial v} &= C_h \left( \frac{\partial^2 Q_b}{\partial \alpha \partial v} + \frac{\partial^2 Q_c}{\partial \alpha \partial v} + \frac{\partial^2 Q_e}{\partial \alpha \partial v} + \frac{\partial^2 Q_f}{\partial \alpha \partial v} + \frac{\partial^2 Q_g}{\partial \alpha \partial v} \right) + C_u \frac{\partial^2 Q_a}{\partial \alpha \partial v} \\ &\quad + C_b \frac{\partial^2 Q_b}{\partial \alpha \partial v} + C_a \frac{\partial^2 Q_c}{\partial \alpha \partial v} + C_d \frac{\partial^2 Q_d}{\partial \alpha \partial v} + C_p \frac{\partial^2 Q_e}{\partial \alpha \partial v} + C_s (\lambda + \delta) \frac{\partial^2 Q_{0,0}}{\partial \alpha \partial v}, \\ \frac{\partial^2 TC}{\partial v^2} &= C_h \left( \frac{\partial^2 Q_b}{\partial v^2} + \frac{\partial^2 Q_c}{\partial v^2} + \frac{\partial^2 Q_e}{\partial v^2} + \frac{\partial^2 Q_f}{\partial v^2} + \frac{\partial^2 Q_g}{\partial v^2} \right) + C_u \frac{\partial^2 Q_a}{\partial v^2} \\ &\quad + C_b \frac{\partial^2 Q_b}{\partial v^2} + C_a \frac{\partial^2 Q_c}{\partial v^2} + C_d \frac{\partial^2 Q_d}{\partial v^2} + C_p \frac{\partial^2 Q_e}{\partial v^2} + C_s (\lambda + \delta) \frac{\partial^2 Q_{0,0}}{\partial v^2}, \end{aligned} \tag{18}$$

where

$$\begin{aligned} \frac{\partial^2 Q_a}{\partial \alpha^2} &= \frac{2(\lambda b + \delta(1 + \rho_1))[\lambda b \delta \rho A]}{[AB + \delta C \rho_1]^3}; \quad \frac{\partial^2 Q_b}{\partial \alpha^2} = \frac{\partial^2 Q_c}{\partial \alpha^2} = 0; \\ \frac{\partial^2 Q_d}{\partial \alpha^2} &= \frac{2\delta(\lambda b + \delta(1 + \rho_1))[\delta^2(\rho + \rho_1) + \lambda b \delta \rho]}{[AB + \delta C \rho_1]^3}; \quad \frac{\partial^2 Q_e}{\partial \alpha^2} = \rho_1 \frac{\partial^2 Q_d}{\partial \alpha^2}; \\ \frac{\partial^2 Q_f}{\partial \alpha^2} &= \frac{[\lambda b + \delta(1 + \rho_1)][(\lambda b)^2 \mu_2 + 2\rho \rho_1] \delta A + 2\lambda b \delta \rho_1 [\lambda b + \delta(1 + \rho_1)]^2}{[AB + \delta C \rho_1]^3}; \\ &\quad \{ \alpha B - \rho BC - \delta C \rho_1 \}^2 (\lambda b)^2 \beta_2^* - (1 - \rho) [\alpha B - \rho BC - \delta C \rho_1] \\ &\quad \times \{ C [2\lambda b \rho + C(\lambda b)^2 \beta_2^*] + \delta [2\lambda b(\rho + \rho_1) + C((\lambda b)^2 \mu_2 + 2\rho \rho_1 + (\lambda b)^2 \beta_2^*)] \} \\ &\quad - [\alpha B - \rho BC - \delta C \rho_1] [\alpha + B - 2\rho C - \delta(\rho + \rho_1)] \left\{ \begin{array}{l} 2\lambda b \rho + 2C(\lambda b)^2 \beta_2^* \\ + \delta((\lambda b)^2 \mu_2 + 2\rho \rho_1 + (\lambda b)^2 \beta_2^*) \end{array} \right\} \\ \frac{\partial^2 Q_g}{\partial \alpha^2} &= \frac{+[\alpha + B - 2\rho C - \delta(\rho + \rho_1)]^2 \left\{ \begin{array}{l} C [2\lambda b \rho + C(\lambda b)^2 \beta_2^*] \\ + \delta [2\lambda b(\rho + \rho_1) + C((\lambda b)^2 \mu_2 + 2\rho \rho_1 + (\lambda b)^2 \beta_2^*)] \end{array} \right\}}{\left\{ BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) \right\}^3}; \\ \frac{\partial^2 Q_{0,0}}{\partial \alpha^2} &= \lambda b (\lambda + \delta) \left\{ \left( \frac{BC}{\alpha(AB + \delta C \rho_1)} \right) \left( \frac{-2\lambda b}{C^3} + \frac{-2\delta \rho_1}{B^3} \right) + \frac{Y \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right)}{(\alpha^4)(AB + \delta C \rho_1)^4} \right. \\ &\quad \left. + 2 \left( \frac{\alpha(AB + \delta C \rho_1)(2C + \delta) - BCX}{\alpha^2(AB + \delta C \rho_1)^2} \right) \left( \frac{\lambda b}{C^2} + \frac{\delta \rho_1}{B^2} \right) \right\}; \end{aligned}$$

such that

$$\begin{aligned} X &= [A(\alpha + B) + \delta(\alpha + C)\rho_1] \\ Y &= \alpha^2(AB + \delta C \rho_1)^2[(AB + \delta C \rho_1)(2\lambda b + 4\alpha + \delta) + (2\lambda b + 2\alpha + \delta)X(\alpha - 1) - BC(2A + 2\delta \rho_1)] \\ &\quad - [\alpha(AB + \delta C \rho_1)(2C + \delta) - BCX] \times [2\alpha(AB + \delta C \rho_1)^2 + 2\alpha^2(AB + \delta C \rho_1)X] \\ \frac{\partial^2 Q_a}{\partial \alpha \partial v} &= \frac{2(\lambda b)^2 \beta_1 \theta [(AB + \delta \rho_1 C) - C(\lambda b + \delta(1 + \rho))]}{v^2(AB + \delta \rho_1 C)^2}; \quad \frac{\partial^2 Q_b}{\partial \alpha \partial v} = 0; \\ \frac{\partial^2 Q_c}{\partial \alpha \partial v} &= \frac{-2\lambda b \beta_1 \theta (\lambda b + \delta(1 + \rho)) [(AB + \delta \rho_1 C) - (\lambda b C + \delta(C \rho_1 + \lambda b + B))]}{v^2(AB + \delta \rho_1 C)^2}; \\ \frac{\partial^2 Q_d}{\partial \alpha \partial v} &= \frac{2\theta \lambda b \delta \beta_1 (AB + \delta \rho_1 C) - \delta(\lambda b + \delta(1 + \rho)) \left[ \lambda b \beta_1 \left( \frac{2\rho}{v^2} \right) (B + \lambda b) \right]}{v^2(AB + \delta \rho_1 C)^2}; \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 Q_e}{\partial \alpha \partial v} &= \frac{2\lambda b \theta \beta_1 \delta \rho_1 [(AB + \delta \rho_1 C) - C(\lambda b + \delta(1 + \rho))]}{v^2 (AB + \delta \rho_1 C)^2}; \\ \frac{\partial^2 Q_{0,0}}{\partial \alpha \partial v} &= (\lambda + \delta) 2\theta (\lambda b)^2 \beta_1 \left\{ \frac{\alpha (AB + \delta \rho_1 C) (B + C) - BC[A(\alpha + B) + \delta \rho_1 (\alpha + C)]}{v^2 \alpha^2 (AB + \delta \rho_1 C)^2} \right\}; \\ \frac{\partial^2 Q_f}{\partial \alpha \partial v} &= \frac{-2\lambda b \theta \beta_1 \delta \rho_1 [(AB + \delta \rho_1 C) - C(\lambda b + \delta(1 + \rho))]}{v^2 (AB + \delta \rho_1 C)^2}; \\ &\quad (4\lambda b \theta \beta_1 BC v^2) \left\{ \begin{aligned} &2[(\alpha + B) - \rho(B + C) - \delta \rho_1] N' + 2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) \frac{\partial N'}{\partial \alpha} \\ &- (4\lambda b \theta \beta_1 (B + C) v^2) N - \frac{\partial N}{\partial \alpha} (4\lambda b \theta \beta_1 BC v^2) \end{aligned} \right\} \\ \frac{\partial^2 Q_g}{\partial \alpha \partial v} &= \frac{-2[(\alpha + B) - \rho(B + C) - \delta \rho_1] \left\{ 2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) N' - (4\lambda b \theta \beta_1 BC v^2) N \right\}}{\left\{ 2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) \right\}^3} \end{aligned}$$

such that

$$\begin{aligned} N &= C \left[ 2(\lambda b)^2 \beta_1 \left( 1 + \frac{2\theta}{v} \right) + C(\lambda b)^2 \left( \beta_2 \left( 1 + \frac{2\theta}{v} \right)^2 + \theta \beta_1 \frac{2}{v^2} \right) \right] \\ &\quad + \delta \left[ 2(\lambda b)^2 \beta_1 \left( 1 + \frac{2\theta}{v} \right) + 2\lambda b \rho_1 + C \left( \begin{aligned} &(\lambda b)^2 \mu_2 + 2(\lambda b)^2 \beta_1 \left( 1 + \frac{2\theta}{v} \right) \rho_1 \\ &+ (\lambda b)^2 \left( \beta_2 \left( 1 + \frac{2\theta}{v} \right)^2 + \theta \beta_1 \frac{2}{v^2} \right) \end{aligned} \right) \right], \\ N' &= (\lambda b)^2 BC \left[ 4\theta \beta_2 \left( 1 + \frac{2\theta}{v} \right) \left( \frac{-1}{v^2} \right) - \frac{4\theta \beta_1}{v^3} \right] + 4\lambda b \theta \beta_1 \left( \frac{-1}{v^2} \right) (\lambda b B + \delta \rho_1 C), \\ \frac{\partial N}{\partial \alpha} &= 2\lambda b \rho + 2C(\lambda b)^2 \beta_2^* + \delta((\lambda b)^2 \mu_2 + 2\rho \rho_1 + (\lambda b)^2 \beta_2^*), \\ \frac{\partial N'}{\partial \alpha} &= (\lambda b)^2 \left[ 4\theta \beta_2 \left( 1 + \frac{2\theta}{v} \right) \left( \frac{-1}{v^2} \right) - \frac{4\theta \beta_1}{v^3} \right] (B + C) + 4\lambda b \theta \beta_1 \left( \frac{-1}{v^2} \right) (\lambda b + \delta \rho_1). \\ \frac{\partial^2 Q_a}{\partial v^2} &= \frac{-4(\lambda b)^2 \beta_1 \theta C}{v^3 (AB + \delta \rho_1 C)}; \quad \frac{\partial^2 Q_b}{\partial v^2} = 0; \\ \frac{\partial^2 Q_c}{\partial v^2} &= \frac{4\lambda b \beta_1 \theta (\lambda b C + \delta(C \rho_1 + \lambda b + B))}{v^3 (AB + \delta \rho_1 C)}; \\ \frac{\partial^2 Q_d}{\partial v^2} &= \frac{-4\lambda b \theta \beta_1 \delta (\lambda b + B)}{v^3 (AB + \delta \rho_1 C)}; \quad \frac{\partial^2 Q_e}{\partial v^2} = \frac{-4\lambda b \delta \theta \beta_1 C \rho_1}{v^3 (AB + \delta \rho_1 C)}; \\ \frac{\partial^2 Q_{0,0}}{\partial v^2} &= -4\theta \lambda b \beta_1 \frac{(\lambda + \delta) \lambda b}{v^3} \frac{BC}{\alpha (AB + \delta \rho_1 C)}; \quad \frac{\partial^2 Q_f}{\partial v^2} = \frac{4\lambda b \theta \beta_1 \delta \rho_1 C}{v^3 (AB + \delta \rho_1 C)}; \\ \frac{\partial^2 Q_g}{\partial v^2} &= \frac{2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) N_1'' - N_1 4BC \theta \lambda b \beta_1 \left( \frac{-2}{v^3} \right)}{\left[ 2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) \right]^2} \\ &\quad - \frac{2 \left[ (2BC)^2 \theta \lambda b \beta_1 \left( \frac{2}{v^2} \right) \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) N_1' - N_1 \left( 2BC \theta \lambda b \beta_1 \left( \frac{2}{v^2} \right) \right)^2 \right]}{\left[ 2BC \left( \frac{\alpha}{C} - \rho - \frac{\delta \rho_1}{B} \right) \right]^3} \end{aligned}$$

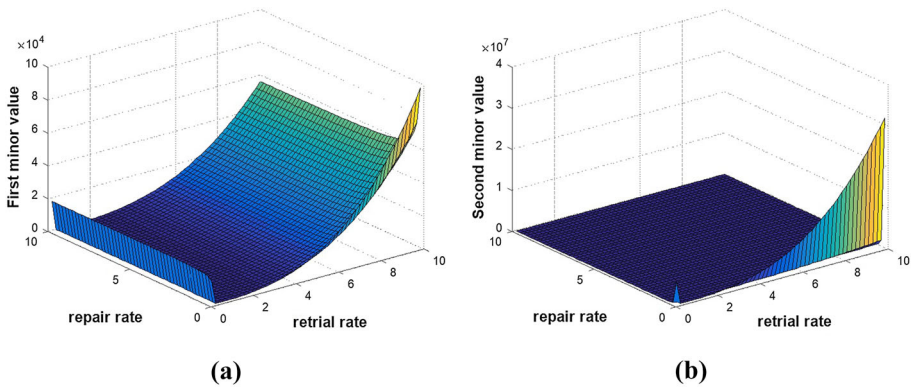


Fig. 7 Surface plot of  $F_1$  and  $G_1$  in (a) and (b) respectively

where

$$\begin{aligned}
 N_1 &= C \left[ 2(\lambda b)^2 \beta_1 \left( 1 + \frac{2\theta}{v} \right) + C(\lambda b)^2 \left( \beta_2 \left( 1 + \frac{2\theta}{v} \right)^2 + \theta \beta_1 \left( \frac{2}{v^2} \right) \right) \right] \\
 &+ \delta \left[ 2(\lambda b)^2 \beta_1 \left( 1 + \frac{2\theta}{v} \right) + 2\lambda b \rho_1 + C \left\{ \begin{aligned} &(\lambda b)^2 \mu_2 + 2\lambda b \beta_1 \rho_1 \left( 1 + \frac{2\theta}{v} \right) \\ &+ (\lambda b)^2 \left[ \beta_2 \left( 1 + \frac{2\theta}{v} \right)^2 + \theta \beta_1 \left( \frac{2}{v^2} \right) \right] \end{aligned} \right\} \right], \\
 N_1' &= (\lambda b)^2 BC \left[ 4\theta \beta_2 \left( \frac{-1}{v^2} \right) \left( 1 + \frac{2\theta}{v} \right) + 2\theta \beta_1 \left( \frac{-2}{v^3} \right) \right] \\
 &+ 4\theta (\lambda b)^2 \beta_1 B \left( \frac{-1}{v^2} \right) + 4\theta \lambda b \beta_1 \rho_1 \delta C \left( \frac{-1}{v^2} \right), \\
 N_1'' &= (\lambda b)^2 BC \left[ 4\theta \beta_2 \left( \frac{2}{v^3} \right) \left( 1 + \frac{2\theta}{v} \right) + 4\theta \beta_1 \left( \frac{3}{v^4} \right) + 8\theta^2 \beta_2 \left( \frac{-1}{v^2} \right)^2 \right] \\
 &+ 4\theta (\lambda b)^2 \beta_1 B \left( \frac{2}{v^3} \right) + 4\theta (\lambda b)^2 \beta_1 \rho_1 \delta C \left( \frac{2}{v^3} \right).
 \end{aligned}$$

Clearly all the partial derivatives of of  $TC$  exists and are continuous over the domain, thus the Hessian matrix of  $TC$  will be

$$M_1 = \begin{bmatrix} \frac{\partial^2 TC}{\partial \alpha^2} & \frac{\partial^2 TC}{\partial \alpha \partial v} \\ \frac{\partial^2 TC}{\partial v \partial \alpha} & \frac{\partial^2 TC}{\partial v^2} \end{bmatrix}.$$

Let the determinant of first and the second principle minor be denoted as

$$F_1 = \left| \frac{\partial^2 TC}{\partial \alpha^2} \right|, \quad G_1 = \left| \left( \frac{\partial^2 TC}{\partial \alpha^2} \right) \left( \frac{\partial^2 TC}{\partial v^2} \right) - \left( \frac{\partial^2 TC}{\partial \alpha \partial v} \right)^2 \right|$$

respectively. As clearly from the 3D plots (Fig. 7a, b) we can see that  $F_1 > 0$  and  $G_1 \geq 0$  so  $M_1$  is a positive semidefinite matrix. Hence  $TC$  is a convex function. Thus  $(CP)$  is a convex programming problem over  $S$ . Therefore the optimal cost  $TC(\alpha, v) = \$17$  obtained for the system is a global minima.

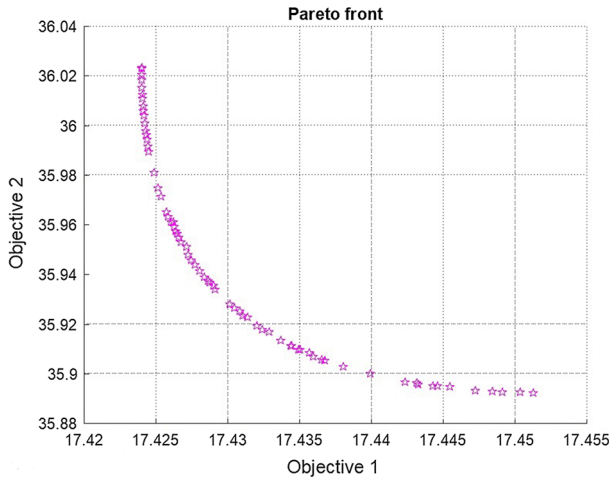


Fig. 8 Efficient solution using MOGA

### 6.2 Bi-objective optimization

Most of the optimization studies of queuing system focus on a single-objective cost optimization problems. However, for real-world problems to make the system robust more than one objective function should be considered which will be conflicting in nature. Thus multi-objective programming problem is of great importance. In this sub-segment, we formulate a bi-objective problem of optimization where the expected cost function ( $TC$ ) and the expected awaiting time in queue ( $W_1$ ) are to be minimized simultaneously over a domain  $S$ . Consider the following bi-objective problem of optimization

$$(BP) \quad (TC^*, W_1^*) = \min(TC, W_1) \tag{19}$$

where  $S : \{(\alpha, v) : \alpha, v \in (0.2, 10)\} \subset R^2$ .

In the problem  $(BP)$  there are two objective functions and thus we try to find the minimum cost to be invested with minimum awaiting time of the clients over the feasible region which is convex.

To attain the efficient solution of the problem  $(BP)$ , the Multi-Objective Genetic Algorithm (MOGA) has been executed. It is to be noted from Pareto frontier that the Pareto optimal solution cannot improve both the objective functions simultaneously. Hence the concept of an efficient solution is studied. Figure 8 represents non-controlling solutions obtained using MOGA where it can be seen that for the Pareto rates  $(\alpha^*, v^*) = (0.9066, 9.9949)$ , the minimum cost and minimum client’s awaiting time are  $TC^* = \$17.4$  and  $W_1^* = 35.9$  mins, respectively. Clearly, the minimum cost obtained for the bi-objective optimization problem  $(BP)$  has slightly increased from the minimum cost obtained for  $(CP)$  but as the minimum awaiting time of the clients is also obtained so, it will be very useful to develop a robust system.

Table 2 highlights the changing pattern of the cost function and the halting time in the queue, simultaneously for changing values of  $\lambda$  and  $\theta$ . It can be seen that as we increase the parametric values, the cost of the system increases while the halting time decreases which is in accordance with the Little’s formula and relates to the realistic scenario.

**Table 2** Using MOGA, the effect of  $\lambda$  and  $\theta$  is obtained on system cost and waiting time

Parameters	Cost set-I	Cost set-II	Cost set-III	Cost set-IV
$\lambda$	$(TC, W_1)$ for $(\alpha^*, v^*) = (0.9066, 9.9949)$			
1.5	(17.42, 35.94)	(18.24, 53.87)	(18.18, 35.97)	(19.04, 53.91)
2	(20.34, 34.12)	(21.67, 50.98)	(21.34, 34.14)	(22.70, 50.97)
2.5	(22.97, 32.46)	(24.76, 48.54)	(24.26, 32.34)	(25.99, 48.50)
3	(25.42, 30.90)	(27.55, 46.38)	(26.88, 30.90)	(29.03, 46.35)
$\theta$	$(TC, W_1)$ for $(\alpha^*, v^*) = (0.9066, 9.9949)$			
0.1	(17.43, 35.94)	(18.43, 54.75)	(18.32, 36.52)	(19.24, 54.72)
0.2	(17.83, 37.56)	(18.79, 56.42)	(18.59, 37.69)	(19.64, 56.36)
0.3	(18.07, 38.71)	(19.16, 58.01)	(18.86, 38.81)	(20.03, 58.07)
0.4	(18.34, 39.81)	(19.51, 59.76)	(19.14, 39.95)	(20.44, 59.68)

To ensure that the minimum value obtained for  $(BP)$  is global Pareto optimal value, we need to check if the minimization problem  $(BP)$  is convex or not for which, both the objective functions need to be shown convex. As discussed in sub-segment. 6.1, the proposed objective  $TC$  is convex. To show the other proposed objective  $W_1$  convex, rewriting Eq. (13) we get

$$W_1(\alpha, v) = \frac{C_h}{\lambda} (Q_f + Q_g + Q_b + Q_c + Q_e). \tag{20}$$

The second order partial derivatives of Eq. (20) will be

$$\begin{aligned} \frac{\partial^2 W_1}{\partial \alpha^2} &= \frac{C_h}{\lambda} \left( \frac{\partial^2 Q_f}{\partial \alpha^2} + \frac{\partial^2 Q_g}{\partial \alpha^2} + \frac{\partial^2 Q_b}{\partial \alpha^2} + \frac{\partial^2 Q_c}{\partial \alpha^2} + \frac{\partial^2 Q_e}{\partial \alpha^2} \right); \\ \frac{\partial^2 W_1}{\partial \alpha \partial v} &= \frac{C_h}{\lambda} \left( \frac{\partial^2 Q_f}{\partial \alpha \partial v} + \frac{\partial^2 Q_g}{\partial \alpha \partial v} + \frac{\partial^2 Q_b}{\partial \alpha \partial v} + \frac{\partial^2 Q_c}{\partial \alpha \partial v} + \frac{\partial^2 Q_e}{\partial \alpha \partial v} \right); \\ \frac{\partial^2 W_1}{\partial v^2} &= \frac{C_h}{\lambda} \left( \frac{\partial^2 Q_f}{\partial v^2} + \frac{\partial^2 Q_g}{\partial v^2} + \frac{\partial^2 Q_b}{\partial v^2} + \frac{\partial^2 Q_c}{\partial v^2} + \frac{\partial^2 Q_e}{\partial v^2} \right); \end{aligned} \tag{21}$$

where  $Q_f, Q_g, Q_b, Q_c, Q_e$  are as defined in sub-segment. 6.1. Clearly the second order partial derivatives of  $W_1$  exists and are continuous over the domain, thus the Hessian matrix is

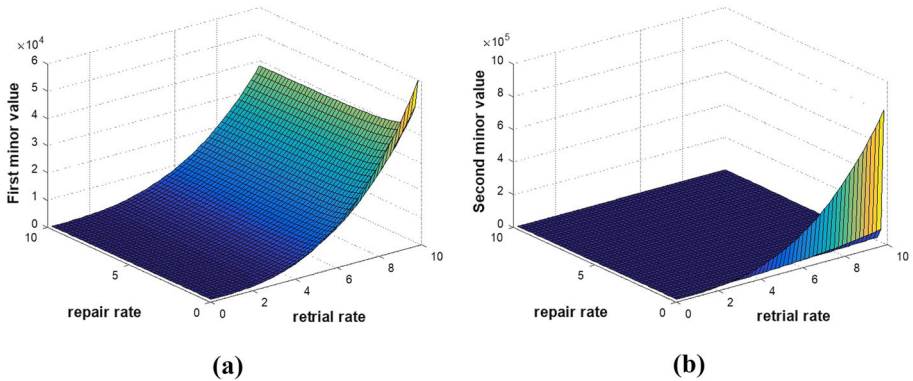
$$M_2 = \begin{bmatrix} \frac{\partial^2 W_1}{\partial \alpha^2} & \frac{\partial^2 W_1}{\partial \alpha \partial v} \\ \frac{\partial^2 W_1}{\partial v \partial \alpha} & \frac{\partial^2 W_1}{\partial v^2} \end{bmatrix}.$$

Consider the determinant of the first principal minor as

$$F_2 = \left| \frac{\partial^2 W_1}{\partial \alpha^2} \right|$$

and the determinant of second principal minor as

$$G_2 = \left| \left( \frac{\partial^2 W_1}{\partial \alpha^2} \right) \left( \frac{\partial^2 W_1}{\partial v^2} \right) - \left( \frac{\partial^2 W_1}{\partial \alpha \partial v} \right)^2 \right|.$$



**Fig. 9** Surface plot of first and second minor of  $W_1$  in (a) and (b) respectively

Figure 9a, b shows that  $F_2 > 0$  and  $G_2 \geq 0$ . So  $M_2$  is a positive semidefinite matrix thus  $W_1$  is also a convex function. Therefore (BP) is a convex programming problem. Thus it can be deduced that the Pareto optimal values  $\mathbf{TC}^* = \$17.4$ ,  $\mathbf{W}_1^* = 35.9$  are global optimal values of a bi-objective optimization problem (BP).

It may be noted that in the optimization problems (CP) and (BP), the interval (0.2, 10) ensures the convexity of the objective functions and thus the problems are convex programming problems. If the lower bound or upper bound is changed, the Hessian matrix of the objective functions remain no longer positive semidefinite hence the convexity of the objective functions is disturbed which is in fact an important sufficiency condition.

## 7 Conclusion

In tackling various real life computing situations easily, cost-effective systems are made for which the optimal values of critical parameters of the system might provide an insight to network system designers, system operators and software system engineers. Application of the developed model is not only limited to mentioned in the paper rather it can be seen in call centres, mobile and communication workplaces or any organization using commuting systems, where erratic queueing system with retrial can play a crucial role in conveying quality help rapidly under decreased expense. The model  $M/G/1$  having erratic server with delayed repair and balking could be adopted by computing managers in getting minimal cost with minimized awaiting time. The presumption of resisting behaviour of clients make the model flexible to work in real life. In real economical world, our model is reliable and more helpful as the minimum cost value obtained is not only found using heuristic technique but also the concept of convexity is applied to show the existence of global minima. In future, the model can be studied further by considering different types of vacations for the server. We can extend the paper in future by applying priority concept with multi optional services or one can think of using different control policies i.e. N-policy, F-policy on this model.

**Acknowledgements** The authors of the paper are thankful for the insightful and valuable comments of the anonymous referees as their suggestions helped in shaping our paper for the better.

## References

- Bazarara, M., Sherali, H., & Shetty, C. M. (2005). *Nonlinear Programming—Theory and Algorithms* (3rd ed.). Wiley.
- Choudhury, G., & Tadj, L. (2009). An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair. *Applied Mathematical Modelling*, 33(6), 2699–2709. <https://doi.org/10.1016/j.apm.2008.08.006>
- Dragieva, V. I., & Phung-Duc, T. (2020). A finite-source M/G/1 retrial queue with outgoing calls. *Annals of Operations Research*, 293, 101–121. <https://doi.org/10.1007/s10479-019-03359-z>
- Gao, S., Zhang, J., & Wang, X. (2020). Analysis of a retrial queue with two-type breakdowns and delayed repairs. *IEEE Access*, 8, 172428–172442. <https://doi.org/10.1109/ACCESS.2020.3023191>
- Gomes Corral, A. (1999). Stochastic analysis of a single server retrial queue with general retrial times. *Naval Research Logistics*, 46(5), 561–581.
- Jin-ting, W. (2006). Reliability analysis of M/G/1 queues with general retrial times and server breakdowns. *Progress in Natural Science: Materials International*, 16(5), 464–473. <https://doi.org/10.1080/10020070612330021>
- Kahag, M. R., Niaki, S. T. A., Seifbarghy, M., & Zabihi, S. (2019). Bi-objective optimization of multi-server intermodal hub-location allocation problem in congested systems: modeling and solution. *Journal of Industrial Engineering International*, 15(2), 221–248. <https://doi.org/10.1007/s40092-018-0288-0>
- Kennedy, J., & Eberhart R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE, Perth, Australia.
- Kim, J., & Kim, B. (2016). A survey of retrial queueing systems. *Annals of Operations Research*, 247, 3–36. <https://doi.org/10.1007/s10479-015-2038-7>
- Lan, S., & Tang, Y. (2020). An unreliable discrete-time retrial queue with probabilistic preemptive priority, balking customers and replacements of repair times. *AIMS Mathematics*, 5(5), 4322–4344. <https://doi.org/10.3934/math.2020276>
- Lee, S. W., Kim, B., & Kim, J. (2020). Analysis of the waiting time distribution in M/G/1 retrial queues with two way communication. *Annals of Operations Research*, 310, 505–518. <https://doi.org/10.1007/s10479-020-03717-2>
- Miettinen, K. (1998). *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers.
- Mohammadi, M., Dehbari, S., & Vahdani, B. (2014). Design of a bi-objective reliable healthcare network with finite capacity queue under service covering uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, 72, 15–41. <https://doi.org/10.1016/j.tre.2014.10.001>
- Pourmohammadi, P., Tavakkoli-Moghaddam, R., Rahimi, Y., & Triki, C. (2021). Solving a hub location-routing problem with a queue system under social responsibility by a fuzzy meta-heuristic algorithm. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-021-04299-3>
- Saggou, H., Lachemot, T., & Ourbih-Tari, M. (2017). Performance measures of M/G/1 retrial queues with recurrent customers, breakdowns, and general delays. *Communications in Statistics - Theory and Methods*, 46(16), 7998–8015. <https://doi.org/10.1080/03610926.2016.1171352>
- Sherman, N. P., Kharoufeh, J. P., & Abramson, M. P. (2009). An M/G/1 retrial queue with unreliable server for streaming multimedia applications. *Probability in the Engineering and Informational Sciences*, 23(2), 281–304. <https://doi.org/10.1017/S0269964809000175>
- Taleb, S., & Aissani, A. (2016). Preventive maintenance in an unreliable M/G/1 retrial queue with persistent and impatient customers. *Annals of Operations Research*, 247(1), 291–317. <https://doi.org/10.1007/s10479-016-2217-1>
- Tavakkoli-Moghaddam, R., Vazifeh-Noshafagh, S., Taleizadeh, A. A., Hajipour, V., & Mahmoudi, A. (2017). Pricing and location decisions in multi-objective facility location problem with M/M/m/k queueing systems. *Engineering Optimization*, 49(1), 136–160. <https://doi.org/10.1080/0305215X.2016.1163630>
- Upadhyaya, S. (2014). Performance analysis of a batch arrival retrial queue with Bernoulli feedback. *International Journal of Mathematics in Operational Research*, 6(6), 680–703. <https://doi.org/10.1504/IJMOR.2014.065423>
- Upadhyaya, S., & Kushwaha, C. (2020). Performance prediction and ANFIS computing for unreliable retrial queue with delayed repair under modified vacation policy. *International Journal of Mathematics in Operational Research*, 17(4), 437–466. <https://doi.org/10.1504/IJMOR.2020.110843>
- Verma, R. K. (1986). Multiobjective optimization of a queueing system. *Optimization*, 17(1), 103–115. <https://doi.org/10.1080/02331938608843107>
- Wu, C. H., & Yang, D. Y. (2021). Bi-objective optimization of a queueing model with two-phase heterogeneous service. *Computers & Operations Research*. <https://doi.org/10.1016/j.cor.2021.105230>
- Zhang, Z. G. (2006). On the convexity of the two-threshold policy for an M/G/1 queue with vacations. *Operations Research Letters*, 34(4), 473–476. <https://doi.org/10.1016/j.orl.2005.07.002>

Zhang, Z. G., Vickson, & Eenige, M. J. A. (1997). Optimal two-threshold policies in an M/G/1 queue with two vacation types. *Performance Evaluation*, 29(1), 63–80. [https://doi.org/10.1016/S0166-5316\(96\)00005-3](https://doi.org/10.1016/S0166-5316(96)00005-3)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.