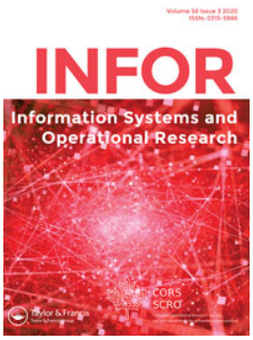

INFOR: Information Systems and Operational Research

ISSN: 0315-5986 (Print) 1916-0615 (Online) Journal homepage: https://www.tandfonline.com/loi/tinf20

Balancing herding and congestion in service systems: a queueing perspective

Hao Zhang, Qi-Ming He & Xiaobo Zhao

To cite this article: Hao Zhang, Qi-Ming He & Xiaobo Zhao (2020) Balancing herding and congestion in service systems: a queueing perspective, INFOR: Information Systems and Operational Research, 58:3, 511-536, DOI: 10.1080/03155986.2020.1734902

To link to this article: https://doi.org/10.1080/03155986.2020.1734902

Published online: 16 Mar 2020.

INFOR: INFORMATION SYSTEMS AND OPERATIONAL RESEARCH
2020, VOL. 58, NO. 3, 511–536
https://doi.org/10.1080/03155986.2020.1734902

Balancing herding and congestion in service systems: a queueing perspective

Hao Zhang[a], Qi-Ming He[b] and Xiaobo Zhao[c]

[a]School of Civil Engineering, Wuhan University, Wuhan, China; [b]Department of Management Sciences, University of Waterloo, Waterloo, Canada; [c]Department of Industrial Engineering, Tsinghua University, Beijing, China

ABSTRACT

In service industries such as restaurants and tourism, empirical findings show that uninformed customers may consider queues as a signal of service quality and choose to join a longer queue. Service managers become aware of this phenomenon and stimulate customer purchase by maintaining a queue. In this paper, we explore issues related to the balance between herding and congestion for service systems using a state-dependent queue. In our model, the herding effect is represented by system idle probability (as opposed to system busy probability) and the congestion is represented by a non-decreasing function of queue length. An optimization problem with the objective of minimizing the long-run average cost and constraints on traffic intensities is formulated, and the structure of its optimal solution is characterized. Further, we find closed-form solutions of the optimal state-dependent traffic intensity and the optimal service rate switching state, and characterize the relationship between the optimal solution and system parameters. Through a series of propositions and numerical examples, we gain insight into the balance between stimulation of herding effect and reduction of customer waiting, and propose that service managers should intentionally slow down when the queue is short and operate at their full speed when the queue is long.

ARTICLE HISTORY
Received 9 April 2018
Accepted 12 December 2019

KEYWORDS
Queue; state-dependent; optimization; herding; waiting time

1. Introduction

For the past years, Service Science as a burgeoning discipline has attracted the attention of scholars and practitioners. It may appear, then, that no stone in the service-management garden has been left unturned, not to mention analyzed, polished and replaced (Chase and Dasu 2001). However, there is one thing carved on the stone that is unturned: customers are not simply consumers of the service but can also be an integral part of production (Frei 2008).

The main function of a service system is to deliver services to customers (Chan and Gao 2013). Duration of service delivery is a primary measure for service quality

CONTACT Qi-Ming He q7he@uwaterloo.ca Department of Management Sciences, University of Waterloo, Waterloo, N2L 3G1, Canada

and a long queue could be a nuisance. This is true for food industries and also true for almost all service systems if we simply consider customers as service receiver. However, temporal aspects of service encounters are subtle. Maglio, Kieliszewski, and Spohrer (2010) wrote "*if OM sees its main function as moving customers in and out of the service facility as quickly as possible then the quality of the service delivered from a customer's standpoint may suffer*." (Note: OM stands for operations management.) Large queues can be interpreted as a proxy for higher quality (Bitran, Ferrer, and Rocha e Oliveira 2008). From a manager's standpoint, queues can be used for product branding. For example, the New York Times and the Los Angeles printed articles about the lines outside a nightclub Studio 54 stretching around the block, which only served to make them longer. More examples come from food industries, such as the line outside Keizo Shimamoto the Brooklyn ramen burger booth in the summer of 2013 and the Manhattan Dominique Ansel Bakery in January 2014 (Mordfin 2014). Xishaoye, a fast-food chain founded by Chinese young entrepreneurs, strategically applied this tactic to sell their burgers (Zhao 2014).

In service systems such as restaurants, a long queue usually implies better service quality and hence stimulates herding effect on uninformed customers, or empty restaurant syndrome (see Debo and Veeraraghavan (2009); Kremer and Debo (2012); Veeraraghavan and Debo (2011)). Lu (2013) suggested that a long queue might induce more customers to join the queue. Raz and Ert (2008) reported that queue length has impact on consumer's choice among restaurants, even those local ones that may be familiar to customers. In restaurants, service time is prolonged to attract more customers. When there are few customers, it is a good time to increase revenue: under low workload waiters and waitress increase service time to make cross-selling and up-selling attempts (see Aksin, Armony, and Mehrotra (2007); Tan and Netessine (2014)). This phenomenon is typical in industries where queues can be a signal of service quality, such as in restaurant industry and tourism (Hernandez-Maskivker, Ryan, and del Mar PAmies Pallis ́e 2012).

On the other hand, waiting in a congested environment can have a negative impact on customer experiences. As a result, there is a trade-off in service systems, where herding and congestion allude to contradictory ideas in operational strategy. To understand the dilemma, we explore the issue from a queueing perspective by optimizing an objective function under a certain cost structure, which incorporates waiting cost induced by congestion and idle cost induced by herding. Maglio, Kieliszewski, and Spohrer (2010) addressed that "*Note that some resources may incur a cost only when they are actively in use by a service process, but others may contribute to the cost of the SDS even when idle*." (Note: SDS stands for service delivery system.)

To thoroughly discuss the above issue, we analyze a related optimization problem based on a queueing model with state-dependent control policy. Here, the state-dependent control policy is associated with the arrival rate and/or the service that can be adjusted according to information about the queue length. In this way, we try to reach the objective of balancing herding and congestion. We show that it is usually necessary to adjust the state-dependent traffic intensity in order to minimize the long-run average cost. It is also shown that even if we are provided with the state-dependent traffic intensity at many possible levels, we only need two levels of traffic

intensity: low traffic intensity (low arrival rate or high service rate) and high traffic intensity (high arrival rate or low service rate) to achieve the minimum costs. Findings in this paper are consistent with the phenomena in reality where "herding" in restaurants is present.

The remainder of this paper is organized as follows. In Section 2, a literature review on the study of related queueing models is conducted. The problem of interest is introduced in Section 3, which includes a queueing model, costs involved, and an optimization problem for which the structure of the optimal solution is characterized. In Section 4, we analyze two optimization problems for the selection of (state-dependent) traffic intensity and the switching state, and for the impact of system parameters on the optimal solution, respectively. Section 5 presents a number of numerical examples to gain insight into the problem of interest. Section 6 concludes the paper. Proofs of all propositions and theorems are collected in the Appendix.

2. Literature review

Studies on how a service system with state-dependent productivity (or service rate in a queueing system) is controlled, operated and performed are versatile and interdisciplinary. Our study is closely related to three streams of literature: 1) optimal control and design in queueing models; 2) cost calibration; and 3) herding behavior and state-dependent productivity in behavioral operations regime.

The literature on the control and design of a queueing system extensively explores various combinations of cost structure and objective functions and related optimal policies (for surveys one can refer to Crabill, Gross, and Magazine (1977); Sobel (1974); Stidham (1974); Stidham and Weber (1993)). Three major types of costs are usually considered: customer waiting cost, server operating cost, and service rate switching cost. The trade-off among them is the main issue to address both in the design and control of queueing systems. In the area of optimal design of queues, for example, Grassmann, Chen, and Kashyap (2001) suggested to adjust the service rate to optimize the waiting cost and operating cost in an $M/G/1$ system with a state-dependent arrival rate. Batta, Berman, and Wang (2007) addressed the trade-off between staffing cost and switching cost. More literature can be found in the area of optimal control of queues. For example, Yadin and Naor (1967) considered an $M/M/1$ system with variable service rates. They studied the joint distribution of phase and queue length induced by a hysteretic state-dependent policy without assumption of any cost function. Lippman (1975) introduced the optimal control of the service rate in an $M/M/1$ queue, where cost is incurred for the used proportion of the potential service rate and showed that the optimal service rate is increasing in the queue length. Later on, numerous studies such as Weber and Stidham (1987), Stidham and Weber (1989), and George and Harrison (2001) considered a cost function with non-decreasing holding cost and proposed monotone policies.

Moreover, there is another line of optimal control problems focusing on the policy of adjusting the number of servers in a multi-server queueing system. Jain (2005) considered an $M/M/r/K$ queue, where a server is always open while extra servers become open only when the queue length exceeds certain thresholds. Two types of

costs (i.e., waiting cost and operating cost) are considered. The paper develops a method for computing the long-run average cost for the system with a given policy and some performance measures. The paper provides a set of inequalities that optimal thresholds have to satisfy and demonstrates them with numerical examples. In general, none of the existing work explicitly explores the trade-off between the system idle cost and customer waiting cost. We would like to point out that idle cost can arguably be considered as special cases of operating cost or waiting cost. However, these costs were generally assumed to be non-decreasing with service rate or queue-length (see Crabill (1972); Sabeti (1970); Lippman (1975)). As a result, the cost structure of our objective function is different from models in the existing literature.

The idea of introducing the balance of idleness and waiting is originated from scheduling and planning in healthcare problems such as outpatient scheduling, in which the reduction of idle cost and waiting cost is the core issue. For an introduction of the cost structure in outpatient scheduling, one can refer to Cayirli and Veral (2003) and Weiss (1990). Our proposed objective function is scarce in queueing models but is analogous to those in scheduling. On the other hand, the literature on healthcare provides guidance on how to calibrate cost parameters. As Fries and Marathe (1981) pointed out, it is easier to estimate the costs relative to the server, which are usually available via standard cost accounting. Keller and Laughhunn (1973) divided the annual salary of a doctor by the hours worked per year to estimate patient waiting cost and used the minimum wage to reflect the opportunity cost of the patient waiting time. Idle cost includes not only the cost of the idle doctor, but also the cost of the idle facility (Yang, Lau, and Quek 1998). Same estimation of costs can also be found in Gupta, Zoreda, and Kramer (1971). We note that there have been experimental studies focusing on idle time and waiting time (Fetter and Thompson 1966).

Literature in behavioral operations management regime (not related to queueing model) shows that people have behaviors in contradiction with the classic non-decreasing assumption. Schultz et al. (1998) first considered the issue of the state-dependent productivity and provided a detailed review (Delasay et al. 2014). Kc and Terwiesch (2009) performed a rigorous econometric analysis and managerially considered the increase in the pressure of hospitals to operate at very high levels of utilization. From the perspective of psychology, Hsee, Yang, and Wang (2010) suggested idleness aversion behavior. Parkinson (1955) indicated that work expands so as to fill the time available for its completion. The literature provides evidence to support that people try to avoid idleness or low queue length for some reasons. Our findings contribute to this line of literature of behavioral effects on productivity by providing another explanation for why workers and/or organizations intentionally slow down and avoid idleness.

The study on the herding effect is rich in economics literature. Herding behavior in queues is generally explored from an information externality perspective (see Debo and Veeraraghavan (2009); Veeraraghavan and Debo (2011)). Kremer and Debo (2012) used a laboratory experiment to test theoretical results. Becker (1991) observed that a popular seafood restaurant in Palo Alto had a long queue while another restaurant across the street did not. Hernandez-Maskivker, Ryan, and del Mar PAmies

Pallis´e (2012) gave an extensive survey on herding behavior in tourism industries. In our study, we measure the cost of herding indirectly by using the system idle cost. In this way, we emphasize the effect of idleness (as opposed to herding) on the design and control of such queueing systems.

As demonstrated above, the literature on queueing control and analysis is mixed with papers focusing on methods that can be used in practice and with papers focusing on gaining insight into systems of interest. This paper focuses on the characterization of the optimal policy and the intrinsic relationship among system parameters and solutions. The results can be pragmatically useful and provide a simple rule of thumb to managers of service systems if they have calibrated cost parameters. We would like to point out that our study is devoted to "prescribe" an optimal policy for a state-dependent queueing system, while in queueing literature there are numerous stochastic models developed to "describe" the properties and performances of the queueing system, which are therefore not surveyed in this paper.

3. Problem formulation and a main result

We consider a state-dependent $M/M/1$ queue. The queueing model has a single queue and a single server. Customers are served on a first-come-first-served basis. Customer arrival rate and server service rate depend on the number of customers in the system (i.e., the queue length). Let λ_n be the arrival rate and μ_n the service rate, if the queue length is n.

Let $q(t)$ be the queue length at time t. If $q(t) = n \geq 0$, the time to the next arrival, if the queue length remains at n, has an exponential distribution with parameter λ_n; and, if $q(t) = n \geq 1$, the time to the next service completion, if the queue length remains at n, has an exponential distribution with parameter μ_n. It is easy to see that $\{q(t),\ t \geq 0\}$ is a continuous time Markov chain with infinitesimal generator

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

It is well-known that the steady state distribution of $\{q(t),\ t \geq 0\}$, if it exists, is given by

$$\begin{aligned} &\pi_n = \pi_0 \rho_1 \rho_2 \cdots \rho_n, \quad n \geq 1; \\ &\pi_0 = (1 + \rho_1 + \rho_1\rho_2 + \ldots + \rho_1\rho_2 \cdots \rho_n + \ldots)^{-1}, \end{aligned} \tag{2}$$

where $\rho_n = \lambda_{n-1}/\mu_n$, for $n = 1, 2, \ldots$, which shall be called the (state-dependent) *traffic intensities* for server utilization. From now on, we represent the model and present results in terms of $\{\rho_n,\ n = 1, 2, \ldots\}$, instead of $\{\lambda_n,\ n = 0, 1, 2, \ldots\}$ and $\{\mu_n,\ n = 1, 2, \ldots\}$. Interpretation of results in terms of $\{\lambda_n,\ n = 0, 1, 2, \ldots\}$ and $\{\mu_n,\ n = 1, 2, \ldots\}$ is given from time to time, though.

In order to study the utilization-waiting conundrum, we introduce two types of costs:

i. system idle cost C_I per unit idle time; and
ii. customer waiting cost $C_w(n)$ per unit time, where n is the queue length.

The long-run average cost (i.e., the expected total cost per unit time) is given by

$$TC(\rho_n, n \geq 1) = C_I\pi_0 + E\big[C_w\big(q(t)\big)\big] = C_I\pi_0 + \sum_{n=1}^{\infty} C_w(n)\pi_n. \tag{3}$$

For given $\{C_I, C_w(.)\}$, we want to find $\{\rho_1, \rho_2, \ldots, \rho_n, \ldots\}$ (or $\{\lambda_0, \lambda_1, \ldots, \lambda_n, \ldots\}$ and $\{\mu_1, \mu_2, \ldots, \mu_n, \ldots\}$) to minimize the long-run average cost, under certain constraints on $\{\rho_1, \rho_2, \ldots, \rho_n, \ldots\}$. Based on the commonly used queueing control schemes (e.g., Conway and Maxwell (1962) and Jain (2005)), we impose constraints on $\{\rho_1, \rho_2, \ldots, \rho_n, \ldots\}$. That is, we assume that the traffic intensities $\{\rho_n, n = 1, 2, \ldots\}$ would vary in queue length but become a constant after the queue length is larger than a threshold. As a result, there are a finite number of choices for traffic intensities, denoted by $\{\gamma_1, \ldots, \gamma_K\}$, where $K \geq 1$ is a finite integer. We denote the queue lengths at which the traffic intensity is switched by $\{q_1, \ldots, q_{K-1}\}$. As the traffic intensity hardly reaches zero, we assume that there is a lower bound of traffic intensities $\gamma_{\min}$, i.e., $\min\{\gamma_1, \ldots, \gamma_K\} \geq \gamma_{\min} \geq 0$. We aim to find the optimal values of $\{q_1, \ldots, q_{K-1}\}$ and $\{\gamma_1, \ldots, \gamma_K\}$ to minimize the long-run average cost given in formula (3). The optimization problem can be formulated as follows:

$$\begin{aligned} & TC^* = \inf_{\{(\gamma_1,\ldots,\gamma_K),(q_1,\ldots,q_{K-1})\}}\{TC(\rho_n, n \geq 1)\} \\ s.t.: \; & \rho_n = \gamma_1, \quad \text{for} \quad 1 \leq n \leq q_1; \\ & \rho_n = \gamma_k, \quad \text{for} \; q_{k-1} < n \leq q_k, \quad k = 2, \ldots, K-1; \\ & \rho_n = \gamma_K, \quad \text{for} \; q_{K-1} < n < \infty; \\ & \min\{\gamma_1, \gamma_2, \ldots, \gamma_K\} \geq \gamma_{\min}. \end{aligned} \tag{4}$$

The above optimization problem is a mixed integer programming problem, which can be solved numerically (Jünger et al. (2009)). In this paper, we use the optimization problem (4) to explore the relationship between server utilization and customer waiting and to gain insight into the balance between them. For that purpose, we first find the structural properties of the optimal solution.

Theorem 1. *Assume that $C_w(n)$ is a nonnegative and non-decreasing function of n. For given finite positive integer K, the optimal solution of (4) has the structure $q_2 = q_3 = \ldots = q_{K-1} = \infty$ and $\gamma_2 = \gamma_3 = \ldots = \gamma_K = \gamma_{\min}$, i.e., $\rho_n = \gamma_1$, for $1 \leq n \leq q_1$, and $\rho_n = \gamma_{\min}$, for $n > q_1$.* □

Theorem 1 implies that the optimal solution of (4) is determined by four parameters: K (= 1 or 2), γ_1, γ_2, and q_1. The traffic intensity γ_1 is chosen properly to keep the idle cost small (or to keep a proper queue length) and $\gamma_2 = \gamma_{\min}$ is required to ensure that the queue length is not long. Intuitively, the trade-off between the idle

cost and waiting cost can be achieved by having a short queue. Thus, ideally, the queue should be shorter than or equal to q_1. Hence, if the queue length is greater than q_1, we should set the traffic intensity as small as possible. The results provides insight into the relationship between customer herding and server utilization. While it is important to reduce the queue length to keep the waiting cost small, it is equally important to maintain a proper traffic intensity (i.e., herding).

To end this section, we would like to point out that Theorem 1 can be applied to multi-server optimal design problem for some special cases. For instance, set $C_I = 0$ and $C_w(n)$ be a piece-wise increasing linear function of n where the jump points are optimal thresholds at which service rates increase, then the cost structure and objective function are equivalent to those of the multi-server model with infinite capacity in Jain (2005).

4. Optimal traffic intensities and switching times

Based on the structure of the optimal solution of the optimization problem (4), in this section, we explore further properties related to the optimal policy and gain insight into the balance between idle time and waiting time. For that purpose, we choose the linear waiting cost function $C_w(n) = nC_w$, which is a typical in the literature. Note that C_w is the waiting cost per customer per unit time. We consider three cases: i) $K = 1$; ii) changing the traffic intensities while keeping switching time constant (Propositions 1, 2, and 3); and iii) changing the switching time while keeping the traffic intensities constant (Propositions 4 and 5). For all cases, the policies under consideration have the same structure as that of the optimal policy characterized in Theorem 1.

First, we consider the case with $K = 1$. For this case, the state-dependent $M/M/1$ queue is reduced to the classical $M/M/1$ queue (see Cohen (2012)), i.e., $\rho_n = \rho$, for $n = 1, 2, \ldots$. We assume $\gamma_{\min} \leq \rho < 1$. The long-run average cost is

$$TC(\rho) = C_I(1-\rho) + C_w\frac{\rho}{1-\rho}. \tag{5}$$

It is easy to show that the function $TC(\rho)$ is convex in ρ and the optimal solution is given by

$$\rho^*_{K=1} = \max\left\{\gamma_{\min}, 1-\sqrt{\frac{C_w}{C_I}}\right\}.$$

$$TC^*_{K=1} = TC\left(\rho^*_{K=1}\right) = \begin{cases} C_I(1-\gamma_{\min}) + C_w\dfrac{\gamma_{\min}}{1-\gamma_{\min}}, & \text{if } 1-\sqrt{\dfrac{C_w}{C_I}} < \gamma_{\min}; \\ 2\sqrt{C_I C_w} - C_w, & \text{if } 1-\sqrt{\dfrac{C_w}{C_I}} \geq \gamma_{\min}. \end{cases} \tag{6}$$

Equation (6) implies that if $C_I(1-\gamma_{\min})^2 \leq C_w$, the traffic intensity should be as small as possible (i.e., $\gamma_{\min}$), which can be achieved by reducing the arrival rate or

increasing the service rate. This is intuitive since, under the condition, we would like to keep the queue as small as possible. If $C_I(1-\gamma_{\min})^2 > C_w$, a proper traffic intensity should be chosen by equation (6).

For the case with $K=2$, we first keep the switching time q constant (i.e., $q=q_1$). Let $TC_q(\gamma)=TC(\rho_n=\gamma$, for $1\le n\le q$; $\rho_n=\gamma_{\min}$, for $n>q)$. By routine calculations, $TC_q(\gamma)$ can be written explicitly as

$$TC_q(\gamma)=\frac{C_I+\left(\gamma+2\gamma^2+\ldots+(q-1)\gamma^{q-1}+q\gamma^q\frac{1}{1-\gamma_{\min}}+\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})^2}\right)C_W}{1+\gamma+\gamma^2+\ldots+\gamma^{q-1}+\gamma^q\frac{1}{1-\gamma_{\min}}}. \tag{7}$$

It is easy to see that $TC_q(0)=C_I$, $TC_q(\infty)=(q+\gamma_{\min}/(1-\gamma_{\min}))C_w$, and $TC_q(\gamma_{\min})=(1-\gamma_{\min})C_I+\gamma_{\min}C_w/(1-\gamma_{\min})$, for all $q\ge 1$. Further properties of $TC_q(\gamma)$, as a function of γ and q, are collected in Propositions 1, 2, and 3.

Proposition 1. *The function $TC_q(\gamma)$ defined in equation (7), as a function of γ, has the following properties.*

i. *If $C_I \le C_w$, the function $TC_q(\gamma)$ is increasing in γ and, consequently, $TC_q(\gamma)\ge TC_q(0)=C_I$, for $\gamma\ge 0$.*
ii. *Assume that $C_I > C_w$ and $q=1$. We have $TC_1(\gamma)\ge C_w$. If $C_w/(1-\gamma_{\min}) > C_I$, $TC_1(\gamma)$ is increasing in γ; Otherwise, $TC_1(\gamma)$ is non-increasing in γ.*
iii. *Assume that $C_I > C_w$ and $q>1$. The function $-TC_q(\gamma)$ is unimodal in γ.*

□

Let $\gamma^*{}_1(q)$ be the optimal traffic intensity, i.e., $TC_q(\gamma)$ is minimized at $\gamma^*{}_1(q)$, and $TC^*{}_{K>1}(q)$ the corresponding minimal cost. Since $-TC_q(\gamma)$ is either monotone or unimodal in γ, it is easy to find the optimal $\gamma^*{}_1(q)$. Based on Proposition 1, we characterize the optimal solution $\{\gamma^*{}_1(q), TC^*{}_{K>1}(q)\}$.

Proposition 2. *For $K=2$, it holds that*

a. *If $C_I \le C_w$, the optimal solution of (4) for given q is $\rho_n=\gamma_{\min}$ for $n\ge 1$, and $TC^*{}_{K>1}(q)=TC^*{}_{K=1}$.*
b. *If $C_w < C_I < C_w/(1-\gamma_{\min})$ and $q=1$, then $\gamma^*{}_1(1)=\gamma_{\min}$ and $TC^*{}_{K>1}(1)=TC^*{}_{K=1}$.*
c. *If $C_I \ge C_w/(1-\gamma_{\min})$ and $q=1$, then $\gamma^*{}_1(1)=\infty$ and $TC^*{}_{K>1}(1)=C_w/(1-\gamma_{\min})$.*
d. *If $C_I > C_w$ and $q>1$, the optimal traffic intensity $\gamma^*{}_1(q)$ is the maximum of $\gamma_{\min}$ and the unique solution in $(0,\infty)$ satisfying*

$$\left(\sum_{i=0}^{q}\gamma^i+\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})}\right)\left(\sum_{i=1}^{q}i^2\gamma^{i-1}+q^2\gamma^{q-1}\frac{\gamma_{\min}}{(1-\gamma_{\min})}+q\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})^2}\right)C_w$$
$$-\left(\sum_{i=1}^{q}i\gamma^{i-1}+q\gamma^{q-1}\frac{\gamma_{\min}}{(1-\gamma_{\min})}\right)\left(C_I+\left(\sum_{i=1}^{q}i\gamma^i+q\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})}+\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})^2}\right)C_w\right)=0. \tag{8}$$

□

Proposition 2 shows that, if $C_I \geq C_w/(1-\gamma_{\min})$ and $q=1$, the optimal solution of equation (7) is $\rho^*{}_1=\infty$, which implies that there is no service if the queue length $q(t)=1$. This optimal solution explains why in many real systems (e.g., restaurants), service slows down when the queue is short. The reason is to keep the system loaded in order to reduce system idle cost. The result implies that, while it is necessary to stimulate herding effect, it is also important to make sure that a congestion is not resulted from reduced efficiency.

The next result characterizes the relationship between the optimal solution $\{\gamma^*{}_1(q), TC^*{}_{K>1}(q)\}$ of (7) and cost parameters $\{C_I, C_w\}$.

Proposition 3. *Consider the optimal policy* $\{\rho_n=\gamma^*{}_1(q)$, *for* $1\leq n\leq q$, *and* $\rho_n=\gamma_{\min}$, *for* $n>q\}$. *Then we have i)* $TC^*{}_{K>1}(q)$ *is increasing in* C_I/C_w, *and ii)* $\gamma^*{}_1(q)$ *is increasing in* C_I/C_w. *In addition, we have*

$$\lim_{\frac{C_I}{C_w}\to\infty} TC^*_{K>1}(q)=\left(q+\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)C_w, \quad \textit{and} \quad \lim_{\frac{C_I}{C_w}\to\infty}\gamma^*_1(q)=\infty. \tag{9}$$

□

Let $\gamma_{\min}$ be a variable instead of a fixed system parameter. Then we find that the queue with minimum $\gamma_{\min}$ has the smallest minimum long-run average cost $TC^*{}_{K>1}(q)$. This is a natural extension of Theorem 1, as shown in Corollary 1. Note that the linear increasing waiting cost in Corollary 1 can be relaxed to be non-decreasing.

Corollary 1. $TC^*{}_{K>1}(q)$ *is an increasing function in* $\gamma_{\min}$, *for* $0\leq\gamma_{\min}<1$.

Proposition 3 implies that, if the system idle cost is higher, then the minimal long-run average cost and the traffic intensity are higher. That implies that the system will choose a slower service rate (when the queue length is small). So, herding appears if the system is managed under the optimal policy, but the system efficiency is compromised since the service rate may be deliberately set to be smaller. This observation indicates that setting a higher idle cost to reduce system idleness also has negative impact on system performance. We remark that the impact of the ratio C_I/C_w on the performance of the system was addressed in some other works (see the survey paper by Cayirli and Veral (2003)) too.

Next, we keep the traffic intensities constant. For the $K=2$ case, we have $\rho_n=\gamma_1$, for $1\leq n\leq q_1$, and $\rho_n=\gamma_2$, for $n>q_1$, where γ_2 can be chosen as $\gamma_{\min}$. As we mention before, γ_2 may be regarded as different minimum traffic intensity $\gamma_{\min}$ of different queueing systems with other things being equal. If γ_1 and γ_2 are fixed, the long-run average cost is a function of $q=q_1$ only. We assume $\gamma_2<1$ to ensure that the queue is stable. By routine calculations, we obtain

$$\pi_0=\left(\frac{1-\gamma_1^{q+1}}{1-\gamma_1}+\gamma_1^q\gamma_2\frac{1}{1-\gamma_2}\right)^{-1}=\begin{cases}\dfrac{(1-\gamma_1)(1-\gamma_2)}{1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q}, & \text{if } \gamma_1\neq 1;\\ \dfrac{1-\gamma_2}{\gamma_2+(q+1)(1-\gamma_2)}, & \text{if } \gamma_1=1;\end{cases}$$

$$\pi_n = \begin{cases} \pi_0 \gamma_1^n, & \text{if } 1 \le n \le q; \\ \pi_0 \gamma_1^q \gamma_2^{n-q}, & \text{if } n > q; \end{cases} \tag{10}$$

The mean queue length, if it exists, can be obtained:

$$E[q(t)] = \begin{cases} \pi_0 \left(\gamma_1 \dfrac{1 - \gamma_1^{q+1} - (q+1)(1-\gamma_1)\gamma_1^q}{(1-\gamma_1)^2} + \gamma_1^q \gamma_2 \dfrac{1 + q(1-\gamma_2)}{(1-\gamma_2)^2} \right), & \text{if } \gamma_1 \ne 1; \\ \pi_0 \left(\dfrac{q(q+1)}{2} + \gamma_2 \dfrac{1 + q(1-\gamma_2)}{(1-\gamma_2)^2} \right), & \text{if } \gamma_1 = 1. \end{cases} \tag{11}$$

Let $TC_{\gamma_1,\gamma_2}(q) = TC(\rho_n,\ n \ge 1)$. Now, we characterize the cost function $TC_{\gamma_1,\gamma_2}(q)$ and find the optimal q for the $K = 2$ case.

Proposition 4. *Assume that $\gamma_2 < 1$. Then $-TC_{\gamma_1,\gamma_2}(q)$ is unimodal in q. The optimal q^* that minimizes $TC_{\gamma_1,\gamma_2}(q)$ is given by*

$$q^* = \begin{cases} \min\left\{ q \ge 1 : \ \dfrac{q+1}{1-\gamma_1} - \dfrac{(\gamma_1 - \gamma_2)\left(1 - \gamma_1^{q+1}\right)}{(1-\gamma_1)^2(1-\gamma_2)} \ge \dfrac{C_I}{C_W} \right\}, & \text{if } \gamma_1 \ne 1; \\ \min\left\{ q \ge 1 : \ \dfrac{q(q+1)}{2} + \dfrac{q+1}{1-\gamma_2} \ge \dfrac{C_I}{C_W} \right\}, & \text{if } \gamma_1 = 1. \end{cases} \tag{12}$$

□

Based on Proposition 4, the optimal q can be found by enumerating $TC_{\gamma_1,\gamma_2}(q)$ for $q = 0, 1, 2, \ldots$, until the first time that $TC_{\gamma_1,\gamma_2}(q)$ increases.

In the next proposition, we explore the relationship between the optimal solution $\{q^*, TC_{\gamma_1,\gamma_2}(q^*)\}$ and system parameters γ_1 and γ_2.

Proposition 5. *If γ_1 is fixed, q^* is a decreasing function in γ_2, for $0 \le \gamma_2 < 1$, which is piecewise constant. If γ_1 is fixed, $TC_{\gamma_1,\gamma_2}(q^*)$ is an increasing function in γ_2, for $0 \le \gamma_2 < 1$.* □

Proposition 5 implies that if the traffic intensity γ_2 is smaller (e.g., the service rate is higher), then the switching time can be set at longer queue length. If γ_2 is increasing, the queue length after switching is getting longer. Thus, it is better to switch at a smaller queue length. When γ_1 is small, a large q^* is chosen to keep the queue length at the right level. Proposition 5 also implies that it is optimal to use the smallest traffic intensity when a switching takes place if the queue length increases. In other words, when the queue length increases and a switch of service rate is warranted, the highest service rate (or the lowest arrival rate) should be selected (while the optimal q^* is kept). When γ_1 is large, it gives more flexibility to adjust the queue length (or waiting time), since one can set the switching time earlier. Consequently, the long-run average cost is reduced. Proposition 5 implies, again, that it is better-off for the

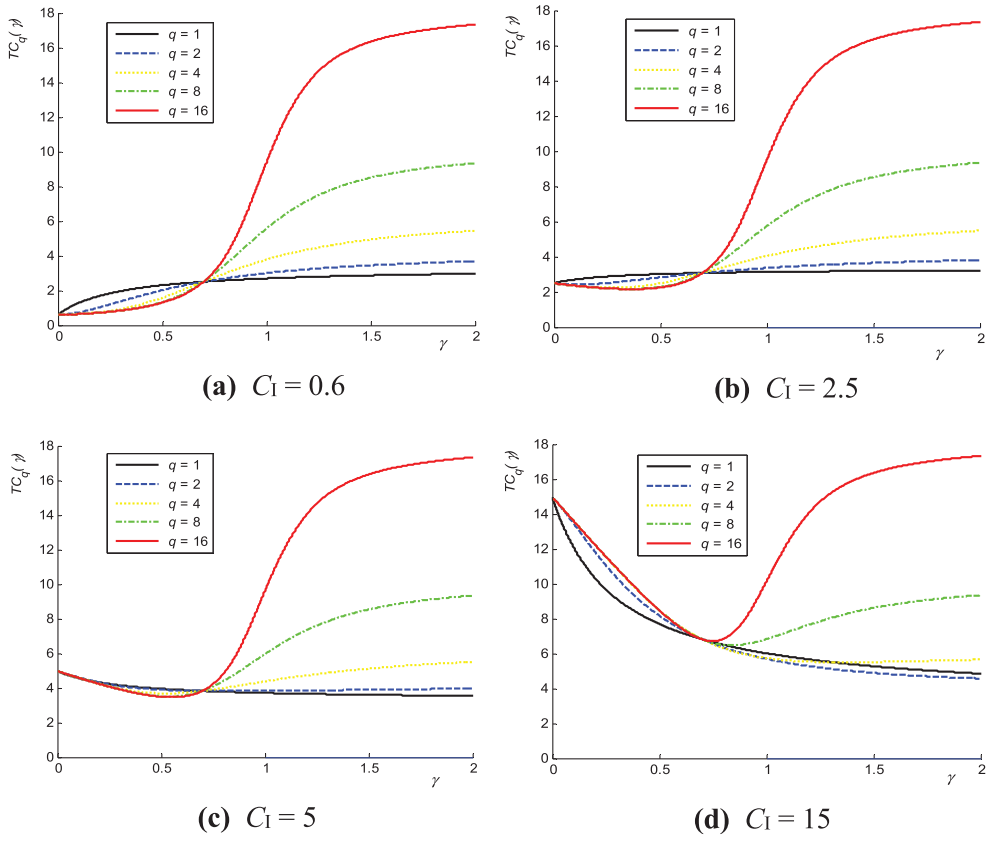

Figure 1. The function $TC_q(\gamma)$ for $C_w = 1$ and $\gamma_{min} = 0.7$ for Example 1.

system to slow down service when the system is less crowded, if the (only) trade-off is between system idleness and customer waiting time. This is consistent with the herding phenomenon in some service industry (e.g., restaurant).

Proposition 6. *Assume that* $\gamma_2 < 1$. *Then we have i)* q^* *is increasing in* C_I/C_w, *and ii) if* C_w *is fixed,* $TC_{\gamma_1,\gamma_2}(q^*)$ *is increasing in* C_I/C_w.

Proposition 7. *Assume that* $\gamma_2 < 1$ *and* γ_2 *is fixed. Then* q^* *is non-increasing in* γ_1, *if* $\gamma_2 \leq \gamma_1$.

Both Propositions 6 and 7 are intuitive. Proposition 7 indicates that q^* is always bounded.

In addition to the discussions and observations following the propositions, further insight and observations are provided in Section 5.

5. Numerical analysis

In this section, we present two sets of numerical examples to extend the insight we have learned from Propositions 1 to 5. Examples 1 to 3 are about the impact of traffic intensity γ, which are related to Propositions 1 to 3. Examples 4 to 6 are about the impact of rate switching state, which are related to Propositions 4 and 5.

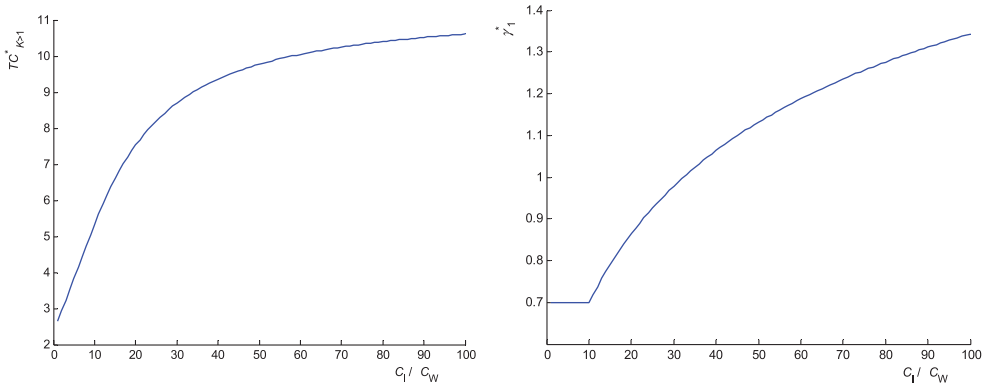

Figure 2. The optimal $TC^*_{K>1}$ and γ^*_1 as a function of C_I/C_W for Example 2.

From a decision-making point of view, especially in our study, it is sufficient to come up with relative values for costs. It is the estimate of the ratio C_I/C_W that we need, rather than the actual monetary values of C_I and C_W, even though they are available via standard cost accounting (Fries and Marathe (1981)). The relative cost ratio of C_I/C_W considered in the studies range from 1 to 100, as pointed out in Cayirli and Veral (2003).

Example 1 (Propositions 1 and 2) Consider a model with $C_w = 1$ and $\gamma_{min} = 0.7$. The function $TC_q(\gamma)$ for $C_I = 0.6$, 2.5, 5, and 15, and $q_1 = q = 1$, 2, 4, 8, and 16, is plotted in Figure 1. A few observations on $TC_q(\gamma)$ can be obtained from Figure 1. By equation (7), $TC_q(\gamma_{min}) = (1-\gamma_{min})C_I + \gamma_{min}C_w/(1-\gamma_{min})$, for all q. Thus, every function $TC_q(\gamma)$ goes through the point $(\gamma_{min}, (1-\gamma_{min})C_I + \gamma_{min}C_w/(1-\gamma_{min}))$. There are four typical cases of $TC_q(\gamma)$, as shown in Figure 1.

1. Figure 1(a) demonstrates that the function $TC_q(\gamma)$ is increasing in γ for all $q \geq 1$, which is consistent with Proposition 1 under condition $C_I \leq C_w$. The optimal γ for this case is the smallest possible traffic intensity (see Proposition 2).
2. Figure 1(b) demonstrates that the function $TC_q(\gamma)$ is increasing for $q = 1$ and $-TC_q(\gamma)$ is unimodal in γ for $q > 1$. Proposition 1 shows this property under condition $C_w < C_I \leq C_w/(1-\gamma_{min})$.
3. Figure 1(c) and Figure 1(d) demonstrate that the function $TC_q(\gamma)$ is decreasing in γ for $q = 1$ and $-TC_q(\gamma)$ is unimodal in γ for $q > 1$. Proposition 1 shows this property under condition $C_I > C_w/(1-\gamma_{min})$. The optimal γ is close to γ_{min} if q is sufficiently large.

Example 2. (Proposition 3) Consider the case with $C_w = 1$, $q = 10$, and $\gamma_{min} = 0.7$. The optimal γ^*_1 as a function of C_I/C_w is plotted in Figure 2.

It is intuitive (which is confirmed by Proposition 3) that both γ^*_1 and $TC_q(\gamma^*_1)$ are increasing in C_I/C_w. It is interesting to see that γ^*_1 is constant for small C_I/C_w, which is also intuitive since a system with a nominal idleness cost would prefer the server to work always at the high speed.

The relationship between $\gamma^*_1(q)$ and q is more complicated. Based on propositions in Section 4 and the above numerical results, we have the following summary (Figure 3).

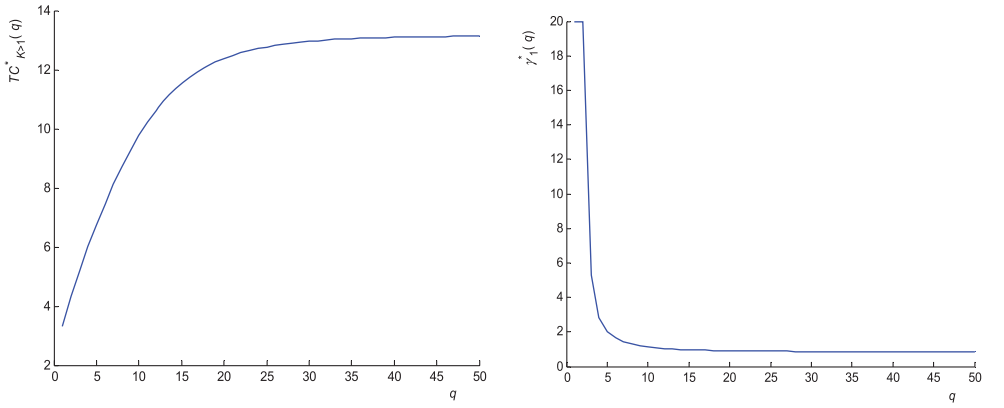

Figure 3. Functions $TC^*_{K>1}(q)$ and $\gamma^*_1(q)$ for Example 3.

a. If $C_I \leq (1-\gamma_{min})C_w$, $\gamma^*_1(q) = \gamma_{min}$ and $TC^*_{K>1}(q) = TC^*_{K=1}$ is constant in q.
b. If $C_I > (1-\gamma_{min})C_w$, $\gamma^*_1(q)$ is decreasing in q and converges to $\rho^*_{K=1}$, and $TC^*_{K>1}(q)$ is increasing in q, and converges to $TC^*_{K=1}$.

Example 3. Consider a model with $C_I = 50$, $C_w = 1$, and $\gamma_{min} = 0.7$. The functions $\gamma^*_1(q)$ and $TC^*_{K>1}(q)$ are plotted in Figure 3. The limit of $\gamma^*_1(q)$ is 0.8586 and the limit of $TC^*_{K>1}(q)$ is 13.1421, as q goes to infinity.

Intuitively, if q increases, the chance for a longer queue increases. Therefore, the traffic intensity can be smaller (to minimize the long-run average cost). On the other hand, the capability in manipulating the system idleness by adjusting γ_1 is reduced. Thus, the minimal long-run average cost is increased.

Example 4 (Proposition 4) Consider a system with $\lambda = 1$, $\mu_1 = 1.1$, $\mu_2 = 3$, $C_I = 20$, and $C_W = 1$. Since the (negative) cost function is unimodal in q, it is easy to see that the optimal switching point is $q^* = 5$, as shown in Figure 4. That is: if the queue length goes from 5 to 6, the service rate should be switched from μ_1 to μ_2.

Example 5 (Proposition 5) Consider a model with $C_I = 20$ and $C_w = 1$. The optimal q^* is plotted as a function of γ_2, for $\gamma_1 = 0.7$ and 2, in Figure 5. As γ_2 increases, the optimal q^* is getting smaller.

The optimal cost function $TC_{\gamma_1,\gamma_2}(q^*)$ is plotted as a function of γ_2 in Figure 6, for $\gamma_1 = 0.7$ and 2, which is increasing in γ_2.

Example 6 (Proposition 5) Fix $\gamma_2 = 0.3$ or 0.7, $C_I = 20$, and $C_w = 1$, the relationship between γ_1 and q^* and $TC_{\gamma_1,\gamma_2}(q^*)$ is exemplified in Figure 7.

As γ_1 increases, both $q_1{}^*$ and $TC_{\gamma_1,\gamma_2}(q^*)$ are decreasing. For smaller γ_2, both $q_1{}^*$ and $TC_{\gamma_1,\gamma_2}(q^*)$ are larger. Thus, γ_1 increases, the system should switch to slower server earlier.

It is clear that the numerical examples support the results obtained in Section 4. In particular, we would like to point out that the propositions indicate that the ratio of

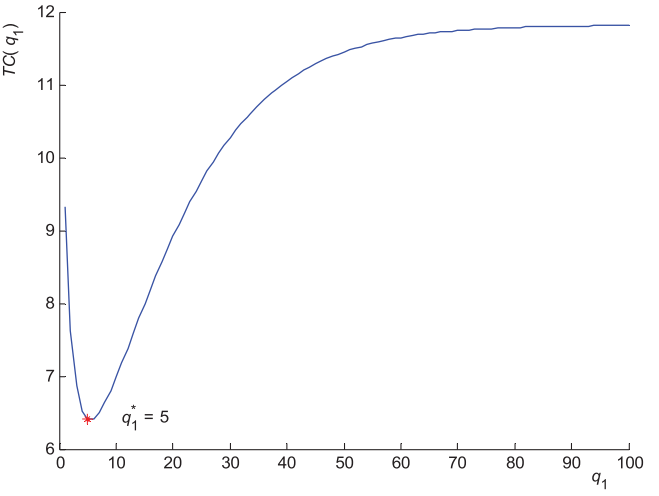

Figure 4. The function $TC_{\gamma_1,\gamma_2}(q)$ for Example 4.

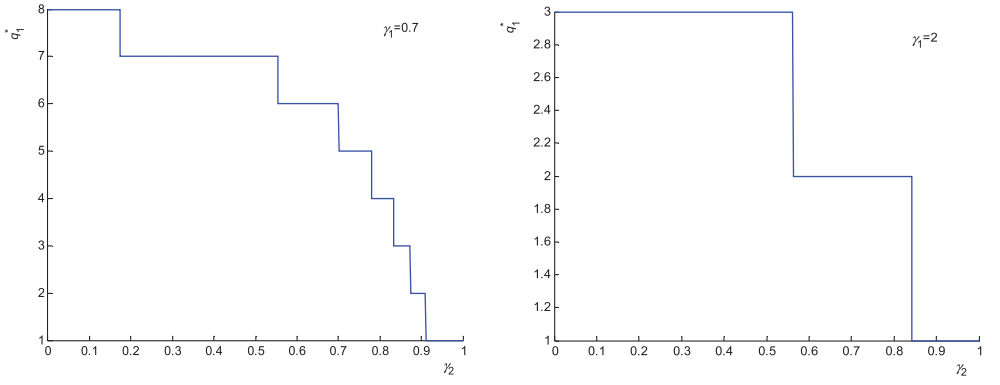

Figure 5. The optimal q^* as a function of γ_2 for Example 5.

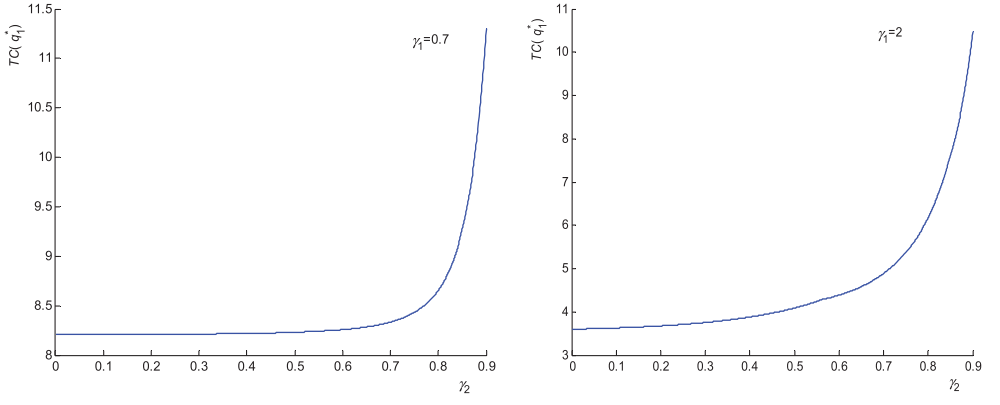

Figure 6. The optimal cost function $TC_{\gamma_1,\gamma_2}(q^*)$ as a function of γ_2 for Example 5.

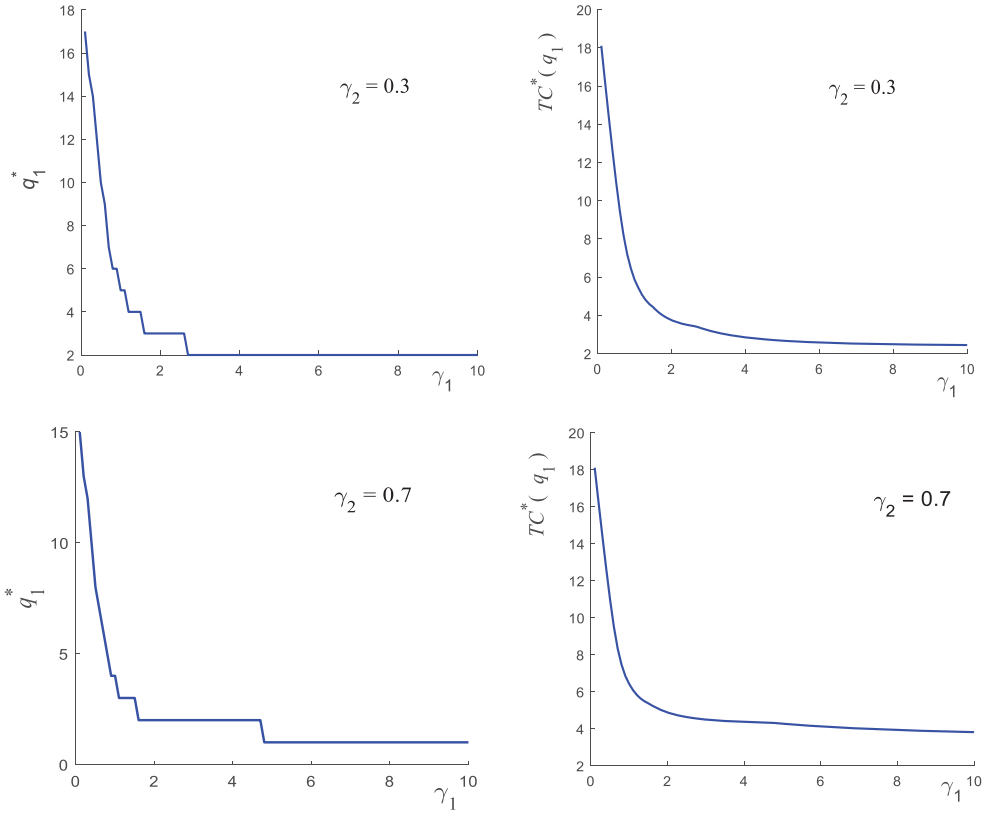

Figure 7. The optimal q^* and $TC_{\gamma_1,\gamma_2}(q^*)$ as a function of γ_1 for Example 6.

C_I and C_w, i.e., the trade-off between idle cost and waiting cost, has significant impact on i) when the traffic intensity should be switched, and ii) the magnitude of the change in traffic intensity. In general, if C_I/C_w is high, switching may occur as late as possible (i.e., at longer queue length), the traffic intensity should be smaller, and the long-run average cost is larger. Thus, reducing idle cost is a more effective way to reduce the total system cost than that of the waiting cost. Results also show that, in order to reduce the long-run average cost, the service speed should be as slow as possible, if the queue length is below the rate switching point; and the service speed should be as fast as possible, if the queue length is above the rate switching state. The results and observations provide strong support for the "non-idleness" and "herding" phenomena in service systems.

Example 7. Consider a model with $C_w = 1$, $C_I = 0.6$, $\gamma_{\min} = 0.7$, and non-linear waiting cost $C_w(n) = n^{0.5}C_w$ or $C_w(n) = n^2 C_w$. The long-run average cost $TC_q(\gamma)$, as a function of γ, is plotted in Figure 8.

Together with Figure 1(a), where $C_w(n) = nC_w$, Figure 8 demonstrates that the cost function $TC_q(\gamma)$ has similar properties for non-decreasing waiting cost functions. Although Sections 4 and 5 of this paper focus on the linear waiting cost case, the results can be obtained for the non-linear waiting cost case.

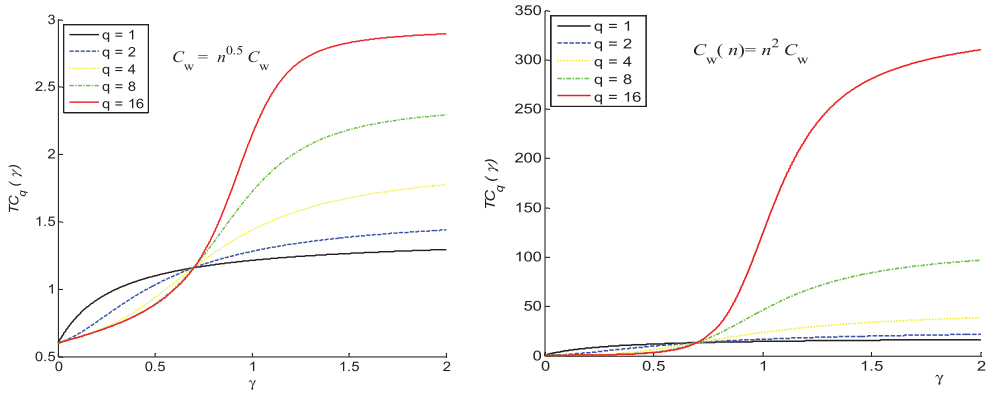

Figure 8. $TC_q(\gamma)$ as a function of γ for Example 7.

6. Conclusion

In this paper, we considered the trade-off between system idleness (herding) and customer waiting (congestion), and gained insight into the issue through a number of propositions and examples. We would like to point out that the insight gained in this paper applies to systems with multiple production/service facilities, if the service capacity in the $M/M/1$ queue can be considered as the aggregation of all service capacities in a stochastic system.

People observe idleness-aversion in service industries and try to explain this phenomenon by introducing conundrums between different costs. Holding cost and switching cost are definitely reasons why workers do not exert all their efforts all the time. Our study pointed out that idleness cost induced by herding may be another explanation for the slowdown and more importantly, for the intentional slowdown. Empirical findings on worker's productivity in service industries consistently indicate the behavior of intentional slowdown when workload is small and our study is the first to propose an explanation from a queueing perspective.

Although we start with waiting cost and idle cost in this paper, the solution approach can be extended to include operating cost, which is incurred whenever a server is working. If the server switches between different levels of traffic intensities when queue length reaches certain thresholds, then in this case, the operating cost can be added to the waiting cost, making it a piece-wise non-decreasing function of queue length.

In the literature, many works have considered the trade-off between customer waiting cost, system operating cost, and other types of costs. It is interesting to study the balance between all of them, i.e., system idleness, system operation, and customer waiting together. Mathematically, it is challenging to solve an optimization problem like (3) with more complex cost structure. Qualitatively, it is harder to gain insight into the problem of interest. Nevertheless, this is a good topic for future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Aksin Z, Armony M, Mehrotra V. 2007. The modern call center: A multi-disciplinary perspective on operations management research. Prod Oper Manage. 16(6):665–688.

Batta R, Berman O, Wang Q. 2007. Balancing staffing and switching costs in a service center with flexible servers. Eur J Oper Res. 177(2):924–938.

Becker GS. 1991. A note on restaurant pricing and other examples of social influences on price. J Political Econ. 99(5):1109–1116.

Bitran GR, Ferrer JC, Rocha e Oliveira P. 2008. Om forum- managing customer experiences: Perspectives on the temporal aspects of service encounters. Manuf Serv Oper Manage. 10(1): 61–83.

Cayirli T, Veral E. 2003. Outpatient scheduling in health care: a review of literature. Prod Oper Manage. 12(4):519–549.

Chan WK, Gao B. 2013. Unfair consequence of fair competition in service systems - An agent-based and queueing approach. Serv Sci. 5(3):249–262.

Chase R, Dasu S. 2001. Want to perfect your company's service? Use behavioral science. Harvard Bus Rev. 79(6), 78–84.

Cohen J. 2012. The single server queue. Amsterdam: North-Holland.

Conway RW, Maxwell WL. 1962. A queuing model with state dependent service rates. J Ind Eng. 12(2):132–136.

Crabill TB, Gross D, Magazine MJ. 1977. A classified bibliography of research on optimal design and control of queues. Oper Res. 25(2):219–232.

Crabill TB. 1972. Optimal control of a service facility with variable exponential service times and constant arrival rate. Manage Sci. 18(9):560–566.

Debo LG, Veeraraghavan SK. 2009. Models of herding behavior in operations management. In: Consumer-Driven Demand and Operations Management Models. Springer;(p. 81–112).

Delasay M, Ingolfsson A, Kolfal B, Schultz KL. (2014). The influence of load on service times. Tepper School of Business, Carnegie Mellon University. Working Paper.

Fetter RB, Thompson JD. 1966. Patients' waiting time and doctors' idle time in the outpatient setting. Health Services Research. 1(1):66–90.

Frei F. 2008. The four things a service business must get right. Harvard Bus Rev. 86(4): 70–80.

Fries BE, Marathe VP. 1981. Determination of optimal variable-sized multiple-block appointment systems. Oper Res. 29(2):324–345.

George JM, Harrison JM. 2001. Dynamic control of a queue with adjustable service rate. Oper Res. 49(5):720–731.

Grassmann WK, Chen X, Kashyap BR. 2001. Optimal service rates for the state-dependent M/G/1 queues in steady state. Oper Res Lett. 29(2):57–63.

Gupta I, Zoreda J, Kramer N. 1971. Hospital manpower planning by use of queueing theory. Health Serv Res. 6(1):76–82.

Hernandez-Maskivker G, Ryan G, del Mar PAmies Pallis ́e M. 2012. Queues as a sign of value in tourism services. In 2nd advances in hospitality and tourism marketing & management conference.

Hsee CK, Yang AX, Wang L. 2010. Idleness aversion and the need for justifiable busyness. Psychol Sci. 21(7):926–930.

Jain M. 2005. Finite capacity M/M/r queueing system with queue-dependent servers. Comput Math Appl. 50(1-2):187–199.

Jünger M, Liebling TM, Naddef D, Nemhauser GL, Pulleyblank WR, Reinelt G, Rinaldi G, Wolsey LA. 2009. 50 years of integer programming 1958-2008: From the early years to the state-of-the-art. Berlin Heidelberg: Springer.

Kc DS, Terwiesch C. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Manage Sci. 55(9):1486–1498.

Keller T, Laughhunn D. 1973. An application of queuing theory to a congestion problem in an outpatient clinic. Decis Sci. 4(3):379–394.

Kremer M, Debo L. 2012. Herding in a queue: A laboratory experiment. Chicago Booth Research Paper, 12–28.
Lindvall T. 2002. Lectures on the coupling method. Mineola, NY: Courier Dover Publications, Inc.
Lippman SA. 1975. Applying a new device in the optimization of exponential queuing systems. Oper Res. 23(4):687–710.
Lu Y. 2013. Data-driven system design in service operations [Unpublished doctoral dissertation]. Columbia University.
Maglio PP, Kieliszewski CA, Spohrer JC. 2010. Handbook of Service Science. New York: Springer.
Marshall AW, Olkin I. Arnold BC. 2011. Inequalities: theory of majorization and its applications. New York: Academic Press.
Mordfin R. 2014. Why long lines can be good for shoppers, and business? Capital Ideas Magazine.
Parkinson CN. 1955. Parkinson's law. *Economist*. Nov 19th.
Raz O, Ert E. 2008. Size counts": The effect of queue length on choice between similar restaurants. Adv Consum Res. 35:803–804.
Sabeti H. 1970. Optimal decision in queueing (Tech. Rep.). DTIC Document.
Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ. 1998. Modeling and worker motivation in JIT production systems. Manage Sci. 44(12-Part-1):1595–1607.
Sobel MJ. 1974. Optimal operation of queues. In Mathematical methods in queueing theory. Dordrecht: Springer; p. 231–261.
Stidham S, Weber R. 1993. A survey of Markov decision models for control of networks of queues. Queueing Syst. 13(1-3):291–314.
Stidham S, Weber RR. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. Oper Res. 37(4):611–625.
Stidham S. 1974. Stochastic clearing systems. Stochastic Processes Appl. 2(1):85–113.
Tan TF, Netessine S. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. Manage Sci. 60(6):1574–1593.
Veeraraghavan SK, Debo LG. 2011. Herding in queues with waiting costs: Rationality and regret. M&SOM. 13(3):329–346.
Weber RR, Stidham S. 1987. Optimal control of service rates in networks of queues. Adv Appl Probab. 19(1):202–218.
Weiss EN. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. IIE Trans. 22(2):143–150.
Yadin M, Naor P. 1967. On queueing systems with variable service capacities. Nav Res Logist Q. 14(1):43–53.
Yang KK, Lau ML, Quek SA. 1998. A new appointment rule for a single-server, multiple-customer service system. Nav Res Logist. 45(3):313–326.
Zhao H. 2014. Grads abandon Tencent, Baidu to sell street food. Beijing Today.

Appendix

Proof of Theorem 1. The result is obvious if $K=1$ (Note that $q_1=\infty$ for this case). Suppose that $K>1$. Without loss of generality, the limiting probabilities can be expressed in terms of γ_1 and $\{\rho_n, n > q_1\}$ as follows:

$$\begin{aligned}
&\pi_0 = \left(\sum_{i=0}^{q_1}\gamma_1^i + \gamma_1^{q_1}\sum_{n=q_1+1}^{\infty}\rho_{q_1+1}\cdots\rho_n\right)^{-1};\\
&\pi_n = \pi_0\gamma_1^n, \quad \text{for} \quad 1 \le n \le q_1;\\
&\pi_n = \pi_0\gamma_1^{q_1}\rho_{q_1+1}\rho_{q_1+2}\cdots\rho_n, \quad \text{for} \quad n > q_1.
\end{aligned} \tag{14}$$

Then the cost function (3) can be written as

$$TC(\rho_1,\rho_2,\ldots\rho_n,\ldots)=\frac{C_I+\left(\sum_{i=0}^{q_1}C_w(i)\gamma_1^i+\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}C_w(i)\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right)\right)}{\sum_{i=0}^{q_1}\gamma_1^i+\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right)}. \tag{15}$$

Next, we show that a solution of the structure $\rho_n=\eta_1$, for $1\le n\le q_1$, and $\rho_n=\gamma_{\min}$, for $n>q_1$, is better. It is easy to see that there exists $\eta_1\ge\gamma_1$ such that

$$\sum_{i=0}^{q_1}\eta_1^i+\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\gamma_{\min}^{i-q_1}\right)=\sum_{i=0}^{q_1}\gamma_1^i+\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right). \tag{16}$$

Since $C_w(n)$ is non-decreasing, Equation (16) leads to

$$\begin{aligned}
&\sum_{i=0}^{q_1}C_w(i)\eta_1^i-\sum_{i=0}^{q_1}C_w(i)\gamma_1^i=\sum_{i=0}^{q_1}C_w(i)\left(\eta_1^i-\gamma_1^i\right)\\
&\le C_w(q_1)\sum_{i=1}^{q_1}\left(\eta_1^i-\gamma_1^i\right)\\
&=C_w(q_1)\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right)-C_w(q_1)\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\gamma_{\min}^{i-q_1}\right)\\
&=\gamma_1^{q_1}\sum_{i=q_1+1}^{\infty}C_w(q_1)\left(\prod_{j=q_1+1}^{i}\rho_j\right)-\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}C_w(q_1)\gamma_{\min}^{i-q_1}\right)\\
&=\left(\gamma_1^{q_1}\sum_{i=q_1+1}^{\infty}C_w(i)\left(\prod_{j=q_1+1}^{i}\rho_j\right)-\eta_1^{q_1}\sum_{i=q_1+1}^{\infty}C_w(i)\gamma_{\min}^{i-q_1}\right)\\
&\quad-\left(\gamma_1^{q_1}\sum_{i=q_1+1}^{\infty}\left(C_w(i)-C_w(q_1)\right)\left(\prod_{j=q_1+1}^{i}\rho_j\right)-\eta_1^{q_1}\sum_{i=q_1+1}^{\infty}\left(C_w(i)-C_w(q_1)\right)\gamma_{\min}^{i-q_1}\right)\\
&=\left(\gamma_1^{q_1}\sum_{i=q_1+1}^{\infty}C_w(i)\left(\prod_{j=q_1+1}^{i}\rho_j\right)-\eta_1^{q_1}\sum_{i=q_1+1}^{\infty}C_w(i)\gamma_{\min}^{i-q_1}\right)-\left(\gamma_1^{q_1}\sum_{i=1}^{\infty}g(i)\left(\prod_{j=1}^{i}\rho_{q_1+j}\right)-\eta_1^{q_1}\sum_{i=1}^{\infty}g(i)\gamma_{\min}^{i}\right),
\end{aligned} \tag{17}$$

where $g(i)=C_w(i+q_1)-C_w(q_1)$, which is also nonnegative and non-decreasing.

Let $a_n=\gamma_1^{q_1}\prod_{j=1}^{n}\rho_{q_1+j}$ and $b_n=\eta_1^{q_1}\gamma_{\min}^{n}$, for $n\ge1$. Since $\rho_n\ge\gamma_{\min}$, for $n\ge q_1$, it is easy to see that $a_1/b_1\le a_2/b_2\le\ldots\le a_n/b_n\le\ldots$. It is routine to show that

$$\frac{a_n}{b_n}\le\frac{\sum_{k=n}^{N}a_k}{\sum_{k=n}^{N}b_k}\le\frac{a_N}{b_N},\quad\text{and}\quad\frac{a_{n-1}}{b_{n-1}}\le\frac{a_n}{b_n}\le\frac{\sum_{k=n}^{\infty}a_k}{\sum_{k=n}^{\infty}b_k}\le\lim_{N\to\infty}\frac{a_N}{b_N}. \tag{18}$$

Since $a_n/b_n\le a_{n+1}/b_{n+1}$, we then obtain, from equation (18), that

$$\frac{\sum_{k=1}^{\infty}a_k}{\sum_{k=1}^{\infty}b_k}\le\ldots\le\frac{\sum_{k=n}^{\infty}a_k}{\sum_{k=n}^{\infty}b_k}\le\frac{\sum_{k=n+1}^{\infty}a_k}{\sum_{k=n+1}^{\infty}b_k},n\ge1. \tag{19}$$

Define discrete random variables X_1 (or X_2) with probability distribution $P\{X_1=n\}=a_n/(a_1+a_2+\dots)$ (or $b_n/(b_1+b_2+\dots)$), for $n=1, 2, \dots$. Equation (19) implies that X_1 is stochastically larger than X_2, which leads to $E[C_w(X_1)] \geq E[C_w(X_2)]$ and $E[g(X_1)] \geq E[g(X_2)]$. Then we obtain

$$\frac{\gamma_1^{q_1}\sum_{i=1}^{\infty}\left(\prod_{j=1}^{i}\rho_{q_1+j}\right)}{\eta_1^{q_1}\sum_{i=1}^{\infty}\gamma_{\min}^{i}}=\frac{\sum_{k=1}^{\infty}a_k}{\sum_{k=1}^{\infty}b_k}\leq\frac{\gamma_1^{q_1}\sum_{i=1}^{\infty}g(i)\left(\prod_{j=1}^{i}\rho_{q_1+j}\right)}{\eta_1^{q_1}\sum_{i=1}^{\infty}g(i)\gamma_{\min}^{i}} \tag{20}$$

Since $\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right)-\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}\gamma_{\min}^{i-q_1}\right)=\sum_{i=0}^{q_1}\eta_1^i-\sum_{i=0}^{q_1}\gamma_1^i\geq 0$, equation (20) implies that $\gamma_1^{q_1}\sum_{i=1}^{\infty}g(i)\left(\prod_{j=1}^{i}\rho_{q_1+j}\right)-\eta_1^{q_1}\sum_{i=1}^{\infty}g(i)\gamma_{\min}^{i}\geq 0$. Then equation (17) implies

$$\sum_{i=0}^{q_1}C_w(i)\eta_1^i+\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}C_w(i)\gamma_{\min}^{i-q_1}\right)\leq\sum_{i=0}^{q_1}C_w(i)\gamma_1^i+\gamma_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}C_w(i)\left(\prod_{j=q_1+1}^{i}\rho_j\right)\right). \tag{21}$$

Define a policy $\rho_n=\eta_1$, for $1\leq n\leq q_1$, and $\rho_n=\gamma_{\min}$, for $n>q_1$. Then the long-run average cost for this policy is given by

$$TC(\eta_1,\eta_1,\dots\eta_1,\gamma_{\min},\gamma_{\min},\dots)=\frac{C_I+\sum_{i=0}^{q_1}C_w(i)\eta_1^i+\eta_1^{q_1}\left(\sum_{i=q_1+1}^{\infty}C_w(i)\gamma_{\min}^{i-q_1}\right)}{\sum_{i=0}^{q_1}\eta_1^i+\eta_1^{q_1}\frac{\gamma_{\min}}{(1-\gamma_{\min})}}. \tag{22}$$

Equations (15) (16) (17) (21), and (22) lead to

$$TC(\eta_1,\dots,\eta_1,\gamma_{\min},\gamma_{\min},\dots)\leq TC(\gamma_1,\dots,\gamma_1,\rho_{q_1+1},\dots\rho_n,\dots). \tag{23}$$

Thus, for any policy $\{\rho_n=\gamma_1,\ 1\leq n\leq q_1,\ \rho_n,\ n>q_1\}$ with $\rho_n\geq\gamma_{\min}$ for $n\geq 1$, there exists a policy of the form $\{\rho_n=\eta_1,\ 1\leq n\leq q_1,\ \rho_n=\gamma_{\min},\ n>q_1\}$ that has a smaller long-run average cost, which leads to the expected result.

Since K is finite and the above property holds for any set of $\{q_1,\ \dots,\ q_{K-1},\ q_K=\infty\}$, the optimal solution must have the desired structure. This completes the proof of Theorem 1. □

Note that, in the following proofs of propositions, the waiting cost function is given as $C_w(n)=nC_w$.

Proof of Proposition 1. Parts i) and ii) can be obtained easily. To prove iii), iv), and v), we find the derivative of $TC_q(\gamma)$ first. Denote the numerator and denominator in equation (7) as $f(\gamma)$ and $g(\gamma)$, respectively. Taking derivatives of both sides of equation (7) with respect to γ, we obtain

$$\begin{aligned}\frac{dTC_q(\gamma)}{d\gamma}&=\frac{f^{(1)}(\gamma)g(\gamma)-f(\gamma)g^{(1)}(\gamma)}{(g(\gamma))^2}\\&=\frac{\left(\sum_{i=0}^{q}\gamma^i+\gamma^q\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)\left(\sum_{i=1}^{q}i^2\gamma^{i-1}+q^2\gamma^{q-1}\frac{\gamma_{\min}}{1-\gamma_{\min}}+q\gamma^{q-1}\frac{\gamma_{\min}}{(1-\gamma_{\min})^2}\right)C_w}{\left(\sum_{i=0}^{q}\gamma^i+\gamma^q\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)^2}\\&\quad-\frac{\left(\sum_{i=1}^{q}i\gamma^{i-1}+q\gamma^{q-1}\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)\left(C_I+\left(\sum_{i=1}^{q}i\gamma^i+q\gamma^q\frac{\gamma_{\min}}{1-\gamma_{\min}}+\gamma^q\frac{\gamma_{\min}}{(1-\gamma_{\min})^2}\right)C_w\right)}{\left(\sum_{i=0}^{q}\gamma^i+\gamma^q\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)^2}\end{aligned} \tag{24}$$

By routine calculations, the numerator of the right hand side of equation (24), which is $f^{(1)}(\gamma)g(\gamma) - f(\gamma)g^{(1)}(\gamma)$, becomes

$$\frac{1}{1-\gamma_{\min}}\left(C_w\frac{1}{1-\gamma_{\min}} - C_I\right), \tag{25}$$

if $q = 1$; and

$$\begin{aligned} &C_w - C_I + C_w\sum_{i=1}^{q-2}\left(\frac{(i+2)(i+3)}{6} - \frac{C_I}{C_w}\right)(i+1)\gamma^i \\ &+ C_w\left(\frac{(q+1)(q+2)}{6} - \frac{C_I}{C_w} + \frac{\gamma_{\min}}{1-\gamma_{\min}}\left(q + \frac{1}{1-\gamma_{\min}} - \frac{C_I}{C_w}\right)\right)q\gamma^{q-1} \\ &+ C_w\sum_{i=q}^{2q-1}\left(\frac{1}{2}\sum_{j=i-q+1}^{q}(2j-i-1)^2 + \frac{\gamma_{\min}}{1-\gamma_{\min}}(2q-1-i)\left(2q-1-i+\frac{1}{1-\gamma_{\min}}\right)\right)\gamma^i \\ &\equiv \sum_{i=0}^{2q-2} a_i\gamma^i, \end{aligned} \tag{26}$$

if $q > 1$. Note that $a_{2q-1} = 0$.

If $C_I \leq C_w$, by equations (25) and (26), the derivative of $TC_q(\gamma)$ is nonnegative. Consequently, $TC_q(\gamma)$ is increasing in γ, which proves ii).

If $C_I > C_w$ and $q = 1$, the expected results are obtained from equation (25).

If $C_I > C_w$ and $q > 1$, we have $a_0 < 0$. It is easy to show that, if $a_i \geq 0$, then $(i+1)(i+2)/6 \geq C_I/C_w$ and $a_{i+1} \geq 0$, for $i < q-2$. Note that $a_i \geq 0$, for $i \geq q$. Consequently, $\{a_0, \ldots, a_{2q-2}\}$ may change sign at most once in $\{a_0, \ldots, a_{q-2}\}$, $\{a_{q-2}, a_{q-1}\}$, and $\{a_{q-1}, a_q\}$. If

$$\frac{q(q+1)}{6} > \frac{C_I}{C_w} \quad \text{and} \quad \frac{(q+1)(q+2)}{6} + \frac{\gamma_{\min}}{1-\gamma_{\min}}\left(q + \frac{1}{1-\gamma_{\min}}\right) < \frac{C_I}{(1-\gamma_{\min})C_w}, \tag{27}$$

the sequence $\{a_1, \ldots, a_{2q-2}\}$ changes sign exactly three times; Otherwise, it changes sign exactly once. Thus, by Descartes' rule of signs, the polynomial in equation (26) has either one positive root or three positive roots.

Next, we show that it is not possible to have three roots. If x is a root of the polynomial in equation (26), then we must have $f^{(1)}(x)/g^{(1)}(x) = f(x)/g(x)$, where $f^{(1)}(x)$ and $g^{(1)}(x)$ are the first derivative of $f(x)$ and $g(x)$. Suppose that there are three positive roots: $x < y < z$. Equation (26) implies that the derivative of $TC_q(\gamma)$ is negative at $\gamma = 0$. Then x and z are local minimums and y is a local maximum. Note that it can be shown that x, y, and z are not saddle points of $TC_q(\gamma)$. We must have $TC_q(x) < TC_q(y)$ and $TC_q(z) < TC_q(y)$, which implies that $f^{(1)}(x)/g^{(1)}(x) < f^{(1)}(y)/g^{(1)}(y)$ and $f^{(1)}(z)/g^{(1)}(z) < f^{(1)}(y)/g^{(1)}(y)$. Thus, the function $f^{(1)}(\gamma)/g^{(1)}(\gamma)$ is not monotone. On the other hand, the derivative of $f^{(1)}(\gamma)/g^{(1)}(\gamma)$ can be obtained as $(f^{(2)}(\gamma)g^{(1)}(\gamma) - f^{(1)}(\gamma)g^{(2)}(\gamma))/(g^{(1,2}$, where $f^{(2)}(\gamma)$ and $g^{(2)}(\gamma)$ are the second derivatives of $f(\gamma)$ and $g(\gamma)$, respectively. The numerator $f^{(2)}(\gamma)g^{(1)}(\gamma) - f^{(1)}(\gamma)g^{(2)}(\gamma)$ can be obtained as, by routine and tedious calculation,

$$
\begin{aligned}
&f^{(2)}(\gamma)g^{(1)}(\gamma)\\
&=\left(\sum_{i=0}^{q-2}(i+2)^2(i+1)\gamma^i+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}\right)\gamma^{q-2}\right)\left(\sum_{i=0}^{q-1}(i+1)\gamma^i+\frac{q\gamma_{\min}}{1-\gamma_{\min}}\gamma^{q-1}\right)\\
&=\sum_{i=0}^{q-2}\left(\sum_{j=1}^{i+1}j(i+3-j)^2(i+2-j)\right)\gamma^i+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}\right)\gamma^{q-2}\\
&\quad+\sum_{i=1}^{q-2}\left(\sum_{j=i+1}^{q}j(q+i+1-j)^2(q+i-j)+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}\right)(i+1)+(i+1)^2i\frac{q\gamma_{\min}}{1-\gamma_{\min}}\right)\gamma^{q-2+i}\\
&\quad+\left(\frac{q(q-1)}{1-\gamma_{\min}}\left(q+\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)\frac{q}{1-\gamma_{\min}}\right)\gamma^{2q-3},
\end{aligned}
\tag{28}
$$

$$
\begin{aligned}
&f^{(1)}(\gamma)g^{(2)}(\gamma)\\
&=\left(\sum_{i=0}^{q-1}(i+1)^2\gamma^i+\frac{q\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}\right)\gamma^{q-1}\right)\left(\sum_{i=0}^{q-2}(i+2)(i+1)\gamma^i+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\gamma^{q-2}\right)\\
&=\sum_{i=0}^{q-2}\left(\sum_{j=1}^{i+1}(j+1)j(i+2-j)^2\right)\gamma^i+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\gamma^{q-2}\\
&\quad+\sum_{i=1}^{q-2}\left(\sum_{j=i+1}^{q}j(j-1)(q+i+1-j)^2+\frac{q\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}\right)(i+1)i+(i+1)^2\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\right)\gamma^{q-2+i}\\
&\quad+\left(\frac{q}{1-\gamma_{\min}}\left(q+\frac{\gamma_{\min}}{1-\gamma_{\min}}\right)\frac{q(q-1)}{1-\gamma_{\min}}\right)\gamma^{2q-3},
\end{aligned}
\tag{29}
$$

and

$$
\begin{aligned}
&f^{(2)}(\gamma)g^{(1)}(\gamma)-f^{(1)}(\gamma)g^{(2)}(\gamma)\\
&=C_w\sum_{i=1}^{q-2}\left(\frac{1}{2}\sum_{j=1}^{i+1}j(i+2-j)\left((i+2-2j)^2+2(j+1)\right)\right)\gamma^i+\frac{q(q-1)\gamma_{\min}}{1-\gamma_{\min}}\left(q+\frac{1}{1-\gamma_{\min}}-1\right)\gamma^{q-2}\\
&\quad+C_w\sum_{i=1}^{q-2}\left(\frac{1}{2}\sum_{j=i+1}^{q}j(q+i+1-j)\left((q+i+2-2j)^2+j-1\right)\right)\gamma^{q-2+i}\\
&\quad+C_w\sum_{i=1}^{q-2}\left(\frac{\gamma_{\min}}{1-\gamma_{\min}}(i+1)q(q-1-i)\left(q+\frac{1}{1-\gamma_{\min}}-i-1\right)\right)\gamma^{q-2+i}>0.
\end{aligned}
\tag{30}
$$

Thus, $f^{(1)}(\gamma)/g^{(1)}(\gamma)$ is increasing in γ, which leads to a contradiction.

Therefore, the polynomial in equation (26) cannot have three positive roots, but has exactly one positive root. Then, if the polynomial in equation (26) is nonnegative at γ, then it is nonnegative for any value greater than γ. That implies that the function $-TC_q(\gamma)$ is unimodal in γ. This completes the proof of Proposition 1. □

Proof of Proposition 2. Part a), b), and c) can be obtained easily. Part d) can be obtained by the unimodality of the cost function $-TC_q(\gamma)$, which has been shown in Proposition 1. This completes the proof of Proposition 2. □

In the proofs of Propositions 3 and 4, we need the following result on the limiting probabilities $\{\pi_n,\ n=0,\ 1,\ 2,\ \dots\}$, which is well-known in the literature of stochastic comparison.

Lemma A.1 (Lindvall (2002)) Consider two state-dependent $M/M/1$ queues with parameters $\{\rho_{1,n},\ n=1,\ 2,\ \dots\}$ and $\{\rho_{2,n},\ n=1,\ 2,\ \dots\}$, respectively. Assume that $\rho_{1,n} \geq \rho_{2,n}$, for $n=1,\ 2,\ \dots$, and the two queues are stable. Then the corresponding limiting probabilities $\{\pi_{1,n},\ n=0,\ 1,\ 2,\ \dots\}$ and $\{\pi_{2,n},\ n=0,\ 1,\ 2,\ \dots\}$ satisfy $\pi_{1,0}+\pi_{1,1}+\ \dots\ +\pi_{1,n} \leq \pi_{2,0}+\pi_{2,1}+\ \dots\ +\pi_{2,n}$, for $n=0,\ 1,\ 2,\ \dots$. That is: the random variable having distribution $\{\pi_{1,n},\ n=0,\ 1,\ 2,\ \dots\}$ is stochastically larger than that having distribution $\{\pi_{2,n},\ n=0,\ 1,\ 2,\ \dots\}$. □

Proof of Proposition 3. Part i) is obtained directly from equation (7). To prove part ii), we write $\gamma^*_1(q)$ as γ^*_1 for convenience. We rewrite (3) as $TC(\gamma)=C_w(\pi_0C_I/C_w+E[q(t)]) \equiv C_w(\pi_0(\gamma)c+E[q(\gamma)])$, where $c=C_I/C_w$. Then $TC(\gamma^*_1)=C_w(\pi_0(\gamma^*_1)c+E[q(\gamma^*_1)])$. Due to the optimality of γ^*_1, we must have, for any $0<\gamma<\gamma^*_1$, $C_w(\pi_0(\gamma)c+E[q(\gamma)]) > C_w(\pi_0(\gamma^*_1)c+E[q(\gamma^*_1)])$. Now, we consider $c_\varepsilon > c$ and any $0<\gamma<\gamma^*_1$. By Lemma A.1, we must have $\pi_0(\gamma) > \pi_0(\gamma^*_1)$. Then

$$\begin{aligned} c_\varepsilon\pi_0(\gamma)+E[q(\gamma)] &= (c_\varepsilon-c)\pi_0(\gamma)+c\pi_0(\gamma)+E[q(\gamma)] \\ &\geq (c_\varepsilon-c)\pi_0(\gamma)+c\pi_0(\gamma^*_1)+E\big[q(\gamma^*_1)\big] \\ &\geq (c_\varepsilon-c)\pi_0(\gamma)+(c-c_\varepsilon)\pi_0(\gamma^*_1)+c_\varepsilon\pi_0(\gamma^*_1)+E\big[q(\gamma^*_1)\big] \\ &\geq (c_\varepsilon-c)\big(\pi_0(\gamma)-\pi_0(\gamma^*_1)\big)+c_\varepsilon\pi_0(\gamma^*_1)+E\big[q(\gamma^*_1)\big] \\ &\geq c_\varepsilon\pi_0(\gamma^*_1)+E\big[q(\gamma^*_1)\big]. \end{aligned} \tag{31}$$

Thus, any γ satisfying $0<\gamma<\gamma^*_1$ cannot be the optimal solution for the case with $c_\varepsilon > c$. Consequently, we must have $\gamma^*_{1,\varepsilon} \geq \gamma^*_1$. This proves ii). This completes the proof of Proposition 3. □

In order to prove Proposition 4, we first show two properties on π_0 and $E[q(t)]$ as a function of q.

Lemma A.2 Assume that $\gamma_2<1$. i) The probability π_0 is decreasing in q. ii) The mean queue length $E[q(t)]$ is increasing in q.

Proof. Part i) is obtained by using the explicit expression of π_0 given in equation (10). If $\gamma_1=1$, the result is obvious. If $\gamma_1\neq 1$, we rewrite the expression of π_0 as

$$\pi_0=\frac{(1-\gamma_1)(1-\gamma_2)}{1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)}=\frac{(\gamma_1-1)(1-\gamma_2)}{\gamma_1^q(\gamma_1-\gamma_2)-(1-\gamma_2)}. \tag{32}$$

Part i) is proved for $\gamma_1<1$ and $\gamma_1>1$.

If q increases, by Lemma A.1, the corresponding limit distribution of the queue length becomes stochastically larger. Thus, the mean queue length becomes larger (Marshall et al. (2011)). This proves part ii). This completes the proof of Lemma A.2. □

Proof of Proposition 4. We write $E[q(t)]$ as $E_q[q(t)]$ to emphasize $E[q(t)]$ as a function of q. We rewrite $E_q[q(t)]$ as $E_q[q(t)]=\pi_0(q)f_E(q)$. Then we have

$$\Delta\pi_0 = \pi_0(q+1) - \pi_0(q) = -\frac{(1-\gamma_1)^2(1-\gamma_2)(\gamma_1-\gamma_2)\gamma_1^q}{\left(1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q\right)\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}, \tag{33}$$

and

$$\begin{aligned}\Delta f_E &= f_E(q+1) - f_E(q)\\ &= \gamma_1\frac{\gamma_1^{q+1}(1-\gamma_1)+(q+1)(1-\gamma_1)\gamma_1^q(1-\gamma_1)-(1-\gamma_1)\gamma_1^{q+1}}{(1-\gamma_1)^2}\\ &\quad+\gamma_1^q\gamma_2\frac{\gamma_1(1+(q+1)(1-\gamma_2)-(1+q(1-\gamma_2))}{(1-\gamma_2)^2}\\ &= (q+1)\gamma_1^{q+1}+\gamma_1^q\gamma_2\frac{\gamma_1-1+(q(\gamma_1-1)+\gamma_1)(1-\gamma_2)}{(1-\gamma_2)^2}\\ &= (\gamma_1-\gamma_2)\gamma_1^q\frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}.\end{aligned} \tag{34}$$

Using equations (33) and (34), we can obtain

$$\begin{aligned}\Delta E\big[q(t)\big] &= E_{q+1}\big[q(t)\big] - E_q\big[q(t)\big]\\ &= \pi_0(q+1)\Delta f_E + \Delta\pi_0 f_E(q)\\ &= \frac{(1-\gamma_1)(1-\gamma_2)}{\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}\gamma_1^q(\gamma_1-\gamma_2)\frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}\\ &\quad-\frac{(1-\gamma_1)^2(1-\gamma_2)(\gamma_1-\gamma_2)\gamma_1^q}{\left(1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q\right)\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}f_E(q)\\ &= \frac{(1-\gamma_1)(1-\gamma_2)(\gamma_1-\gamma_2)\gamma_1^q}{\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}\left(\frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}-\frac{(1-\gamma_1)f_E(q)}{1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q}\right).\end{aligned} \tag{35}$$

For the long-run average cost function, we obtain

$$\begin{aligned}\Delta TC &= TC(q+1) - TC(q) = \Delta\pi_0 C_I + \Delta E\big[q(t)\big]C_W\\ &= -\frac{(1-\gamma_1)^2(1-\gamma_2)(\gamma_1-\gamma_2)\gamma_1^q}{\left(1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q\right)\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}C_I\\ &\quad+\frac{(1-\gamma_1)(1-\gamma_2)\gamma_1^q(\gamma_1-\gamma_2)}{\left(1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}\right)}\left(\frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}-\frac{(1-\gamma_1)f_E(q)}{1-\gamma_2-\gamma_1^{q+1}+\gamma_2\gamma_1^q}\right)C_W\\ &= \frac{(1-\gamma_1)(1-\gamma_2)(\gamma_1-\gamma_2)\gamma_1^q}{1-\gamma_2-\gamma_1^{q+2}+\gamma_2\gamma_1^{q+1}}\left(\frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}C_W-\frac{(1-\gamma_1)\big(f_E(q)C_W+C_I\big)}{1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)}\right).\end{aligned} \tag{36}$$

Now, we focus on the following function

$$\Delta h(q) = \frac{1+q(1-\gamma_2)}{(1-\gamma_2)^2}-\frac{(1-\gamma_1)\big(f_E(q)+C_I/C_W\big)}{1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)}. \tag{37}$$

If $\gamma_1 > 1$, it is easy to show that $1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)<0$. Then we always have $(1-\gamma_1)\left(1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)\right) \geq 0$. Then $\Delta h(q) \geq 0$ if and only if

$$\frac{1+q(1-\gamma_2)}{(1-\gamma_1)(1-\gamma_2)^2}\left(1-\gamma_2-\gamma_1^q(\gamma_1-\gamma_2)\right)-f_E(q) \geq \frac{C_I}{C_W}, \tag{38}$$

By routine calculation, it can be shown that $\Delta h(q) \geq 0$ if and only if

$$\begin{cases} \dfrac{q+1}{1-\gamma_1}-\dfrac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q+1}\right)}{(1-\gamma_1)^2(1-\gamma_2)} \geq \dfrac{C_I}{C_W}, & \text{if } \gamma_1 \neq 1; \\ \dfrac{q(q+1)}{2}+\dfrac{q+1}{1-\gamma_2} \geq \dfrac{C_I}{C_W}, & \text{if } \gamma_1 = 1; \end{cases} \tag{39}$$

The left hand side of equation (39) is increasing in q. Therefore, by equation (37), $\Delta h(q)$ is increasing in q. Thus, $\Delta TC_{\gamma_1,\gamma_2}(q)$ changes it sign at most once. Consequently, $-TC_{\gamma_1,\gamma_2}(q)$ is unimodal in q. This completes the proof of Proposition 4. □

Proof of Proposition 5. For the first part of Proposition 5, we consider three cases: $\gamma_1 < 1$, $\gamma_1 = 1$, and $\gamma_1 > 1$. Suppose that $\gamma_1 < 1$. Suppose that γ_2 increases by δ (>0) and $\gamma_2 + \delta < 1$. Then for any $q \geq q_1^*(\gamma_2)$, we have

$$\begin{aligned} &\frac{q+1}{1-\gamma_1} \geq \frac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q+1}\right)}{(1-\gamma_2)(1-\gamma_1)^2}+\frac{C_I}{C_W} \\ &\Rightarrow \frac{q+1}{1-\gamma_1} \geq \frac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q+1}\right)}{(1-\gamma_2)(1-\gamma_1)^2}+\frac{C_I}{C_W} \geq \frac{(\gamma_1-\gamma_2-\delta)\left(1-\gamma_1^{q+1}\right)}{(1-\gamma_2-\delta)(1-\gamma_1)^2}+\frac{C_I}{C_W}, \end{aligned} \tag{40}$$

since $(\gamma_1-\gamma_2)/(1-\gamma_2) > (\gamma_1-\gamma_2-\delta)/(1-\gamma_2-\delta)$. Therefore, by equation (12), we must have $q_1^*(\gamma_2+\delta) \leq q_1^*(\gamma_2)$. If $\gamma_1 = 1$, the result is obtained directly from equation (12). Finally, suppose that $\gamma_1 > 1$. For any $q \geq q_1^*(\gamma_2)$, we have

$$\begin{aligned} &\frac{(\gamma_1-\gamma_2)\left(\gamma_1^{q+1}-1\right)}{(1-\gamma_2)(1-\gamma_1)^2} \geq \frac{q+1}{\gamma_1-1}+\frac{C_I}{C_W} \\ &\Rightarrow \frac{(\gamma_1-\gamma_2-\delta)\left(\gamma_1^{q+1}-1\right)}{(1-\gamma_2-\delta)(1-\gamma_1)^2} \geq \frac{(\gamma_1-\gamma_2)\left(\gamma_1^{q+1}-1\right)}{(1-\gamma_2)(1-\gamma_1)^2} \geq \frac{q+1}{1-\gamma_1}+\frac{C_I}{C_W}, \end{aligned} \tag{41}$$

since $(\gamma_1-\gamma_2-\delta)/(1-\gamma_2-\delta) > (\gamma_1-\gamma_2)/(1-\gamma_2)$. Therefore, by equation (12), we must have $q_1^*(\gamma_2+\delta) \leq q_1^*(\gamma_2)$.

Next, we show the second part of Proposition 5. The first part of Proposition 5 implies that for given γ_2, q_1^* remains the same in an interval covering γ_2. Suppose that q_1^* remains the same in $[\gamma_2, \gamma_2+\delta)$ for $\delta > 0$. By routine calculations, the (right) derivative of $TC_{\gamma_1,\gamma_2}(q^*)$ with respect to γ_2 can be obtained as follows:

$$\begin{aligned} \frac{dTC_{\gamma_1,\gamma_2}(q_1^*)}{d\gamma_2} = {} & C_w \frac{\left((1-\gamma_1)^2-q_1^*(\gamma_1-1)(1-\gamma_2)^2+\left(\gamma_1^{q_1^*}-1\right)(\gamma_1-\gamma_2)^2\right)}{(1-\gamma_2)^2\left(1-\gamma_2-\gamma_1{}^{q_1^*+1}+\gamma_1{}^{q_1^*}\gamma_2\right)^2}\gamma_1{}^{q_1^*} \\ & - C_I \frac{(1-\gamma_1)^2}{\left(1-\gamma_2-\gamma_1{}^{q_1^*+1}+\gamma_1{}^{q_1^*}\gamma_2\right)^2}\gamma_1{}^{q_1^*}. \end{aligned} \tag{42}$$

By Proposition 5, if $\gamma_1 \neq 1$, $q_1{}^*$ satisfies $\frac{q_1^*+1}{1-\gamma_1} - \frac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q_1^*+1}\right)}{(1-\gamma_1)^2(1-\gamma_2)} > \frac{C_I}{C_W}$,

$$\begin{aligned}\frac{dTC_{\gamma_1,\gamma_2}(q_1^*)}{d\gamma_2} &\geq C_w \frac{\left((1-\gamma_1)^2 - q_1^*(\gamma_1-1)(1-\gamma_2)^2 + \left(\gamma_1^{q_1^*}-1\right)(\gamma_1-\gamma_2)^2\right)}{(1-\gamma_2)^2\left(1-\gamma_2-\gamma_1{}^{q_1^*+1}+\gamma_1{}^{q_1^*}\gamma_2\right)^2}\gamma_1{}^{q_1^*} \\ &\quad - C_w\left(\frac{q_1+1}{1-\gamma_1} - \frac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q_1+1}\right)}{(1-\gamma_1)^2(1-\gamma_2)}\right)\frac{(1-\gamma_1)^2\gamma_1{}^{q_1^*}}{\left(1-\gamma_2-\gamma_1{}^{q_1^*+1}+\gamma_1{}^{q_1^*}\gamma_2\right)^2} \\ &= C_w \frac{(\gamma_1-1)\gamma_2\left(\gamma_1^{q_1^*+1}+\gamma_2-\gamma_1^{q_1^*}\gamma_2-1\right)}{(1-\gamma_2)^2\left(1-\gamma_2-\gamma_1{}^{q_1^*+1}+\gamma_1{}^{q_1^*}\gamma_2\right)^2}\gamma_1{}^{q_1^*}.\end{aligned} \tag{43}$$

If $\gamma_1 > 1$, we have $\gamma_1{}^{1+q_1{}^*} + \gamma_2 - \gamma_1{}^{q_1{}^*}\gamma_2 - 1 > 0$, and if $\gamma_1 < 1$, we have $\gamma_1{}^{1+q_1{}^*} + \gamma_2 - \gamma_1{}^{q_1{}^*}\gamma_2 - 1 < 0$. For both cases, we have shown that $dTC_{\gamma_1,\gamma_2}(q_1^*)/d\gamma_2 > 0$.

If $\gamma_1 = 1$, equation (42) becomes

$$\frac{dTC_{\gamma_1,\gamma_2}(q_1^*)}{d\gamma_2} = C_w\frac{\left(1+2q_1^*(1-\gamma_2)+(q_1^*-1)q_1^*(1-\gamma_2)^2/2\right)}{(1-\gamma_2)^2(q_1^*+1-q_1^*\gamma_2)^2} - C_I\frac{1}{(q_1^*+1-q_1^*\gamma_2)^2}. \tag{44}$$

By Proposition 4, we obtain

$$\begin{aligned}\frac{dTC_{\gamma_1,\gamma_2}(q_1^*)}{d\gamma_2} &\geq C_w\frac{\left(1+2q_1^*(1-\gamma_2)+(q_1^*-1)q_1^*(1-\gamma_2)^2/2\right)}{(1-\gamma_2)^2(q_1^*+1-q_1^*\gamma_2)^2} \\ &\quad - C_w\left(\frac{q_1^*(q_1^*+1)}{2}+\frac{q_1^*+1}{1-\gamma_2}\right)\frac{(1-\gamma_2)^2}{(1-\gamma_2)^2(q_1^*+1-q_1^*\gamma_2)^2} \\ &= C_w\frac{\gamma_2}{(1-\gamma_2)^2(q_1^*+1-q_1^*\gamma_2)^2} > 0\end{aligned} \tag{45}$$

Next, we consider the situation if $\frac{q_1^*+1}{1-\gamma_1} - \frac{(\gamma_1-\gamma_2)\left(1-\gamma_1^{q_1^*+1}\right)}{(1-\gamma_1)^2(1-\gamma_2)} = \frac{C_I}{C_W}$. For this case, $q_1{}^*$ is optimal in $(\gamma_2-\delta, \gamma_2]$ and $q_1{}^*-1$ is optimal in $(\gamma_2, \gamma_2+\delta)$ for (sufficiently) small $\delta > 0$. That indicates that $q_1{}^*-1$ is not optimal at γ_2. Therefore, we must have $\lim_{\delta\to 0+} TC_{\gamma_1,\gamma_2}(q_1^*-1) - \lim_{\delta\to 0-} TC_{\gamma_1,\gamma_2}(q_1^*) \geq 0$. In fact, under the condition, we have $\lim_{\delta\to 0+} TC_{\gamma_1,\gamma_2}(q_1^*-1) = \lim_{\delta\to 0-} TC_{\gamma_1,\gamma_2}(q_1^*)$. Therefore, $TC_{\gamma_1,\gamma_2}(q^*)$ is increasing in γ_2 for this case.

Summarizing all the cases, function $TC_{\gamma_1,\gamma_2}(q^*)$ is continuous and piecewise increasing in γ_2. Consequently, $TC_{\gamma_1,\gamma_2}(q^*)$ is increasing in γ_2. This completes the proof of Proposition 5. □

Proof of Proposition 6. For part i), we first rewrite the inequality in equation (12) as follow:

$$\frac{q+1}{1-\gamma_1} + \frac{(\gamma_2-\gamma_1)\left(\gamma_1^q+\gamma_1^{q-1}+\ldots+\gamma+1\right)}{(1-\gamma_1)(1-\gamma_2)} \geq \frac{C_I}{C_W} \tag{46}$$

The result is obtained if $\gamma_1 \leq \gamma_2$ (< 1). If $\gamma_2 < \gamma_1$, we further rewrite equation (46) as

$$\frac{q+1}{1-\gamma_2} + \frac{(\gamma_1-\gamma_2)\left(\gamma_1^{q-1}+2\gamma_1^{q-2}+\ldots+(q-1)\gamma_1+q\right)}{(1-\gamma_2)} \geq \frac{C_I}{C_W} \tag{47}$$

which leads to the desired result. Part ii) is obvious from equation (7). This completes the proof of Proposition 6. □

Proof of Proposition 7. The result is obvious from equation (47). This completes the proof of Proposition 7. □