Strategic Dynamic Jockeying Between Two Parallel Queues

Amin Dehghanian[1] and Jeffrey P. Kharoufeh[2]
Department of Industrial Engineering
University of Pittsburgh
1048 Benedum Hall
3700 O'Hara Street
Pittsburgh, PA 15261 USA

and

Mohammad Modarres[3]
Department of Industrial Engineering
Sharif University of Technology
Tehran, Iran

Final version appears in
Probability in the Engineering and Informational Sciences, 30 (1), 41-60, 2016.

Abstract

Consider a two-station, heterogeneous parallel queueing system in which each station operates as an independent $M/M/1$ queue with its own infinite-capacity buffer. The input to the system is a Poisson process that splits among the two stations according to a Bernoulli splitting mechanism. However, upon arrival, a strategic customer initially joins one of the queues selectively and decides at subsequent arrival and departure epochs whether to jockey (or switch queues) with the aim of reducing her own sojourn time. There is a holding cost per unit time, and jockeying incurs a fixed nonnegative cost while placing the customer at the end of the other queue. We examine individually optimal joining and jockeying policies that minimize the strategic customer's total expected discounted (or undiscounted) costs over finite and infinite time horizons. The main results reveal that, if the strategic customer is in station 1 with ℓ customers in front of her, and q_1 and q_2 customers in stations 1 and 2, respectively (excluding herself), then the incentive to jockey increases as either ℓ increases or q_2 decreases. Numerical examples reveal that it may not be optimal to join, and/or jockey to, the station with the shortest queue or the fastest server.

Keywords: Jockeying; optimal dynamic control; Markov decision processes.

[1]Email: amd120@pitt.edu.
[2]Corresponding author. Email: jkharouf@pitt.edu; Ph: +1 412 624-9832.
[3]Email: modarres@sharif.edu; Ph: +98 21 6616-5719.

1 Introduction

This paper examines a two-station, heterogeneous parallel queueing system in which each station possesses a single server and an infinite-capacity buffer. Customers arrive to the system according to a Poisson process with rate λ ($\lambda > 0$) and split among the two stations according to a Bernoulli splitting mechanism, i.e., an arriving customer joins station i with probability λ_i/λ, $i = 1, 2$, where $\lambda = \lambda_1 + \lambda_2$. Service times at station i are independent and identically distributed (i.i.d.) exponential random variables with rate μ_i ($\mu_i > 0$), and it is assumed that $\lambda_i < \mu_i$ for $i = 1, 2$ (i.e., both queueing stations are stable). Consider now a strategic customer who, upon arrival, must decide which station to join initially and, once there, may decide to switch (or jockey) to the other queue an indefinite number of times. Whenever the jockeying option is exercised, the strategic customer must join the end of the other queue. Additionally, there is a fixed, bounded cost c_{ij} ($c_{ij} \geq 0$, $j \neq i$) associated with jockeying from station i to j, as well as a bounded cost per unit time h ($h \geq 0$) of being held in the system.

The queueing model is motivated by stochastic service systems in which a customer can fully observe the system state – namely her position in the queue and the queue lengths and service rates of all stations – and utilize this information to minimize the time spent in the system. A quintessential example is that of an individual customer checking out of a supermarket. Once shopping is completed, the customer must decide which waiting line to join initially. Subsequently, while waiting to receive service, the customer may choose to join a different waiting line if it is perceived that the waiting time can be reduced by switching lines. However, if the jockeying option is taken, the customer must join the end of the other queue and forfeit her position in the current queue. Other motivating examples are prevalent in the telecommunications setting. For instance, Zhao and Grassman [22] and Xu and Zhao [20] (among others) described the jockeying phenomenon in multi-beam satellite systems. In these systems, ground-based stations that are organized in disjoint regions of the earth relay data packets to overhead satellites for processing and/or transmission. The packets can join one of (possibly) several buffers to await processing by the satellite, which uses multiple beams to transmit packets to other destinations. Packets can be moved to buffers with shorter queues in an effort to reduce transmission delays and/or prevent buffer overflow (in the case of finite-capacity buffers).

The study of queueing systems with jockeying has a long history dating back to the 1950s; however, most prior work has focused on the performance analysis of parallel queueing stations in which a queue joining and/or jockeying rule is assumed – the vast majority of existing models do not prescribe either socially or individually optimal policies. One of the earliest papers dealing with jockeying in a two-station, parallel system is due to Haight [7] who analyzed a system in which arriving customers always join the shortest queue initially. The author then provides a steady state analysis for two distinct cases: (1) when jockeying is not allowed; and (2) when jockeying is allowed, but only for the customer at the end of the waiting line. In this case, jockeying only occurs when the other line is shorter. Similar models involving the performance analysis of systems with two (or more) parallel stations have appeared in the applied probability and operations research literature. Some notable contributions are due to Kingman [12], Koenigsberg [13], Disney and Mitchell [2], Gertsbakh [5], Elsayed and Bastani [4], Kao and Lin [11], Zhao and Grassman [21, 22], Adan et al. [1] and, more recently, Tarabia [17]. The majority of these models consider two-station, parallel Markovian queueing stations wherein a customer is transferred from one queue to another when the difference between the two queue lengths exceeds a certain threshold. Subsequently, the steady state distribution of the number of customers in the system, and other standard performance parameters, are derived.

The problem of determining which queue to join initially has been studied in some depth over the past few decades. Weber [18] considered a queueing system with multiple identical servers, each with its own queue. Customers arrive according to a general stochastic process, and upon arrival, each customer is assigned to one of the queues. Once assigned, no jockeying is allowed. Weber showed that, if the service times have a nondecreasing hazard rate, then the policy of assigning an arriving customer to the shortest queue maximizes the number of customers served by a certain time (a socially optimal consideration). Later, Whitt [19] illustrated by way of counterexamples that the strategy of joining the shortest queue may be suboptimal for either social or individual objectives if the service time condition of Weber [18] is not met. In fact, Whitt shows that joining the *longer* queue can actually decrease the total expected sojourn time for an individual customer. Glazer and Hassin [6] studied a multi-station queueing system with i.i.d. exponential service times to which customers arrive according to a Poisson process. Upon arrival, the number of customers in each station is unobservable, so an arriving customer must visit a station (without necessarily joining the queue) to determine the number of customers present at that station. However, visiting a station consumes time and incurs a certain cost. Once the customer decides to join a station, switching stations is not permitted. Assuming the number of stations is large, the authors determine the customer's optimal search behavior that minimizes the total costs incurred. Haviv and Hassin [8] consider a two-station Markovian system with homogeneous servers. Neither queue length is observable to an arriving customer; however, information about which queue is shorter can be purchased at a fixed cost. Informed customers join the shortest queue, and uninformed customers (i.e., those who do not purchase information) join one of the queues at random. Information can be acquired once, and only upon arrival. Customers (whether informed or uninformed) jockey from the last position of one line to the end of the other line (without cost) whenever the difference between two queue lengths reaches a given threshold. The customers' costs include the cost of purchasing information upon arrival and the holding cost in the system. Because the decision of one customer influences the decisions of others, it is shown how to obtain Nash-equilibrium policies. Hassin and Haviv [9] provide a comprehensive survey of equilibrium behavior of servers and customers in queueing systems.

As for individually optimal policies, Mandelbaum and Yechiali [15] considered a strategic customer seeking service at an $M/G/1$ queue. General (or non-strategic) customers join the system immediately; however, a strategic customer chooses among three alternatives: (1) enter the system (join); (2) leave the system (balk); or (3) wait outside of the system and observe. The strategic customer incurs a cost associated with each action. If she enters or leaves the system, the decision is final and no further actions are taken. If the strategic customer chooses to wait and observe, another decision is made at the next service completion epoch. Assuming linear holding time costs, the authors show that for any finite n-period horizon, the customer's (individually) optimal strategy is a three-region policy. Specifically, it is optimal for the strategic customer to enter a short queue, balk from a large queue and wait when the queue size is intermediate. Necessary and sufficient conditions for waiting are also provided. Hlynka et al. [10] examined a queueing system comprised of two heterogeneous, exponential servers with respective rates μ_1 and μ_2. The input to the system is a renewal process, and no customer jockeying is allowed. General customers join the shortest queue upon arrival, but a strategic customer, who cannot observe service rates, may delay joining the system and instead choose to observe the system's operation for some time. The authors obtain sufficient conditions for which the strategic customer can reduce her expected sojourn time by waiting rather than immediately joining the shortest queue. However, they do not prescribe an optimal policy as was done in [15]. Lippman and Stidham [14] studied a single-server Markovian queueing system with a state-dependent service rate, i.e., the service rate is μ_i when there are i customers in the system, and μ_i is a bounded, non-decreasing, concave function of i. Associated with each arriving customer is a nonnegative reward r, and rewards of successive customers are i.i.d. random variables. The system is controlled by accepting or rejecting arriving customers. An arriving customer receives the reward r immediately if accepted into the system and receives nothing if rejected. An accepted customer joins the system and incurs a waiting cost h per unit time until departure. Lippman and Stidham [14] established that under both the total expected discounted and undiscounted reward criteria (over both finite and infinite horizons), the individually optimal policy calls for the customer to enter the system whenever the socially optimal policy does (i.e., individual and optimal policies coincide).

The literature related to the optimal control of parallel queueing stations with jockeying is relatively sparse. Specifically, most contributions are focused on socially optimal policies, i.e., those considering system-wide objectives. Most relevant to our work here, Xu and Zhao [20] investigated dynamic routing and jockeying policies in a two-station parallel queueing system with the objective of minimizing the total expected (holding plus jockeying) costs of the system, and established the optimality of monotone threshold policies for routing and jockeying. In their model, a controller makes routing and jockeying decisions, and they proved the optimality of a threshold routing policy, i.e., it is optimal to route to station 2 if the number in station 2 is less than a nondecreasing function of the number in station 1. The optimal jockeying policy is similarly a threshold policy. That is, if it is optimal for the controller to move a job from one station to the other in a given state, then it must also be optimal to move the job when the other station is less congested. The policy is symmetric (i.e., it holds for whichever queue the job is located). Additionally, they proved that, under some general conditions, these thresholds have finite asymptotes. Another model examining socially-optimal policies is due to Down and Lewis [3] who considered an N-station, parallel queueing system with independent arrivals streams at each queue. A controller may move customers from one queue to another, and each time customers are moved, a fixed cost and a variable cost (that depends on the number of customers moved) are incurred. They examined conditions for stability and instability, and for stable systems it was shown that, for a two-station system, it is optimal to store customers in the least-cost queue (under the long-run average cost criterion). Their model is more general than that of Xu and Zhao [20] in that it considers multiple stations and general arrival and service processes.

The primary objective of our work is to analyze the optimal, dynamic behavior of an individual, strategic customer who seeks service at a two-station, parallel queueing system. To that end, we present a classical Markov decision process (MDP) model in which, at each arrival and departure epoch of either station, the strategic customer has the option of either remaining in the current queue or jockeying to the other queue. Within the MDP framework, we additionally characterize the optimal joining policy for the strategic customer at the time of arrival to the system. The objective is to minimize the total expected discounted (or undiscounted) jockeying and holding costs over both finite and infinite time horizons. The main contributions of our work are as follows. First, we prove that, given the two queue lengths, there exists a threshold number of customers *ahead* of the strategic customer above which it is optimal to jockey to the other station. Likewise, there exists a threshold number of customers in the other queue below which it is optimal to jockey. For both cases, the monotonicity of the thresholds is established. Second, we show that the joining policy is also a monotone threshold policy, namely that for each number of customers in station 2, there exists a threshold number of customers in station 1 above which it is optimal to join station 2. Interestingly, numerical results reveal that, for individual optimization, it may not be optimal to join, and/or jockey to, the station with the shortest queue or the fastest server. Finally, we establish that the optimal joining and jockeying policies are valid under the total (undiscounted) cost criterion for both finite and infinite time horizons.

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the queueing model, and Section 3 presents the Markov decision process model. In Section 4, we establish the optimality of threshold jockeying and joining policies, as well as the monotonicity of the thresholds. Section 5 demonstrates that the same results are valid under the total (undiscounted) cost criterion over an infinite horizon. Finally, in Section 6 we illustrate the main results by way of two numerical examples that reveal interesting insights, namely that it may not be optimal join, and/or jockey to, the station with the shortest queue or the fastest server.

2 Queueing Model Description

Consider a queueing system comprised of two parallel, heterogeneous stations, each with its own dedicated infinite waiting room. External customers arrive to the system according to a Poisson process with rate λ ($\lambda > 0$), and customers can be processed at either station. The primary arrival stream splits according to a Bernoulli splitting mechanism, i.e., there exist positive rates λ_1 and λ_2 such that $\lambda = \lambda_1 + \lambda_2$, and an arbitrary arrival elects to join station i with probability λ_i/λ, $i = 1, 2$. The service times at station i are i.i.d. exponential random variables with rate parameter μ_i ($\mu_i > 0$); therefore, each station operates as an $M/M/1$ queue in which customers are served in the order that they arrive to the system. Figure 1 depicts the two-station, parallel system.

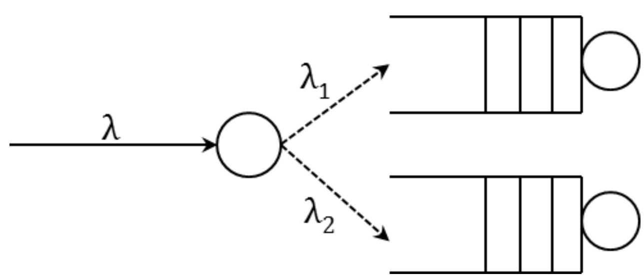

Figure 1: Graphical depiction of the two-station queueing system.

Consider now a strategic customer who joins one of the queues *selectively* at the time of arrival. Subsequently, at the arrival and departure epochs of either queue, the strategic customer may choose to jockey (or join the other queue) at a nonnegative fixed cost; however, the customer must join the end of the other queue if the jockeying option is selected. If the customer is currently in station i, then the cost of jockeying to station j ($j \neq i$) is c_{ij} ($c_{ij} \geq 0$). (It is assumed that the time needed to jockey is negligible.) In addition to the fixed jockeying cost, the customer also incurs a holding cost h ($h \geq 0$) per unit of time spent in the system (either in queue or in service). Both c_{ij} and h are bounded. It is assumed that preemption is allowed, i.e., if the strategic customer is currently being served, the customer may choose to stop service and join the end of the other queue. The terminal cost (whether service is completed or not) is assumed to be zero. The objective is to determine optimal joining and jockeying policies that minimize the strategic customer's total expected discounted holding and jockeying costs. Additionally, we extend the results to the total (undiscounted) cost criterion. Considered are both finite and infinite time horizons. Section 3 presents the main problem formulation, as well as the optimality equations. Our main results are stated in Section 4.

3 Discounted Markov Decision Process Model

In this section, we first analyze the problem of optimally jockeying between the two stations and then use this framework to decide which line the strategic customer should initially join upon arriving to the system. To this end, we formulate the control problem as a Markov decision process (MDP) model. The majority of the results will be developed under the discounted cost criterion, and these will be shown to carry over to the total (undiscounted) cost criterion as well.

The state of the system at time t is the vector $X(t) = (Q_1(t), Q_2(t), L(t))$, where $Q_i(t)$ denotes the number of customers in station i at time t (in the queue or being served and excluding the strategic customer), and $L(t)$ denotes the number of customers ahead of the strategic customer, irrespective of the station currently occupied by the strategic customer. (To distinguish the strategic customer's location, we later introduce separate notation for the value function corresponding to each case.) Denote the state space of $\{X(t) : t \geq 0\}$ by the denumerable set S. Because arrivals occur according to a Poisson process, and service times are i.i.d. exponential random variables at both stations, it is clear that $\{X(t) : t \geq 0\}$ is a continuous-time Markov chain on S. By employing the method of uniformization, we convert the continuous-time model to an equivalent discrete-time MDP model to solve the problem of minimizing the strategic customer's discounted holding and jockeying costs over both finite and infinite horizons. The decision epochs are customer arrival and departure times (service completion epochs) in either queue, and the time period between any two consecutive decision epochs is referred to as a period. Specifically, period n refers to the time interval between decision epochs $n-1$ and n. At decision epochs, the strategic customer observes a state $x = (q_1, q_2, \ell) \in S$, where q_i is the total number of customers in station i, including any in service but excluding the so-called strategic customer, and ℓ is the number of customers ahead of the strategic customer in her current queue. Obviously, it must be the case that $\ell \leq q_i$ if the strategic customer is currently in queue i.

The strategic customer, whether in station 1 or station 2, can choose at each arrival and departure epoch (of either station) to either remain in the current station or jockey to the other station. Denote the finite action set by $A \equiv \{0, 1\}$, where 0 means 'remain in the current queue' and 1 means 'jockey to the other queue'. Recall that, if the decision to jockey is made, the customer must join the end of the line in the other station. The total jockeying (c_{ij}) and holding (h) costs are continuously discounted at rate α ($\alpha > 0$) over the time horizon. The objective is to determine optimal joining and jockeying strategies that minimize the strategic customer's total expected discounted jockeying and holding costs over finite and infinite time horizons.

We now define the cost functions. Starting in state (q_1, q_2, ℓ), let $V_n(q_1, q_2, \ell)$ be the minimal total expected discounted cost with n periods remaining in the time horizon when the strategic customer is in station 1. The initial condition is $V_0(q_1, q_2, \ell) = 0$. Moreover, when the customer is in station 1, let $V_n[(q_1, q_2, \ell), a]$ be the minimal total expected n-period discounted cost with initial state (q_1, q_2, ℓ) when action $a \in A$ is taken. To help facilitate our proofs later, if the strategic customer is in station 2, let $V_n^{'}(q_1, q_2, \ell)$ be the minimal total expected n-period discounted cost starting in state (q_1, q_2, ℓ), where in this case ℓ refers to the number of customers ahead of the strategic customer in station 2. Additionally, let $V_n^{'}[(q_1, q_2, \ell), a]$ be defined in a manner similar to $V_n[(q_1, q_2, \ell), a]$. For the sake of convenience, let $\Lambda = \mu_1 + \mu_2 + \lambda_1 + \lambda_2 + \alpha$ and $y^+ = \max\{y, 0\}$ is the positive part of $y \in \mathbb{R}$. Then for $(q_1, q_2, \ell) \in S$, the dynamic programming optimality equations are given by

$$V_n(q_1, q_2, \ell) = \min_{a \in A}\{V_n[(q_1, q_2, \ell), a]\}, \tag{1}$$

$$V_n^{'}(q_1, q_2, \ell) = \min_{a \in A}\{V_n^{'}[(q_1, q_2, \ell), a]\}, \tag{2}$$

where for $\ell = 0$ and $a = 0$,

$$\begin{aligned} V_n[(q_1, q_2, 0), 0] = \Lambda^{-1}\big[h + \mu_2 V_{n-1}\big(q_1, (q_2 - 1)^+, 0\big) + \lambda_1 V_{n-1}(q_1 + 1, q_2, 0) \\ + \lambda_2 V_{n-1}(q_1, q_2 + 1, 0)\big], \end{aligned} \tag{3}$$

$$\begin{aligned} V_n^{'}[(q_1, q_2, 0), 0] = \Lambda^{-1}\Big[h + \mu_1 V_{n-1}^{'}\big((q_1 - 1)^+, q_2, 0\big) + \lambda_1 V_{n-1}^{'}(q_1 + 1, q_2, 0) \\ + \lambda_2 V_{n-1}^{'}(q_1, q_2 + 1, 0)\Big], \end{aligned} \tag{4}$$

for $\ell \geq 1$ and $a = 0$,

$$\begin{aligned} V_n[(q_1, q_2, \ell), 0] = \Lambda^{-1}\big[h + \mu_1 V_{n-1}(q_1 - 1, q_2, \ell - 1) + \mu_2 V_{n-1}\big(q_1, (q_2 - 1)^+, \ell\big) \\ + \lambda_1 V_{n-1}(q_1 + 1, q_2, \ell) + \lambda_2 V_{n-1}(q_1, q_2 + 1, \ell)\big], \end{aligned} \tag{5}$$

$$\begin{aligned} V_n^{'}[(q_1, q_2, \ell), 0] = \Lambda^{-1}\Big[h + \mu_1 V_{n-1}^{'}\big((q_1 - 1)^+, q_2, \ell\big) + \mu_2 V_{n-1}^{'}(q_1, q_2 - 1, \ell - 1) \\ + \lambda_1 V_{n-1}^{'}(q_1 + 1, q_2, \ell) + \lambda_2 V_{n-1}^{'}(q_1, q_2 + 1, \ell)\Big], \end{aligned} \tag{6}$$

and for $\ell \geq 0$ and $a = 1$,

$$V_n[(q_1, q_2, \ell), 1] = c_{12} + V_n^{'}[(q_1, q_2, q_2), 0], \tag{7}$$

$$V_n^{'}[(q_1, q_2, \ell), 1] = c_{21} + V_n[(q_1, q_2, q_1), 0]. \tag{8}$$

Equations (4), (6) and (8) correspond to equations (3), (5) and (7) for the states in which the strategic customer is in station 2. Clearly, V_n is a nonnegative function and, using an induction argument, it can be shown that $V_n(q_1, q_2, \ell)$ is nondecreasing in q_1, q_2 and ℓ. For convenience, we additionally introduce the following incremental cost of $V_n(q_1, q_2, \ell)$:

$$\Delta V_n(q_1, q_2, \ell) = V_n(q_1, q_2 + 1, \ell) - V_n(q_1, q_2, \ell).$$

That is, ΔV_n represents the incremental cost of adding one customer to station 2 when the strategic customer is in station 1.

Finally, define $V(q_1, q_2, \ell)$ and $V'(q_1, q_2, \ell)$ as the infinite-period costs corresponding to V_n and V'_n, respectively. Because $V_0(q_1, q_2, \ell) = V'_0(q_1, q_2, \ell) = 0$, by Proposition 3.1 of Ross [16, Chapter II, Section 3], we can assert that

$$V(q_1, q_2, \ell) = \lim_{n\to\infty} V_n(q_1, q_2, \ell),$$

$$V'(q_1, q_2, \ell) = \lim_{n\to\infty} V'_n(q_1, q_2, \ell).$$

Therefore, the results in Section 4 that hold for V_n and V'_n also hold for V and V'.

The next section sets out to establish the optimality of threshold jockeying and joining policies. That is, it will be proved that the strategic customers optimal behavior is dictated by monotone thresholds that are a function of the number of customers ahead of the strategic customer in the current queue, and the number of customers present in the other station.

4 Structure of the Optimal Policy

This section aims to characterize the optimal jockeying (and joining) policies by analyzing the optimal cost functions $V_n(q_1, q_2, \ell)$ and $V'_n(q_1, q_2, \ell)$. Specifically, Lemma 1 asserts several inequalities that are needed to establish the existence and monotonicity of an optimal threshold-type policy.

Lemma 1 *For all $n \in \mathbb{N}_0$, $q_1 \geq 0$, $q_2 \geq 0$, and $\ell \geq 0$, the following inequalities hold simultaneously:*

$$\Delta V_n(q_1, q_2, \ell) \leq \Delta V_n(q_1, q_2, \ell + 1), \quad (9)$$

$$\Delta V_n(q_1, q_2, q_1) \leq V'_n(q_1, q_2 + 1, q_2 + 1) - V'_n(q_1, q_2, q_2), \quad (10)$$

$$\Delta V_n(q_1, q_2, \ell) \leq V'_n(q_1, q_2 + 1, q_2 + 1) - V'_n(q_1, q_2, q_2), \quad (11)$$

$$V'_n(q_1, q_2 + 1, \ell + 2) - V'_n(q_1, q_2, \ell + 1) \leq V'_n(q_1, q_2 + 1, \ell + 1) - V'_n(q_1, q_2, \ell), \quad (12)$$

$$\Delta V_n(q_1, q_2, q_1) \leq V'_n(q_1, q_2 + 1, \ell + 1) - V'_n(q_1, q_2, \ell), \quad (13)$$

$$\Delta V_n(q_1, q_2 + 1, \ell) \leq V'_n(q_1, q_2 + 2, q_2 + 1) - V'_n(q_1, q_2 + 1, q_2), \quad (14)$$

$$V_{n+1}[(q_1, q_2, \ell), 1] - V_{n+1}[(q_1, q_2, \ell), 0] \leq V_{n+1}[(q_1, q_2 + 1, \ell), 1] - V_{n+1}[(q_1, q_2 + 1, \ell), 0], \quad (15)$$

$$V'_{n+1}[(q_1, q_2 + 1, \ell + 1), 1] - V'_{n+1}[(q_1, q_2 + 1, \ell + 1), 0] \leq V'_{n+1}[(q_1, q_2, \ell), 1] - V'_{n+1}[(q_1, q_2, \ell), 0], \quad (16)$$

$$V_{n+1}[(q_1, q_2, \ell + 1), 1] - V_{n+1}[(q_1, q_2, \ell + 1), 0] \leq V_{n+1}[(q_1, q_2, \ell), 1] - V_{n+1}[(q_1, q_2, \ell), 0]. \quad (17)$$

The proof of Lemma 1 is provided in the Appendix. Some useful insights are provided by a few of these inequalities. Let us first consider inequality (9), which asserts the supermodularity of $V_n(q_1, q_2, \ell)$ with respect to (q_2, ℓ). Suppose the strategic customer is in station 1 at the arrival time of a new customer in station 2. Then (9) asserts that the marginal increase in cost induced by the new arrival in station 2 is nondecreasing in the number of customers ahead of the strategic customer (ℓ). Furthermore, inequalities (15) and (17) show that $V_n[(q_1, q_2, \ell), 0] - V_n[(q_1, q_2, \ell), 1]$ is nonincreasing in q_2 and nondecreasing in ℓ. These facts will help establish that jockeying is only optimal if the other queue length is sufficiently short, or if the number of customers ahead of the strategic customer is sufficiently large. Specifically, it can be inferred from (15) and (17) that, if the strategic customer is in station 1 with ℓ customers in front of her, and q_1 and q_2 customers in stations 1 and 2, respectively (excluding herself), then the incentive to jockey from station 1 to 2 increases, all else being equal, as either ℓ increases or q_2 decreases. Theorem 1 formalizes these assertions.

Theorem 1 *Suppose the strategic customer is currently in station 1. Then the optimal jockeying policy is one of monotone thresholds. That is,*

(i) For given values of q_1 and ℓ, there exists a number $q^*(q_1, \ell)$ such that if $q_2 < q^*(q_1, \ell)$, then it is optimal for the customer to jockey to station 2. Otherwise, the customer should remain in station 1. Furthermore, the threshold $q^*(q_1, \ell)$ is nondecreasing in ℓ.

(ii) For given values of q_1 and q_2, there exists a number $\ell^*(q_1, q_2)$ such that if $\ell \geq \ell^*(q_1, q_2)$, then it is optimal for the customer to jockey to station 2. Otherwise, the customer should remain in station 1. Furthermore, the threshold $\ell^*(q_1, q_2)$ is nondecreasing in q_2.

Proof. To prove part (i), define

$$q^*(q_1, \ell) = \min_{q_2 \in \mathbb{N}_0} \{V_n[(q_1, q_2, \ell), 0] - V_n[(q_1, q_2, \ell), 1] \leq 0\} \tag{18}$$

If the set defined on the right-hand side of (18) is empty, then set $q^*(q_1, \ell) = -\infty$. By the definition of $q^*(q_1, \ell)$, for any state (q_1, q_2, ℓ) with $q_2 < q^*(q_1, \ell)$,

$$V_n[(q_1, q_2, \ell), 0] - V_n[(q_1, q_2, \ell), 1] > 0,$$

i.e., it is optimal to jockey to station 2. Furthermore, using inequality (15), it is seen that if $q_2 \geq q^*(q_1, \ell)$, then

$$\begin{aligned} V_n[(q_1, q_2, \ell), 0] - V_n[(q_1, q_2, \ell), 1] &\leq V_n[(q_1, q^*(q_1, \ell), \ell), 0] - V_n[(q_1, q^*(q_1, \ell), \ell), 1] \\ &\leq 0. \end{aligned}$$

Thus, the optimal jockeying policy is a threshold policy with respect to the number of customers in station 2. It remains to prove that the threshold is nondecreasing in ℓ. Based on (17), it is evident that

$$\begin{aligned} &V_n[(q_1, q^*(q_1, \ell+1), \ell), 0] - V_n[(q_1, q^*(q_1, \ell+1), \ell), 1] \\ &\qquad \leq V_n[(q_1, q^*(q_1, \ell+1), \ell+1), 0] - V_n[(q_1, q^*(q_1, \ell+1), \ell+1), 1] \\ &\qquad \leq 0. \end{aligned}$$

Therefore, we conclude that $q^*(q_1, \ell) \le q^*(q_1, \ell+1)$, i.e., the threshold is nondecreasing in ℓ.

Next, to prove (ii), define

$$\ell^*(q_1, q_2) = \min_{\ell \in \mathbb{N}_0} \left\{ V_n \left[(q_1, q_2, \ell), 0 \right] - V_n \left[(q_1, q_2, \ell), 1 \right] > 0 \right\}. \tag{19}$$

If the set defined on the right-hand side of (19) is empty, then set $\ell^*(q_1, q_2) = +\infty$. It follows from the definition of $\ell^*(q_1, q_2)$ that, for any state (q_1, q_2, ℓ) with $\ell < \ell^*(q_1, q_2)$,

$$V_n \left[(q_1, q_2, \ell), 1 \right] - V_n \left[(q_1, q_2, \ell), 0 \right] \ge 0,$$

i.e., it is optimal to remain in station 1. Conversely, if $\ell \ge \ell^*(q_1, q_2)$, then by (17) and (19), it follows that

$$\begin{aligned} V_n \left[(q_1, q_2, \ell), 1 \right] - V_n \left[(q_1, q_2, \ell), 0 \right] &\le V_n \left[(q_1, q_2, \ell^*(q_1, q_2)), 1 \right] - V_n \left[(q_1, q_2, \ell^*(q_1, q_2)), 0 \right] \\ &< 0, \end{aligned}$$

i.e., it is optimal to jockey to station 2. To show that $\ell^*(q_1, q_2)$ is nondecreasing in q_2, note that by inequality (15) and the definition of $\ell^*(q_1, q_2+1)$, we have

$$\begin{aligned} & V_n \left[(q_1, q_2, \ell^*(q_1, q_2+1)), 1 \right] - V_n \left[(q_1, q_2, \ell^*(q_1, q_2+1)), 0 \right] \\ & \qquad \le V_n \left[(q_1, q_2+1, \ell^*(q_1, q_2+1)), 1 \right] - V_n \left[(q_1, q_2+1, \ell^*(q_1, q_2+1)), 0 \right] \\ & \qquad < 0. \end{aligned}$$

By the definition of $\ell^*(q_1, q_2)$, it follows that $\ell^*(q_1, q_2) \le \ell^*(q_1, q_2+1)$. ■

Theorem 1 characterizes the structure of the optimal jockeying policy; however, it does not prescribe the optimal joining policy for the strategic customer. Specifically, which of the two queues should the customer join initially? Intuitively, if joining station 1 is preferred to joining queue 2, then joining station 1 is also preferred if the number in station 1 is smaller (or the number in station 2 is larger). Suppose that upon arrival of the strategic customer, the number in stations 1 and 2 are q_1 and q_2, respectively. If the customer joins station 1 or 2 and then follows the optimal jockeying policy of Theorem 1, the expected costs are, $V_n \left[(q_1, q_2, q_1), 0 \right]$ and $V_n^{'} \left[(q_1, q_2, q_2), 0 \right]$, respectively. Consequently, the customer joins station 1 if

$$V_n \left[(q_1, q_2, q_1), 0 \right] \le V_n^{'} \left[(q_1, q_2, q_2), 0 \right];$$

otherwise, the customer joins station 2. According to (7), we have

$$V_n \left[(q_1, q_2, q_1), 0 \right] - V_n^{'} \left[(q_1, q_2, q_2), 0 \right] = V_n \left[(q_1, q_2, q_1), 0 \right] - V_n \left[(q_1, q_2, q_1), 1 \right] + c_{12}. \tag{20}$$

On the basis of (15), and due to the problem's symmetry, equation (20) is nonincreasing in q_2 and nondecreasing in q_1. Using an approach similar to that used in the proof of Theorem 1, we next establish the optimality of a threshold-type joining policy. The result is formalized in Theorem 2.

Theorem 2 *Suppose that, upon arrival, the strategic customer observes q_1 and q_2 customers in stations 1 and 2, respectively. Then, for a given q_2, there is a number $\varepsilon(q_2)$ such that if $q_1 \ge \varepsilon(q_2)$, it is optimal to initially join station 2; otherwise, the customer should join station 1. Moreover, the threshold $\varepsilon(q_2)$ is monotone nondecreasing in q_2.*

Before concluding this section, we pause here briefly to comment on models involving three or more parallel stations. For such cases, we conjecture that the optimal jockeying and joining policies will mirror those described in Theorems 1 and 2, respectively. Consider the case of three parallel stations. Let q_i denote the number of customers in station i, excluding the strategic customer, and suppose the strategic customer is located in station 1. The state is described by the four-dimensional vector (q_1, q_2, q_3, ℓ). We conjecture that, for given values (q_1, q_3, ℓ), there exists a number $q_2^*(q_1, q_3, \ell)$ such that jockeying to station 2 is optimal if and only if its queue length is below this number. Similarly, for given values (q_1, q_2, ℓ), there exists a number $q_3^*(q_1, q_2, \ell)$ so that it is optimal to jockey to station 3 if its queue length is below this threshold. With regard to the number of customers ahead of the strategic customer, we conjecture that, for given values (q_1, q_2, q_3), there exists a number $\ell^*(q_1, q_2, q_3)$ such that jockeying is optimal if and only if $\ell \geq \ell^*(q_1, q_2, q_3)$. For joining, we conjecture that it is optimal to initially join station j if and only if $q_j \leq \varepsilon(q_i, q_k)$ for $i, k \neq j$. The verification of these conjectures is left for future work.

Finally, we conclude this section by summarizing insights gained from Lemma 1 and Theorems 1 and 2. If the strategic customer is in station 1, Lemma 1 characterizes how the strategic customer should react to arrivals or departures at station 2. Inequality (15) asserts that the arrival (departure) of a customer at station 2 may diminish (enhance) the strategic customer's incentive to jockey to station 2. Furthermore, by inequality (16), service completions at station 1 incentivize the strategic customer to remain in station 1. However, these results do not characterize how the strategic customer should behave when new customers join station 1 (behind the customer). Intuitively, one expects the strategic customer to be indifferent to these arrivals. On the other hand, it may be the case that the strategic customer is interested in maintaining the current position in station 1 rather than forfeiting this position by jockeying to station 2. The current model does not provide a definitive answer for the strategic customer's behavior under these circumstances, so it will be instructive to pursue such a characterization in future work. The next section demonstrates that the optimal policies described here carry over to the total cost criterion in a natural way.

5 Total (Undiscounted) Cost Criterion

Here, we discuss the structure of the optimal jockeying and joining policies under the total cost criterion (i.e., when the costs are not discounted). For the sake of brevity, we provide only a sketch of the analysis. This section emphasizes the dependence of the optimal cost functions V_n and V_n' on the discount factor α by writing $V_n^\alpha(q_1, q_2, \ell)$ and $V_n'^\alpha(q_1, q_2, \ell)$, respectively. Now, denote by $V_n^0(q_1, q_2, \ell)$ (respectively $V_n'^0(q_1, q_2, \ell)$), the minimum total expected cost with n periods remaining in the time horizon (cf. Ross [16, Chapter III] for a formal definition). It is clear that, for each $n \geq 0$, $V_n^0(q_1, q_2, \ell)$ and $V_n'^0(q_1, q_2, \ell)$ are well-defined with $\alpha = 0$ in equations (1) through (8). Moreover, the proof of Lemma 1 is valid with $\alpha = 0$; therefore, Theorems 1 and 2 are valid under the total expected cost criterion. Additionally, if we consider the infinite-horizon version of the problem under the total expected cost criterion, an optimal stationary policy exists by Theorems 1.1 and 1.2 of Ross [16, Chapter III, Section 1] since the jockeying and holding costs (c_{ij} and h, respectively) are non-negative and bounded. Denoting the infinite-horizon optimal cost function by $V^0(q_1, q_2, \ell)$, it follows that for each $(q_1, q_2, \ell) \in S$,

$$V^0(q_1, q_2, \ell) = \lim_{n \to \infty} V_n^0(q_1, q_2, \ell) \quad \text{and} \quad V'^0(q_1, q_2, \ell) = \lim_{n \to \infty} V_n'^0(q_1, q_2, \ell)$$

by Theorem 4.2 of Ross [16, Chapter III, Section 4]. Hence, the monotone threshold structures proved in Theorems 1 and 2 hold under the total expected cost criterion over an infinite horizon as well.

6 Numerical Examples

This section illustrates the main results of Section 4 and reveals that individually optimal joining and/or jockeying policies may be somewhat counterintuitive. Specifically, for certain parameter values, it may not be optimal for the strategic customer to join the shortest queue upon arrival, or to join the station with the fastest server.

Example 1. Consider a two-station, parallel queueing system in which customers arrive according to a Poisson process with rate $\lambda = 8$. Normal customers join station 1 (station 2) with probability 0.5 (0.5), so the arrival rate to station 1 (station 2) is $\lambda_1 = 4$ ($\lambda_2 = 4$). Station 1 (Station 2) serves customers in an exponentially distributed amount of time with parameter $\mu_1 = 8$ ($\mu_2 = 7.5$). Therefore, the traffic intensities at stations 1 and 2 are $\rho_1 = 0.500$ and $\rho_2 = 0.533$, respectively. The holding cost per unit time is $h = 1$, and the jockeying costs are, respectively, $c_{12} = 0.15$ and $c_{21} = 0.002$. The discount factor is $\alpha = 1$, and to illustrate the optimal jockeying policy, we fix the number ahead of the strategic customer at $\ell = 9$.

Figure 2 graphically depicts the optimal joining policy for the strategic customer. In this example, if the strategic customer arrives to find station 1 empty, then this station should be joined, irrespective of the number in station 2. However, the policy is not intuitive for some states in which $q_1 > 0$. For instance, when $q_1 = 4$ and $q_2 = 5$, it is optimal for the strategic customer to initially join station 2, despite the fact that it is more congested and has a slower server. Therefore, when $(q_1, q_2) = (4, 5)$, we observe that joining the station with the longer queue and the slower server is optimal. This may be attributed to the fact that the fixed cost of jockeying from station 2 to station 1 (c_{21}) is relatively small as compared to the holding cost rate h and the jockeying cost c_{12}.

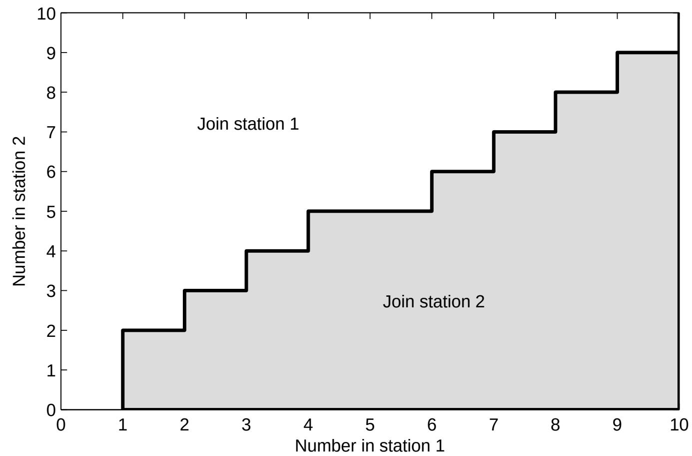

Figure 2: Optimal joining policy in Example 1.

Figure 3 illustrates the optimal jockeying policy when the strategic customer is in station 1 and $\ell = 9$. We fix the number of customers ahead of the strategic customer in order to graphically depict the optimal jockeying policy in two dimensions; hence, values of q_1 less than 9 are not displayed. Given that the strategic customer is in station 1 with $\ell = 9$ customers ahead of her, she should jockey to station 2 if $q_2 \leq 2$. The optimal policy in this example reveals that the incentive to jockey from station 1 to station 2 is not strong. This may be the case because: (a) jockeying from station 1 to 2 is relatively costly, and (b) station 1 possesses a faster server; therefore, switching to station 2 is not justified unless its queue length is relatively small.

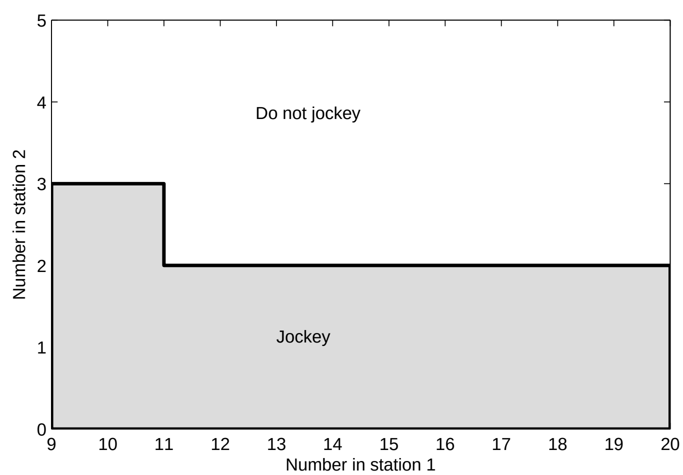

Figure 3: Optimal jockeying policy in Example 1 ($\ell = 9$).

Example 2. Now consider the same system with Poisson arrival rate $\lambda = 14$. Arriving customers join station 1 (station 2) with probability 5/14 (9/14), so the corresponding arrival rate is $\lambda_1 = 5$ ($\lambda_2 = 9$). The station 1 (station 2) service rate is $\mu_1 = 20$ ($\mu_2 = 10$); therefore, the traffic intensities at stations 1 and 2 are $\rho_1 = 0.25$ and $\rho_2 = 0.90$, respectively. The holding cost per unit time is $h = 1$, and the jockeying costs are $c_{12} = 0.005$ and $c_{21} = 0.001$. The discount factor is $\alpha = 1$, and we fix $\ell = 12$ to illustrate the jockeying policy in Figure 5.

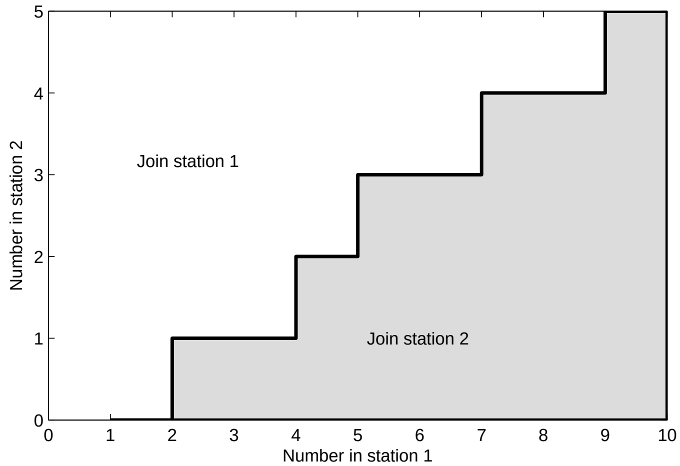

Figure 4: Optimal joining policy in Example 2.

Figure 4 graphically depicts the optimal joining policy for the strategic customer. As in the first example, it is optimal to join station 1 if it is empty upon arrival, irrespective of the number of customers in station 2. However, it may be optimal to join station 1 even when it is more congested than station 2, e.g. when $(q_1, q_2) = (6, 4)$. This behavior may be attributed to the fact that the service rate in station 1 is twice that of station 2 and the holding cost is relatively high.

Finally, the optimal jockeying policy of Example 2 is illustrated in Figure 5. It is interesting to note that, when $\ell = 12$, the strategic customer should always jockey to station 2 if $q_2 \leq 4$, despite the fact that $\mu_1 = 2\mu_2$. However, if the number in station 2 is at least as high as 7, then the strategic customer is better off maintaining her current position in station 1. This example serves to illustrate the point that it may not be optimal to remain at the station with the fastest server if the other queue length is sufficiently short.

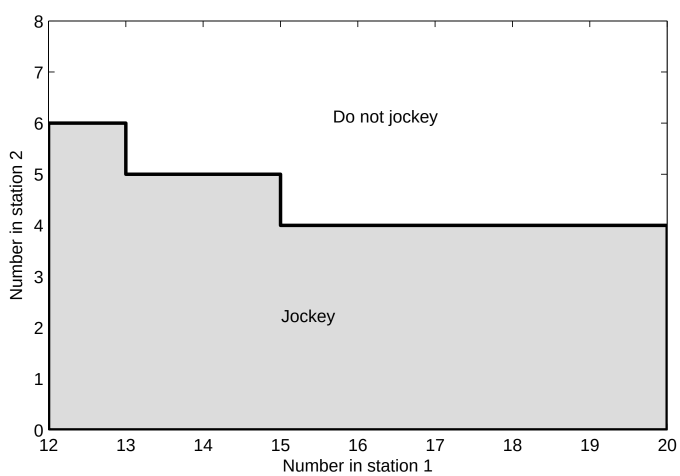

Figure 5: Optimal jockeying policy in Example 2 ($\ell = 12$).

Appendix: Proof of Lemma 1

Before proceeding to the proof of Lemma 1, we first state Lemma 2, which will be used in the proof that follows.

Lemma 2 *If $a_1, a_2, b_1, b_2 \in \mathbb{R}$ and $a_2 - a_1 \le b_2 - b_1$, then,*

$$(a_2 \wedge b_2) - (a_1 \wedge b_1) \le b_2 - b_1,$$

where $a \wedge b = \min\{a, b\}$.

Proof. (Lemma 1). The lemma will be proved using an induction argument. First note that inequalities (9)–(17) certainly hold for $n = 0$. For the induction hypothesis, assume they hold for some $n - 1 > 0$. To prove (9), by equation (1), we need to establish that

$$\begin{aligned}\min\{V_n[(q_1, q_2+1, \ell), 0], V_n[(q_1, q_2+1, \ell), 1]\} - \min\{V_n[(q_1, q_2, \ell), 0], V_n[(q_1, q_2, \ell), 1]\} \\ \le \min\{V_n[(q_1, q_2+1, \ell+1), 0], V_n[(q_1, q_2+1, \ell+1), 1]\} \\ - \min\{V_n[(q_1, q_2, \ell+1), 0], V_n[(q_1, q_2, \ell+1), 1]\} \quad (21)\end{aligned}$$

We focus on combinations of optimal actions in the right-hand side (r.h.s.) of inequality (21). Specifically, an optimal combination of the form (a, b) implies that action a is optimal in the first minimum on the r.h.s., and action b is optimal in the second minimum. For example, if the optimal combination on the r.h.s. of (21) is $(0, 0)$, then the minimum in the first expression is $V_n[(q_1, q_2+1, \ell+1), 0]$, and the minimum in the second expression of the r.h.s. is $V_n[(q_1, q_2, \ell+1), 0]$. We employ this convention throughout the remainder of the proof.

Case (9a): The optimal combination of the r.h.s. is $(0, 0)$. By the induction hypothesis for (17), the optimal combination of the left-hand side (l.h.s.) of inequality (21) is $(0, 0)$. By expanding $V_n[(q_1, q_2+1, \ell), 0]$, $V_n[(q_1, q_2, \ell), 0]$, $V_n[(q_1, q_2+1, \ell+1), 0]$, and $V_n[(q_1, q_2, \ell+1), 0]$ using equations (3) and (5), and applying the induction hypothesis for (9), it follows that inequality (21) holds for this case.

Case (9b): The optimal combination of the r.h.s. is $(0, 1)$. Using the induction hypothesis for (17), the optimal combination of the l.h.s. could be $(0, 0)$ or $(0, 1)$. By the induction hypothesis on (9), it is seen that

$$\begin{aligned}V_n[(q_1, q_2+1, \ell), 0] - V_n[(q_1, q_2, \ell), 0] \le V_n[(q_1, q_2+1, \ell+1), 0] - V_n[(q_1, q_2, \ell+1), 0] \\ \le V_n[(q_1, q_2+1, \ell+1), 0] - \min\{V_n[(q_1, q_2, \ell+1), 0], V_n[(q_1, q_2, \ell+1), 1]\} \\ \le V_n[(q_1, q_2+1, \ell+1), 0] - V_n[(q_1, q_2, \ell+1), 1], \quad (22)\end{aligned}$$

and it follows from (7), and the fact that $V_n[(q_1, q_2+1, \ell), 0]$ is nondecreasing in ℓ, that

$$V_n[(q_1, q_2+1, \ell), 0] - V_n[(q_1, q_2, \ell), 1] \le V_n[(q_1, q_2+1, \ell+1), 0] - V_n[(q_1, q_2, \ell+1), 1]. \quad (23)$$

Combining inequalities (22) and (23), we conclude that

$$\begin{aligned}V_n[(q_1, q_2+1, \ell), 0] - \min\{V_n[(q_1, q_2, \ell), 0], V_n[(q_1, q_2, \ell), 1]\} \\ \le V_n[(q_1, q_2+1, \ell+1), 0] - V_n[(q_1, q_2, \ell+1), 1].\end{aligned}$$

Case (9c): The optimal combination of the r.h.s. is $(1,1)$. Using Lemma 2 and the induction hypothesis for inequality (15), we have

$$\begin{aligned}\min\left\{V_n\left[\left(q_1,q_2+1,\ell\right),0\right],V_n\left[\left(q_1,q_2+1,\ell\right),1\right]\right\}-\min\left\{V_n\left[\left(q_1,q_2,\ell\right),0\right],V_n\left[\left(q_1,q_2,\ell\right),1\right]\right\}\\ \le V_n\left[\left(q_1,q_2+1,\ell\right),1\right]-V_n\left[\left(q_1,q_2,\ell\right),1\right]\\ =V_n\left[\left(q_1,q_2+1,\ell+1\right),1\right]-V_n\left[\left(q_1,q_2,\ell+1\right),1\right],\end{aligned}$$

where the last equality follows from (7).

Case (9d): The optimal combination of the r.h.s. is $(1,0)$. This combination is not feasible by the induction hypothesis of (15). Therefore, the proof of inequality (9) is hereby complete.

Next, to establish inequality (10), it must be shown by equations (1) and (2) that

$$\begin{aligned}\min\left\{V_n\left[\left(q_1,q_2+1,q_1\right),0\right],V_n\left[\left(q_1,q_2+1,q_1\right),1\right]\right\}-\min\left\{V_n\left[\left(q_1,q_2,q_1\right),0\right],V_n\left[\left(q_1,q_2,q_1\right),1\right]\right\}\\ \le\min\left\{V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),0\right],V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),1\right]\right\}\\ -\min\left\{V_n^{'}\left[\left(q_1,q_2,q_2\right),0\right],V_n^{'}\left[\left(q_1,q_2,q_2\right),1\right]\right\}.\end{aligned}\tag{24}$$

Similar to the proof of inequality (21), we examine optimal combinations on the r.h.s of (24).

Case (10a): The optimal combination of the r.h.s. is (0,0). By the induction hypothesis for (15) and Lemma 2, we have

$$\begin{aligned}\min\left\{V_n\left[\left(q_1,q_2+1,q_1\right),0\right],V_n\left[\left(q_1,q_2+1,q_1\right),1\right]\right\}-\min\left\{V_n\left[\left(q_1,q_2,q_1\right),0\right],V_n\left[\left(q_1,q_2,q_1\right),1\right]\right\}\\ \le V_n\left[\left(q_1,q_2+1,q_1\right),1\right]-V_n\left[\left(q_1,q_2,q_1\right),1\right]\\ =V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),0\right]-V_n^{'}\left[\left(q_1,q_2,q_2\right),0\right],\end{aligned}$$

where the last equality follows from (7).

Case (10b): The optimal combination of the r.h.s. is (0,1). Such a combination is not feasible since, by the induction hypothesis for (16),

$$V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),0\right]-V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),1\right]\ \ge\ V_n^{'}\left[\left(q_1,q_2,q_2\right),0\right]-V_n^{'}\left[\left(q_1,q_2,q_2\right),1\right].$$

Case (10c): The optimal combination of the r.h.s. is (1,1). As a consequence of equations (7) and (8), and the fact that the jockeying costs c_{12} and c_{21} are non-negative, the optimal combination of the l.h.s. of inequality (24) must be (0,0). By (8), it follows that

$$\begin{aligned}V_n\left[\left(q_1,q_2+1,q_1\right),0\right]-V_n\left[\left(q_1,q_2,q_1\right),0\right]\\ =\left(V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),1\right]-c_{21}\right)-\left(V_n^{'}\left[\left(q_1,q_2,q_2\right),1\right]-c_{21}\right)\\ =V_n^{'}\left[\left(q_1,q_2+1,q_2+1\right),1\right]-V_n^{'}\left[\left(q_1,q_2,q_2\right),1\right].\end{aligned}$$

Case (10d): The optimal combination of the r.h.s. is (1,0). By equations (7) and (8), the optimal combination of the l.h.s of inequality (24) could be $(0,0)$ or $(0,1)$. Moreover, it follows from (8) that

$$\begin{aligned} V_n\left[\left(q_1, q_2+1, q_1\right), 0\right] - V_n\left[\left(q_1, q_2, q_1\right), 0\right] &= V_n^{'}\left[\left(q_1, q_2+1, q_2+1\right), 1\right] - V_n^{'}\left[\left(q_1, q_2, q_2\right), 1\right] \\ &\leq V_n^{'}\left[\left(q_1, q_2+1, q_2+1\right), 1\right] - V_n^{'}\left[\left(q_1, q_2, q_2\right), 0\right], \end{aligned} \tag{25}$$

and

$$\begin{aligned} V_n\left[\left(q_1, q_2+1, q_1\right), 0\right] &- V_n\left[\left(q_1, q_2, q_1\right), 1\right] \\ &= \left(V_n^{'}\left[\left(q_1, q_2+1, q_2+1\right), 1\right] - c_{21}\right) - \left(V_n^{'}\left[\left(q_1, q_2, q_2\right), 0\right] + c_{12}\right) \\ &\leq V_n^{'}\left[\left(q_1, q_2+1, q_2+1\right), 1\right] - V_n^{'}\left[\left(q_1, q_2, q_2\right), 0\right], \end{aligned} \tag{26}$$

where the last inequality follows from the fact that jockeying costs are non-negative. Inequalities (25) and (26) imply that

$$\begin{aligned} V_n\left[\left(q_1, q_2+1, q_1\right), 0\right] - \min\left\{V_n\left[\left(q_1, q_2, q_1\right), 0\right], V_n\left[\left(q_1, q_2, q_1\right), 1\right]\right\} \\ \leq V_n^{'}\left[\left(q_1, q_2+1, q_2+1\right), 1\right] - V_n^{'}\left[\left(q_1, q_2, q_2\right), 0\right], \end{aligned}$$

and thus, inequality (10) is established.

We now establish inequality (11). By inequality (9), and using the fact that $\ell \leq q_1$, it follows that

$$\Delta V_n\left(q_1, q_2, \ell\right) \leq \Delta V_n\left(q_1, q_2, q_1\right), \tag{27}$$

and by inequality (10), it follows that

$$\Delta V_n\left(q_1, q_2, q_1\right) \leq V_n^{'}\left(q_1, q_2+1, q_2+1\right) - V_n^{'}\left(q_1, q_2, q_2\right). \tag{28}$$

The result follows immediately by comparing inequalities (27) and (28).

To prove inequality (12), by equation (2), it must be shown that

$$\begin{aligned} \min\Big\{V_n^{'}\left[\left(q_1, q_2+1, \ell+2\right), 0\right], V_n^{'}\left[\left(q_1, q_2+1, \ell+2\right), 1\right]\Big\} \\ - \min\Big\{V_n^{'}\left[\left(q_1, q_2, \ell+1\right), 0\right], V_n^{'}\left[\left(q_1, q_2, \ell+1\right), 1\right]\Big\} \\ \leq \min\Big\{V_n^{'}\left[\left(q_1, q_2+1, \ell+1\right), 0\right], V_n^{'}\left[\left(q_1, q_2+1, \ell+1\right), 1\right]\Big\} \\ - \min\Big\{V_n^{'}\left[\left(q_1, q_2, \ell\right), 0\right], V_n^{'}\left[\left(q_1, q_2, \ell\right), 1\right]\Big\}. \end{aligned} \tag{29}$$

Similar to the proof for inequality (21), we examine optimal combinations of the r.h.s of (29).

Case (12a): The optimal combination of the r.h.s is (0,0). Using Lemma 2 and the induction hypothesis for (16), we have

$$\min\left\{V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),0\right],V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),1\right]\right\} - \min\left\{V_n^{'}\left[\left(q_1,q_2,\ell+1\right),0\right],V_n^{'}\left[\left(q_1,q_2,\ell+1\right),1\right]\right\} \le V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),0\right] - V_n^{'}\left[\left(q_1,q_2,\ell+1\right),0\right]. \quad (30)$$

Moreover,

$$V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),0\right] - V_n^{'}\left[\left(q_1,q_2,\ell+1\right),0\right] \le V_n^{'}\left[\left(q_1,q_2+1,\ell+1\right),0\right] - V_n^{'}\left[\left(q_1,q_2,\ell\right),0\right], \quad (31)$$

where we have made use of the expansion of $V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),0\right]$, $V_n^{'}\left[\left(q_1,q_2,\ell+1\right),0\right]$, $V_n^{'}\left[\left(q_1,q_2+1,\ell+1\right),0\right]$, and $V_n^{'}\left[\left(q_1,q_2,\ell\right),0\right]$ by using equations (4) and (6), and the induction hypothesis for (12). The result follows by comparing inequalities (30) and (31).

Case (12b): The optimal combination of the r.h.s is (0,1). This combination is not feasible due to the induction hypothesis for inequality (16).

Case (12c): The optimal combination of the r.h.s could be $(1,0)$ or $(1,1)$. By the induction hypothesis (17), the possible optimal combinations of the l.h.s of inequality (29) are $(1,0)$ and $(1,1)$, and by equation (7), we have

$$V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),1\right] = V_n^{'}\left[\left(q_1,q_2+1,\ell+1\right),1\right]. \quad (32)$$

Note that $V_n^{'}\left(q_1,q_2,\ell\right)$ is nondecreasing in ℓ, i.e.,

$$V_n^{'}\left(q_1,q_2,\ell\right) \le V_n^{'}\left(q_1,q_2,\ell+1\right). \quad (33)$$

Equation (32) and inequality (33) imply that

$$V_n^{'}\left[\left(q_1,q_2+1,\ell+2\right),1\right] - V_n^{'}\left(q_1,q_2,\ell+1\right) \le V_n^{'}\left[\left(q_1,q_2+1,\ell+1\right),1\right] - V_n^{'}\left(q_1,q_2,\ell\right);$$

therefore, the proof of inequality (12) is hereby complete.

Now, we establish inequality (13). By inequality (12), and the fact that $\ell \le q_2$, we have

$$V_n^{'}\left(q_1,q_2+1,q_2+1\right) - V_n^{'}\left(q_1,q_2,q_2\right) \le V_n^{'}\left(q_1,q_2+1,\ell+1\right) - V_n^{'}\left(q_1,q_2,\ell\right). \quad (34)$$

Moreover, it follows from inequality (10), that

$$\Delta V_n\left(q_1,q_2,q_1\right) \le V_n^{'}\left(q_1,q_2+1,q_2+1\right) - V_n^{'}\left(q_1,q_2,q_2\right). \quad (35)$$

The result follows immediately by comparing inequalities (34) and (35).

Next, we establish inequality (14). By inequality (11), we have

$$\Delta V_n\left(q_1,q_2+1,\ell\right) \le V_n^{'}\left(q_1,q_2+2,q_2+2\right) - V_n^{'}\left(q_1,q_2+1,q_2+1\right), \quad (36)$$

and by inequality (12), we have

$$V_n^{'}(q_1, q_2+2, q_2+2) - V_n^{'}(q_1, q_2+1, q_2+1) \leq V_n^{'}(q_1, q_2+2, q_2+1) - V_n^{'}(q_1, q_2+1, q_2). \quad (37)$$

Therefore, the result is proved.

Next, we will establish inequality (15). For this part of the proof, suppose that $\ell \geq 1$ and $q_2 \geq 1$. For either $\ell = 0$ or $q_2 = 0$, the proof technique is similar, except that in some places, we appeal to the nonnegativity of V_n in lieu of the induction hypothesis. Due to inequality (11), and the fact that μ_1, μ_2, λ_1 and λ_2 are all positive, the following inequalities hold:

$$\mu_1 \Delta V_n(q_1-1, q_2, \ell-1) \leq \mu_1 \left(V_n^{'}(q_1-1, q_2+1, q_2+1) - V_n^{'}(q_1-1, q_2, q_2)\right), \quad (38)$$

$$\mu_2 \Delta V_n(q_1, q_2-1, \ell) \leq \mu_2 \left(V_n^{'}(q_1, q_2, q_2) - V_n^{'}(q_1, q_2-1, q_2-1)\right), \quad (39)$$

$$\lambda_1 \Delta V_n(q_1+1, q_2, \ell) \leq \lambda_1 \left(V_n^{'}(q_1+1, q_2+1, q_2+1) - V_n^{'}(q_1+1, q_2, q_2)\right). \quad (40)$$

Moreover, by inequality (14), we have

$$\lambda_2 \Delta V_n(q_1, q_2+1, \ell) \leq \lambda_2 \left(V_n^{'}(q_1, q_2+2, q_2+1) - V_n^{'}(q_1, q_2+1, q_2)\right). \quad (41)$$

The result follows by summing the right- and left-hand sides of (38)–(41) and simplifying each side using their equivalents in equations (5) through (8).

The proof of inequality (16) mirrors that of (15). Assume that $\ell \geq 1$ and $q_2 \geq 1$. It follows from (13) that

$$\mu_1 \Delta V_n(q_1-1, q_2, q_1-1) \leq \mu_1 \left(V_n^{'}(q_1-1, q_2+1, \ell+1) - V_n^{'}(q_1-1, q_2, \ell)\right), \quad (42)$$

$$\mu_2 \Delta V_n(q_1, q_2-1, q_1) \leq \mu_2 \left(V_n^{'}(q_1, q_2, \ell) - V_n^{'}(q_1, q_2-1, \ell-1)\right), \quad (43)$$

$$\lambda_2 \Delta V_n(q_1, q_2+1, q_1) \leq \lambda_2 \left(V_n^{'}(q_1, q_2+2, \ell+1) - V_n^{'}(q_1, q_2+1, \ell)\right). \quad (44)$$

Moreover, inequalities (9) and (13) imply that

$$\lambda_1 \Delta V_n(q_1+1, q_2, q_1) \leq \lambda_1 \left(V_n^{'}(q_1+1, q_2+1, \ell+1) - V_n^{'}(q_1+1, q_2, \ell)\right). \quad (45)$$

By summing inequalities (42)–(45) and replacing the l.h.s and r.h.s of the derived inequality by their equivalents in equations (5) through (8), the result is derived.

Finally, we establish inequality (17). By equation (7), it is clear that

$$V_{n+1}\left[(q_1, q_2, \ell), 1\right] = V_{n+1}\left[(q_1, q_2, \ell+1), 1\right]. \quad (46)$$

Note that V_n is nondecreasing in ℓ, i.e.,

$$V_{n+1}\left[(q_1, q_2, \ell), 0\right] \leq V_{n+1}\left[(q_1, q_2, \ell+1), 0\right]. \quad (47)$$

Therefore, inequality (17) follows directly from (46) and (47). ■

Acknowledgements. The authors are grateful to an anonymous referee and Professor Sheldon Ross for insightful questions and constructive comments that have improved this work.

References

[1] I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. Matrix-geometric analysis of the shortest queue problem with threshold jockeying. *Operations Research Letters*, 13:107–112, 1993.

[2] R.L. Disney and W.E. Mitchell. A solution for queues with instantaneous jockeying and other customer selection rules. *Naval Research Logistics Quarterly*, 17(3):315–325, 1970.

[3] D.G. Down and M.E. Lewis. Dynamic load balancing in parallel queueing systems: Stability and optimal control. *European Journal of Operational Research*, 168(2):509–519, 2006.

[4] E.A. Elsayed and A. Bastani. General solutions of the jockeying problem. *European Journal of Operational Research*, 22(3):387–396, 1985.

[5] I. Gertsbakh. The shorter queue problem: A numerical study using matrix-geometric solution. *European Journal of Operational Research*, 15:374–381, 1984.

[6] A. Glazer and R. Hassin. Search among queues. Technical Report 406, Institute for Mathematical Studies in the Social Sciences, Stanford University, Palo Alto, CA, 1983.

[7] F.A. Haight. Two queues in parallel. *Biometrika*, 45(3/4):401–410, 1958.

[8] R. Hassin and M. Haviv. Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Communications in Statistics: Stochastic Models*, 10(2):415–435, 1994.

[9] R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA, 2003.

[10] M. Hlynka, D.A. Stanford, W.H. Poon, and T. Wang. Observing queues before joining. *Operations Research*, 42(2):365–371, 1994.

[11] E.P.C. Kao and C. Lin. A matrix-geometric solution of the jockeying problem. *European Journal of Operational Research*, 44(1):67–74, 1990.

[12] J.F.C. Kingman. Two similar queues in parallel. *Annals of Mathematical Statistics*, 32:1314–1323, 1961.

[13] E. Koenigsberg. On jockeying in queues. *Management Science*, 12(5):412–436, 1966.

[14] S.A. Lippman and S. Stidham. Individual versus social optimization in exponential congestion systems. *Operations Research*, 25(2):233–247, 1977.

[15] A. Mandelbaum and U. Yechiali. Optimal entering rules for a customer with wait option at an M/G/1 queue. *Management Science*, 29(2):174–187, 1983.

[16] S.M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, Inc., New York, NY, 1983.

[17] A.M.K. Tarabia. Analysis of two queues in parallel with jockeying and restricted capacities. *Applied Mathematical Modelling*, 32(5):802–810, 2008.

[18] R.R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15(2):406–413, 1978.

[19] W.O. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34(1):55–62, 1986.

[20] S.H. Xu and Y.Q. Zhao. Dynamic routing and jockeying controls in a two-station queueing system. *Advances in Applied Probability*, 28(4):1201–1226, 1996.

[21] Y. Zhao and W.K. Grassmann. The shortest queue model with jockeying. *Naval Research Logistics*, 37(5):773–787, 1990.

[22] Y. Zhao and W.K. Grassmann. Queueing analysis of a jockeying model. *Operations Research*, 43(3):520–529, 1995.