User optimal policies for a stochastic transportation network

Heti Afimeimounga

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Statistics,

The University of Auckland, 2011

Abstract

This thesis studies user optimal policies (or user equilibria) for routing in queueing networks, with particular emphasis on networks where the Downs-Thomson paradox appears. We consider here a queueing network with two modes of transport – private (road) transportation and public (train) transportation – which are modelled as a single server first-in-first-out $\cdot|M|1$ queue and infinite server batch-service $\cdot|G^{(N)}|\infty$ queue respectively, where N is the size of the batch. The waiting time at the single server queue increases as traffic increases while the service frequency at the batch-service queue increases with demand, and hence waiting time there decreases. User equilibria or user optimal policies may occur when users in a network individually choose the best (shortest) route for themselves. The Downs-Thomson paradox occurs when improvements in service produce an overall *decline* in performance as user equilibria adjust.

We study this network system in the setting of both state-dependent routing and probabilistic routing. We use coupling arguments to show that under state-dependent routing a unique user optimal policy exists and is monotone. We also analyze the user equilibria (user optimal policies) in the setting of probabilistic routing when the road is modelled as an $M|G|1$ queue instead of

an $M|M|1$ queue. This analysis shows that for some parameter values there exist multiple equilibria (not all stable) for $N \geq 2$. Furthermore, we analyze the properties of the user optimal policies for the same network with $M|M|1$ queue versus $M|M^{(N)}|1$ queue in the settings of both state-dependent routing and probabilistic routing.

Acknowledgments

It is a great pleasure to thank those who made this thesis possible.

First, I would like to express my deep and sincere gratitude to my supervisor, Dr. Ilze Ziedins, for her encouragement, guidance and support all through from the initial stage to this final stage. I would also like to thank my advisor, Dr. Wiremu Solomon for his support and help, and to an anonymous referee for suggesting the idea that led to the result on page 75.

Finally, I would like to thank the Department of Statistics and The New Zealand Federation of Graduate Women for their support, and to my friends and family for all their love, encouragement and support.

Contents

Chapter 1

Introduction

User optimal policies or user equilibria have been studied in a variety of queueing networks including transportation networks, both in the contexts of state-dependent routing and probabilistic routing. A user optimal policy occurs when users in a network individually choose the best (shortest) route for themselves regardless of its negative effect on other users. This phenomenon is also known as selfish routing or individually optimal routing (see, for example, Arnott and Small [6], Roughgarden [53, 54] and Correa, Schulz and Stier-Moses [18]). It has long been known that, in general, delays in networks where users choose their own routes through the system may be considerably greater than in systems where a system coordinator directs users to the system optimal routes (see, for example, Naor [47], Hassin and Haviv [27], Correa, Schulz and Stier-Moses [18], Calvert [15], Cohen and Kelly [17], Whitt [69], Bell and Stidham [9]) and Ziedins [74]). This thesis studies the user optimal policy and the Downs-Thomson paradox for a network where users have a choice of two routes available to them, and wish to minimize the delay incurred when trav-

elling from a source to a destination (see Figure 1.1). One route is modelled as a first-come-first-served (FCFS) single server queue and the other route as a batch service queue. This network model is motivated by the idea of comparing private with public transportation and it was first proposed by Calvert [15] as a stochastic version of a widely-studied transportation model (see e.g. Mogridge [44], Arnott and Small [6] and Abraham and Hunt [1]). The FCFS single server queue represents private transportation where delays are increasing as the load increases. The batch service queue represents public transportation such as small shuttle buses awaiting passengers at airports or a train, which wait until they are full and then depart. In transportation networks the Downs-Thomson paradox refers to an increase in the expected delay in the system as users shift from public to private transportation when the capacity of private transport is increased.

We consider here two ways for individual users to choose their mode of transport. The first is when individual users have full knowledge of the state of the system and make their routing decisions based on that knowledge. We refer to this route choice as state-dependent routing. The second is when the users choose their routes without knowing the instantaneous state of the system, that is, they choose a route probabilistically. We refer to this route choice as state-independent (probabilistic) routing. Furthermore, we assume for this routing case that users know the arrival rates into the system and service rates in each queue when choosing their route. We then seek routing policies with the property that no user has any incentive to deviate from such a policy if all others follow it. Such policies are variously known as user optimal policies, user equilibria, Wardrop [68], user-determined or symmetric equilibria. Under state-dependent routing we show that for the model we consider here such

policies exist and that they are unique and monotone in structure. We give numerical examples of the existence of multiple user equilibria under probabilistic routing and also numerical examples that show that state-dependent routing mitigates the Downs-Thomson effect observed under probabilistic routing for the system model we consider here. State-dependent routing is also known as dynamic routing and probabilistic routing as static routing (see [62]). We will use interchangeably the terms "user optimal policy" and "user equilibrium" throughout this thesis.

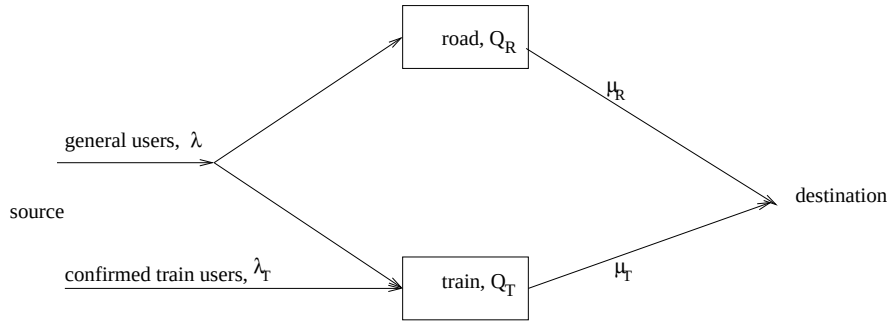


Figure 1.1: The network model: Q_R is a single server queue and Q_T is a batch service queue.

The network we consider here consists of two routes as described above (see Figure 1.1) with the FCFS single server queue (Q_R) being modelled as a $\cdot|M|1$ queue under state-dependent routing and a $\cdot|G|1$ instead of $\cdot|M|1$ queue under probabilistic routing. The batch service queue (Q_T) is modelled as a $\cdot|G^{(N)}|\infty$ queue with batch size $N \geq 2$ and service rate μ_T for both routing cases in Chapters 2 and 3. The two queues behave very differently as their load increases. The expected delay for Q_R increases while that for Q_T decreases as the load is increased. This would not be the case (for large enough arrival rates) if the number of servers were finite. For probabilistic routing we consider $\cdot|G|1$ and $\cdot|G^{(N)}|\infty$ queues since, $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues were completely

analysed in [2]. The network has two independent Poisson arrival processes into it. The first, at rate λ_T, consists of committed public transportation users who always join Q_T – we will also refer to them as train users. The second, at rate λ, consists of general users (commuters) who can join either Q_R or Q_T. Arrivals to Q_R will be served according to the order of their arrival while arrivals to Q_T have to wait for completion of the batch before their service commences. We assume that general users choose the queue that will minimize their own expected transit time and whenever they join a queue they cannot switch queues. We also consider the case where Q_T has a single server instead of infinitely many servers, with Q_R as a $\cdot|M|1$ queue in Chapters 4 and 5.

Calvert [15] studied the existence of a user optimal policy, system optimal policy and Downs-Thomson paradox under state-dependent routing as well as probabilistic routing. His analysis was based on specific parameter values, that is, $\lambda = \lambda_T = \mu_R = \mu_T = 1$ and $N = 2$. Here we extend his analysis to any values of $\lambda > 0, \lambda_T > 0, \mu_R > 0, \mu_T > 0$ and $N \geq 2$. Also unlike his model, we assume that an arrival in states where the expected delays via both routes are the same will join Q_T. This assumption is sufficient to ensure uniqueness properties of the user optimal policy for all parameters in Chapter 2. This system model was also completely analyzed in the setting of probabilistic routing by Afimeimounga, Solomon and Ziedins [2], while here we consider generally instead of exponentially distributed service times for this routing case in Chapter 3. The material in this chapter has appeared in Afimeimounga, Solomon and Ziedins [3].

It is known that the user optimal policy need not coincide with the socially optimal policy. The socially or system optimal policy determines the choice of

route that will minimize the overall average time spent in the system for all users. Naor [47], Correa, Schulz and Stier-Moses [18], Whitt [69] and Ziedins [74] compared user optimal and socially optimal policies for the system they studied in the context of state-dependent routing, while Hassin and Haviv [27] compared them in the context of state-dependent routing as well as probabilistic routing. Although Calvert's work concentrated on state-dependent routing, he also provided a single example of the existence of the user equilibrium and socially optimal policy for the probabilistic routing case. Cohen and Kelly [17] and Bell and Stidham [9] studied both user optimal and socially optimal policies in the context of probabilistic routing only. Afimeimounga, Solomon and Ziedins [2] extended the model considered by Calvert [15] in the context of probabilistic routing. We will analyze the relationship of the user equilibria and social optimal policies under probabilistic routing in Chapter 3. Kumar and Walrand [34] studied user optimal and system optimal policies for a buffer feeding several alternative routes where customers can choose either one of the routes currently available or wait until one with lower delay is free.

Under state-dependent routing in our model, a user's choice of queue will depend not only on the current state of the system but also on the choices made by later arrivals, since service in the batch service queue does not commence until the batch is completed and fewer arrivals into the batch service queue leads to a longer waiting time for service. Calvert, Solomon and Ziedins [16] considered a model with the same property when they studied a network of queues with a single mode of transport illustrating Braess's paradox, while the system we study here consists of two modes ([7, 33]) of transport. Other authors who have considered systems of parallel queues where future arrivals may affect the delays of the customers currently in the system include Altman

and Shimkin [4], Ben-Shahar, Orda and Shimkin [10], Parlaktürk and Kumar [50], Brooms [14], Hassin [24] and Hassin and Haviv [25, 26, 27].

Altman and Shimkin [4] considered a system where users have to choose between a processor-sharing queue and an infinite server queue and they showed the existence and uniqueness of the user equilibrium using stochastic coupling arguments as we do here. Ben-Shahar et al. [10] extended the result in [4] from single class (homogeneous) customers to multiple class (heterogeneous) customers. Brooms [14] considered a similar model to Altman and Shimkin [4] and used coupling arguments to establish stochastic order results for the sojourn time in the system he studied with respect to the entry state. Hassin [24] considered a system with a FCFS single-server service queue and a single-server with random order. An arriving customer joins the random order service queue if and only if its length is less than a given constant called a threshold, given the state of the FCFS service queue. He showed the existence of the user optimal policy for this system and conjectured the uniqueness of it as a result of numerical examples. Hassin and Haviv [26, 27] studied a system where a single class of commuters have to choose between a bus service with infinite capacity and an infinite server batch service queue. The expected waiting time for the bus service is five minutes and since each batch has no limit then it is sufficient to represent the state of the system as one dimensional while the state of the system we study here is two-dimensional as will be seen in Chapter 2. Hassin and Haviv [25] considered a queueing model in which two priority levels, high and low, can be purchased and they showed that a multiplicity of equilibria is a common phenomenon in a class of model including the one they studied. Parlaktürk and Kumar [50] considered a system with two queues where arriving customers are allowed to choose between two sequences of oper-

ations, and they gave stability results for this system under various scheduling rules.

Calvert [15] and Ho [29] considered a $\cdot|M|1$ and a $\cdot|G^{(N)}|\infty$ queue as we do here. Calvert [15] showed that user optimal and socially optimal policies exist for this system with some specific parameter values. For this routing case, Ho [29] gave some monotonicity properties that a user optimal policy must possess, if it exists. He also obtained the user optimal policies for numerical examples in his dissertation. In addition he conjectured monotonicity properties of the user optimal policy, which we prove in Chapter 2, and gave numerical examples in support of those conjectures. Calvert [15] and Ho [29] proved their results by solving recursive equations for the state-dependent delay. Here we use stochastic comparison and coupling arguments to prove the existence, uniqueness and monotonicity of the user optimal policy for this routing case in Chapter 2. Furthermore, we use the same arguments to prove these three properties of the user optimal policy for the same system with Q_T modelled as a $\cdot|M^{(N)}|1$ instead of a $\cdot|G^{(N)}|\infty$ queue in Chapter 4. That is, the batch service queue has a single server instead of infinitely many servers.

Generally, monotonicity is one of the structural properties of a value function that is often studied and used to characterize optimal policies (see for example, Hajek [21], Stidham and Weber [62], Koole [32], Sparaggis and Cassandras [60] and Zhuang and Li [73]). Sample path, coupling and stochastic comparison methods are widely used in Markov decision models to establish monotonicity properties for such policies (see for example Ross [52], Lindvall [40], Thorisson [65], Shaked and Shanthikumar [55], Last and Szekli [35] and Hakej [21]). See Lindvall [40] and Thorisson [65] for introductions to these

methods.

Unlike the state-dependent policy we discussed above, where the delay experienced by an arriving commuter depends not only on the current state of the system but also on the choices made by later arrivals, Spicer and Ziedins [61] considered a state-dependent policy where the delay experienced by a new arrival to the system depends only on the current state of the system at the time of arrival when choosing the route to use (see also Naor [47], Bell and Stidham [9], Cohen and Kelly [17], Whitt [69], Winston [71], Veeranraghav and Debo [66], Ziedins [74] and Sheu and Ziedins [57]).

Under probabilistic routing arriving commuters do not observe the current queue length of each queue but they know the arrival rates into the system and the service rates in each queue when choosing the queue to join or route to use. We assume for this case that each user chooses their route probabilistically according to the well-known Wardrop principle [68], so that the delays on all routes actually in use between a source-destination pair are equal or less than those on any unused routes. Probabilistic routing also can be interpreted as the model where individuals only have knowledge of the mean length of the two queues, rather than the instantaneous state of the system (see for example, Bell and Stidham [9]). Bell and Stidham [9] considered a system with K servers where arriving customers cannot observe the current queue length but they know the service-time distributions and waiting costs at the various servers. Each server was modelled as an $M|G|1$ queue as we do here for our single-server queue. They demonstrated that there are systematic differences between individual and social optimization. Afimeimounga, Solomon and Ziedins [2] showed that multiple user equilibria exist for some parameter values but not

all of them are stable. We will explore in Chapter 3 these properties of the user equilibria for the same network but with a general service time distribution at the single-server queue, Q_R. For this routing case we investigate the existence, uniqueness and stability properties of the user-equilibrium and its relationship to evolutionary stable strategies (ESS), avoid the crowd (ATC) and follow the crowd (FTC) strategies (see for example, Hassin and Haviv [25, 27] and Haviv and Kerner [28]). Furthermore, we investigate the existence, uniqueness and stability properties of the user equilibrium for the same system with Q_T as an $M|M^{(N)}|1$ instead of $M|M^{(N)}|\infty$ queue.

For the system we study here an equilibrium is said to be locally stable if small deviations from it do not cause the system to diverge from the original equilibrium state. A formal definition of the stability property for this system is in [2] p. 325. In the concept of ESS defined by Maynard-Smith [41] stability is a global property, and Hassin and Haviv [27] discuss ESS in the context of queueing problems. The stable user equilibria we obtained for our $\cdot|G|1$ and $\cdot|G^{(N)}|\infty$ queues are not always ESS. Various other definitions of stability in transportation networks with a choice of route can also be found in for example, Smith [59, 58], Horowitz [31], Borst, Jonckheere and Leskelä [11].

According to Hassin and Haviv [25, 27] a naturally good policy for an individual whose objective is to reduce the time spent in a system is to act differently from others, that is to avoid the crowd. That is, an individual tendency to select an action decreases with the tendency of the others in the population to do it. For this model with both single-server and infinite-server batch service queues available to choose from, ATC refers to the case where a marked general user's tendency to use the single-server service (action) *de-*

creases with the tendency to use that service by other general users, and FTC refers to the case where a marked general user's tendency to use the single server service *increases* with the tendency to use that service by other general users. We formally define ESS, ATC and FTC properties for this system in Chapter 3.

In transportation networks, the Downs-Thomson paradox refers to a situation where increasing the capacity of a transportation system, which is modelled here by increasing the service rate for private transport (the FCFS single server queue), does not always decrease the long-run expected transit time for the system as users switch from private transport to public transport or vice versa. This phenomenon was first introduced by Downs [19] in 1962 and has been studied in a variety of settings by Thomson [64], Arnott and Small [6], Abraham and Hunt [1], Brady [12], Holden [30], Mogridge [45], Sharp [56], Calvert [15], Afimeimounga, Solomon and Ziedins [2], Lee, Lee and Lee [37], Lee, Ryu and Lee [38] and the references therein. Calvert [15] was the first to formulate the Downs-Thomson paradox in the queueing network we study here. There are other well-known paradoxes in queueing network such as Braess's paradox (see for example, Cohen and Kelly [17], Calvert, Solomon and Ziedins [16], Bean, Kelly and Taylor [8], Lin and Lo [39]), and see for example, Arnott and Small [6], Ziedins [74] and Ye [72] for others.

Calvert [15] observed the Downs-Thomson effect more clearly in the probabilistic routing than in the state-dependent routing case. We observe that the Downs-Thomson effect is slight under state-dependent routing in our numerical examples and this effect is much less than that observed under probabilistic routing. That is, performance may improve if additional information is avail-

able to users. That is an interesting observation, since it suggests that traffic networks where users have information about the instantaneous state of the network may sometimes perform better than networks where only average delay is available. The effects of differing levels of information on users are often of interest in systems where customers may renege from the queue (see for example, Whitt [70], Guo and Zipkin [20] and Armony, Shimkin and Whitt [5]). They all make comparisons between systems where users have information only about the mean delay in equilibrium and systems where more detailed information is available, either the number of users already in the system, or an estimate of the delay given the current level of congestion. In some cases providing more accurate information improves performance, whereas in others it may harm performance (see also Hassin [22, 23] and Hassin and Haviv [25]). We make similar observations here. The worst effect of the Downs-Thomson paradox disappears under state-dependent routing, however the system does not always perform better under state-dependent routing than probabilistic routing, as we will see in Chapter 2.

The main contributions of this thesis are the proofs of the existence, uniqueness and various monotonicity properties of the user optimal policy under state-dependent routing for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues in Chapter 2 as well as the system with $\cdot|M|1$ and $\cdot|M^{(N)}|1$ queues in Chapter 4. These proofs use stochastic comparison and coupling arguments. We also provide numerical examples which illustrate some important features of expected delays under both state-dependent and probabilistic routing for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues in Chapter 2. In Chapter 3 we provide numerical examples of the existence of multiple user optimal policies (user equilibria) under probabilistic routing for the system with $\cdot|G|1$ and $\cdot|G^{(N)}|\infty$ queues and

batch size $N \geq 2$. In contrast, from [2] if the single server queue is modelled as a $\cdot|M|1$ queue instead of a $\cdot|G|1$ queue while the batch service queue is a $\cdot|G^{(N)}|\infty$ queue then there always exists a unique, stable user equilibrium if the batch size $N > 2$. For this routing case we also analysed the relationship of the social optimum values and the user equilibria if multiple user equilibria exist. Finally, for this routing case we also provide numerical examples that multiple user equilibria exist if we consider the network with $\cdot|M|1$ and $\cdot|M^{(N)}|1$ queues.

The structure of this thesis is as follows. First we consider the case when the batch service queue has an infinite number of servers and study the properties of the user optimal policy or user equilibrium under state-dependent routing and probabilistic routing in chapters 2 and 3, respectively. We then consider the case when the batch service has one server and study analogous properties of the user equilibrium for each routing case in chapters 4 and 5. We conclude and discuss results and possible extensions for this model in Chapter 6.

Chapter 2

State-dependent routing for $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues

2.1 Introduction

In this chapter we consider the network defined in Chapter 1 with Q_R as a $\cdot|M|1$ queue with service rate μ_R, and Q_T as a $\cdot|G^{(N)}|\infty$ queue with service rate μ_T. These queues model possible routes from the same source to the same destination. The expected transit time for Q_R increases as the arrival rate into it increases, since arrivals are served in a first come first served fashion. The expected transit time for Q_T, however, decreases as its arrival rate increases, since the faster arrivals occur, the faster the batch fills up, and the more frequently service takes place. There are two streams of independent Poisson arrivals to this network. The first, at rate λ_T, consists of the committed public transportation (train) users who will always join Q_T. The second, at rate λ, is

the general users who can choose which queue to join. We assume that arriving general users know the state of the system, the arrival rates and service rates before they make their choice of the route to use, and they choose the quickest one. Their decision also depends on that of later arriving general commuters. Once a commuter has joined a queue they cannot switch queues. That is, we model a route choice that is the same for all general commuters, as a function of the state seen by the arriving general commuter on arrival. The service times at Q_R are exponentially distributed with mean $1/\mu_R$. For Q_T, since it has an infinite number of servers, as soon as a batch is full the service on that batch starts immediately. Furthermore since we are only interested in comparing the expected delay of Q_T to that of Q_R, we can consider generally distributed service times with mean $1/\mu_T$ for Q_T. All inter-arrival times and service times are independent of one another.

In this chapter we show that for any parameter values $\lambda > 0$, $\mu_R > 0$, $\lambda_T > 0$, $\mu_T > 0$ and $N \geq 2$ there exists a unique user optimal policy and that it possesses various monotonicity properties.

Calvert [15] considered this system model with fixed parameter values $N = 2, \lambda = \lambda_T = \mu_T = 1$ and $\mu_R \in (0, \sqrt{5} - 1)$. He showed that a user optimal policy exists and is unique for $\mu_R \in (0, 2/3) \cup (1, \sqrt{5} - 1)$ and that multiple user optimal policies exist if $\mu_R = 2/3$ and 1. Note that in Calvert's definition of the user optimal he did not specify which queue to join if the expected transit times in both queues are the same while here we assume that they will join Q_T. Under our joining assumptions we show that the user optimal policy for this routing case is always unique.

The structure of this chapter is as follows. Section 2.2 gives relevant def-

initions and some preliminary results. In section 2.3, we provide the proof of the monotonicity property of the expected transit times via Q_T conditional on the state of the system seen by an arriving general user and show that a user optimal policy exists, is unique and possesses certain monotonicity properties. In section 2.4 we show that the expected transit times for general users going to the destination via Q_T conditional on the state of the system and user optimal policy change monotonically with respect to the arrival and service rate parameters. We conclude and discuss the results we obtain in this chapter in section 2.5.

2.2 Definitions and preliminaries

Consider the network of two queues in parallel described above. We are interested in comparing the delay between the two queues. Since the number of servers at Q_T is infinite, the time spent in this queue waiting for service by an arriving commuter is just the time until their own batch is completed and their service then starts immediately. The service time for Q_T does not affect the waiting time or service time of later arriving general users. The expected service time for this queue is known and equal to $1/\mu_T$. Therefore, we can consider a reduced description of the state of this system as follows. Let $X_1(t) \in \mathbb{Z}^+$ denote the total number of commuters in Q_R including any commuters in service. Let $X_2(t) \in \{0, 1, 2, \ldots, N-1\}$ denote the number of commuters waiting for service in Q_T. Since the number of servers in Q_T is infinite, service on a batch commences immediately as soon as it is completed and then X_2 moves from state $N-1$ to 0. Note that this state description omits the number of commuters in service at Q_T since the number of servers in it is

infinite. However, this reduction and simplification of the state space still gives sufficient information for the calculation of the expected delay at that queue. Therefore we can consider a general service distribution for Q_T with known expected service time $1/\mu_T$. The state space of the process $X(t) = (X_1(t),\, X_2(t))$ of this system is given by $\Omega = \{(a,\, n) :\ a \in \mathbb{Z}^+,\, n \in \{0, 1, \ldots, N-1\}\}$. Let $\boldsymbol{x} = (a,\, n),\ e_1 = (1, 0)$ and $e_2 = (0, 1)$. We extend Calvert's [15] (and Ho's [29]) state-dependent decision policy for general users formally in the following definition.

Definition 2.2.1 *A policy $D = (R,\, T)$ is a partition of Ω into two sets R and T. An arriving general commuter who sees $\boldsymbol{x} \in R$ will go via Q_R. An arriving general commuter who sees $\boldsymbol{x} \in T$ will go via Q_T. Let $\mathcal{D}$ be the class of all decision policies. We will write X_D to denote the process X operating under the decision policy $D \in \mathcal{D}$.*

The process $\{X_D(t),\, t \geq 0\}$ is a Markov process with transition rates as follows:-

$$\boldsymbol{x} \longrightarrow \begin{cases} \boldsymbol{x} - e_1 & \text{at rate } \mu_R \text{ if } a \geq 1 \\ \boldsymbol{x} + e_1 & \text{at rate } \lambda I_{\{\boldsymbol{x} \in R\}} \\ \big(a,\, (n+1) \mathrm{mod} N\big) & \text{at rate } \lambda_T + \lambda I_{\{\boldsymbol{x} \in T\}} \end{cases} \tag{2.1}$$

where $I_{\{\boldsymbol{x} \in R\}}$ and $I_{\{\boldsymbol{x} \in T\}}$ are indicator functions as follows

$$I_{\{\boldsymbol{x} \in R\}} = \begin{cases} 1 & \text{if } \boldsymbol{x} \in R \\ 0 & \text{if } \boldsymbol{x} \notin R \end{cases} \qquad I_{\{\boldsymbol{x} \in T\}} = \begin{cases} 1 & \text{if } \boldsymbol{x} \in T \\ 0 & \text{if } \boldsymbol{x} \notin T \end{cases}$$

Before proceeding, we need to define a number of random variables associated with the system operating under a decision policy $D \in \mathcal{D}$.

We begin with the standard definition of the reaching or hitting time to a set of states. For any $\boldsymbol{x} \in \Omega$, $H \subset \Omega$, let $\tau_D(\boldsymbol{x}, H; s) = \inf_{t \geq 0} \{t : X_D(t+s) \in H | X_D(s) = \boldsymbol{x}\}$ be the first reaching time to the set of states H after time s, given that X_D is in state $\boldsymbol{x}$ at time s. Let $m_D(\boldsymbol{x}, H; s) = E(\tau_D(\boldsymbol{x}, H; s))$ be the expected reaching time to the set of states H starting from state $\boldsymbol{x}$.

Since for Q_T the time spent by an arriving general commuter in the system is the time that they are waiting for the completion of their batch of N commuters, we also need the following related definitions. Let $H_T = \{\boldsymbol{x} \in \Omega : n = 0\}$. Also let $\omega_D(\boldsymbol{x}, s)$ be the time spent waiting for service by a general commuter who arrives at time s to find the system in state $\boldsymbol{x}$ and joins Q_T. Since service for this general commuter commences as soon as the batch they have joined is completed then

$$\begin{aligned} \omega_D(\boldsymbol{x}, s) &= \inf_{t \geq 0} \{t : X(t+s) \in H_T | X(s-) = \boldsymbol{x},\, X(s) = (a,\, (n+1) \bmod N)\} \\ &\sim \tau_D((a,\, (n+1) \bmod N),\, H_T; s) \end{aligned}$$

by the Markov property, where $X(s-) = \lim_{t \nearrow s} X(t)$. By time homogeneity, the distributions of $\omega_D(\boldsymbol{x}, s)$ and $\tau_D(\boldsymbol{x},\, H_T; s)$ do not depend on s, therefore we will omit the dependence on s and write $\omega_D(\boldsymbol{x})$ and $\tau_D(\boldsymbol{x},\, H_T)$. Since we are only interested in the first reaching time to the set H_T we will write $\tau_D(\boldsymbol{x})$ and $m_D(\boldsymbol{x})$ for $\tau_D(\boldsymbol{x},\, H_T)$ and $m_D(\boldsymbol{x},\, H_T)$, respectively. Note that in the special case where $n = N - 1$, we have $\omega_D(a,\, N-1) = \tau_D(a,\, N \bmod N) = 0$ with probability 1 as required.

We can now make the following definition:-

Definition 2.2.2 *For any policy $D = (R, T) \in \mathcal{D}$ and $\boldsymbol{x} \in \Omega$, let $y(\boldsymbol{x})$ and $z_D(\boldsymbol{x}) = E\big(\omega_D(\boldsymbol{x})\big) + \frac{1}{\mu_T}$ be the expected time to reach the destination via Q_R and Q_T respectively, for an arriving general commuter who finds the system in state $\boldsymbol{x}$ on arrival.*

Since $\omega_D(\boldsymbol{x}) = \tau_D\big(a,\, (n+1)\mathrm{mod}\, N\big)$ we have $z_D(\boldsymbol{x}) = m_D\big(a,\, (n+1)\mathrm{mod}\, N\big) + \frac{1}{\mu_T}$. Thus when $n < N-1$, $z_D(\boldsymbol{x})$ can also be interpreted as the expected time until the current batch completes service given the process X_D is in state $\boldsymbol{x} + e_2$. Note that, in general, the value of $z_D(\boldsymbol{x})$ depends on the policy D.

If arriving general commuters see the system is in state $\boldsymbol{x}$ and join Q_R, they reach their destination after $a+1$ services, regardless of the policy D (see Calvert [15]). Therefore

$$y(\boldsymbol{x}) = \frac{a+1}{\mu_R} \quad \text{for all } a \geq 0. \tag{2.2}$$

The expected delay for those general users who join Q_T is more complicated. If an arriving general commuter sees state $(a,\, N-1)$ on arrival and joins Q_T then service of the completed batch starts immediately. Therefore,

$$z_D(a, N-1) = \frac{1}{\mu_T} \qquad \text{for } a \geq 0. \tag{2.3}$$

If $n \leq N-2$ then service does not commence immediately. The arrival of a general user who joins Q_T moves the system from state $\boldsymbol{x}$ into state $\boldsymbol{x} + e_2$, so that, conditioning on the first transition after the general commuter has joined Q_T we obtain the following equations (see Calvert [15] and Ho [29]).

Let $\eta = \lambda + \lambda_T$ and $\Lambda = \eta + \mu_R$. Now, if $\boldsymbol{x} + e_2 \in T$ then

$$z_D(\boldsymbol{x}) = \begin{cases} 1/\eta + z_D(\boldsymbol{x} + e_2), & \text{if } a = 0 \\ \Big(1 + \eta\, z_D\,(\boldsymbol{x} + e_2) + \mu_R\, z_D(\boldsymbol{x} - e_1)\Big)/\Lambda, & \text{if } a \geq 1 \end{cases} \tag{2.4}$$

Next, if $\boldsymbol{x} + e_2 \in R$ for $0 \leq n \leq N - 2$ then

$$z_D(\boldsymbol{x}) = \begin{cases} \big(1 + \lambda_T\, z_D(\boldsymbol{x} + e_2) + \lambda\, z_D(\boldsymbol{x} + e_1)\big)/\eta & \text{if } a = 0 \\ \big(1 + \lambda_T\, z_D\,(\boldsymbol{x} + e_2) + \lambda\, z_D(\boldsymbol{x} + e_1) + \mu_R\, z_D(\boldsymbol{x} - e_1)\big)/\Lambda, & \text{if } a \geq 1 \end{cases} \tag{2.5}$$

We formally extend the definition of a user optimal policy given in Calvert [15] and Ho [29] by introducing the concept of $\xi-$level optimality. The concept of $\xi-$level optimality refers to the number of space or spaces left to fill in a batch. A $\xi-$level optimal policy, if such exists, is user optimal for $n \geq N - \xi$. We note also that Calvert did not specify a decision in state $\boldsymbol{x}$ if $y(\boldsymbol{x}) = z_D(\boldsymbol{x})$, while in the definition here, if $y(\boldsymbol{x}) = z_D(\boldsymbol{x})$ then $\boldsymbol{x} \in T$ as Ho specified. That is, an arriving general commuter joins Q_T if the expected transit times for both queues are the same.

Definition 2.2.3 *1. We say a policy $D \in \mathcal{D}$ is a ξ-level user optimal policy for some ξ such that $1 \leq \xi \leq N$ if*

$$\boldsymbol{x} \in R \Leftrightarrow y(\boldsymbol{x}) < z_D(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega \textit{ such that } n \geq N - \xi. \tag{2.6}$$

Let $\mathcal{D}^{\star}_{\xi}$ be the class of all ξ-level user optimal policies.

2. We say a policy $D \in \mathcal{D}$ is a user optimal policy if it is an $N-$level user

optimal policy, that is

$$\boldsymbol{x} \in R \Leftrightarrow y(\boldsymbol{x}) < z_D(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega. \tag{2.7}$$

Let $\mathcal{D}^\star = \mathcal{D}_N^\star$ be the class of all user optimal policies.

A policy which is user optimal under Definition 2.2.3 is also user optimal under the definition given in Calvert [15], but the converse does not always hold.

Since we are interested in the existence, uniqueness and monotonicity properties of the user optimal policy $D^\star$, we also provide the following definitions. First, we extend the definition of the ($N-$level) monotonicity property of a decision policy D found in Ho [29] as follows.

Definition 2.2.4 *Let $1 \leq \xi \leq N$. A policy $D \in \mathcal{D}$ is $\xi-$level monotone if D satisfies the following two conditions:*

A) $\boldsymbol{x} \in T \Rightarrow \boldsymbol{x} + e_1 \in T$ *for all* $\boldsymbol{x} \in \Omega$ *such that* $N - \xi \leq n \leq N-1,\ a \geq 0.$

B) $\boldsymbol{x} \in T \Rightarrow \boldsymbol{x} + e_2 \in T$ *for all* $\boldsymbol{x} \in \Omega$ *such that* $N - \xi \leq n < N-1,\ a \geq 0.$

In the special case when D is $N-$level monotone we say the policy is monotone.

We will show the existence of a unique policy, and that it is monotone in the next section. Ho [29] showed that if a user optimal policy exists then the number of states in R is finite. He also provided some numerical examples of user optimal policies $D^\star$ that are monotone.

For $D \in \mathcal{D}$, we formally define the switching points $a^\star(n)$.

Definition 2.2.5 *For $D \in \mathcal{D},\ 0 \leq n \leq N-1$, let $a^{\star}(n) = \inf\{a : (a,n) \in T\}$. If $\{a : (a,n) \in T\} = \phi$ then $a^{\star}(n) = \infty$.*

Definition 2.2.5 extends that given in Ho [29] to general policies, that need not be monotone. We call $a^{\star}(n)$ the $T-$marker or switching point for a fixed n. The vector $a^{\star} = \{a^{\star}(n),\ 0 \leq n \leq N-1\}$ is the switching curve for the policy D. Switching curves commonly appear in Markov decision policies, where the aim is to find the system optimal policy (see, for instance, Stidham and Weber [62], Hajek [21], Koole [32] and Zhuang and Li [73]). Note that the condition B in Definition 2.2.4 of a $\xi-$level monotone policy is equivalent to the condition that $a^{\star}(n+1) \leq a^{\star}(n)$ for $n \geq N - \xi$ which was observed by Ho [29] in a more restricted definition. Figure 2.1 gives an example of a monotone policy, $D = (R,\ T)$.

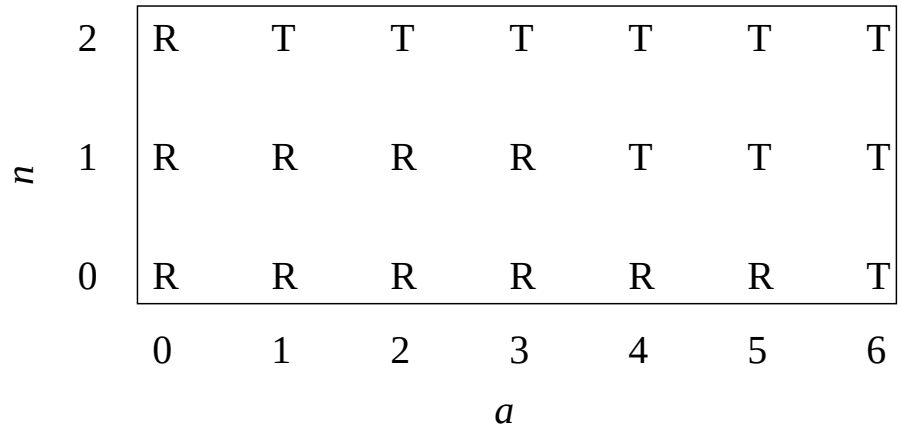


Figure 2.1: A monotone policy when $N = 3$. $R(T)$ denote the states $\boldsymbol{x} \in R$ ($\boldsymbol{x} \in T$) for all $\boldsymbol{x} \in \Omega$. The switching curve in this example is $a^{\star} = (6, 4, 1)$. This is the unique user optimal policy when $\lambda = \lambda_T = 1, \mu_R = 4, \mu_T = 2$.

For any user optimal policy $D^{\star} \in \mathcal{D}^{\star}$, R is finite. To see this, observe that for any state $\boldsymbol{x} \in \Omega$ and any $D \in \mathcal{D}$, if an arriving general user joins Q_T then their expected waiting time until service commences is bounded above by the

expected time it would take to complete the batch if no further general arrivals join Q_T, $(N-1-n)/\lambda_T$. Therefore,

$$z_D(\boldsymbol{x}) \leq \frac{N-1-n}{\lambda_T} + \frac{1}{\mu_T}$$

Also note that $(N-1-n)/\lambda_T + 1/\mu_T \leq (N-1)/\lambda_T + 1/\mu_T$ for all $\boldsymbol{x} \in \Omega$, and $y(\boldsymbol{x})$ is an increasing function of a and does not depend on n. Therefore $y(\boldsymbol{x}) < z_D(\boldsymbol{x})$ for a finite number of states. As a consequence there exists $\bar{a}$ such that $z_D(\boldsymbol{x}) \leq y(\boldsymbol{x})\,(\Rightarrow \boldsymbol{x} \in T)$ for all $a \geq \bar{a}$. Thus a policy D cannot be user optimal if $\boldsymbol{x} \in R$ for any $\boldsymbol{x} \in \Omega$ such that $a \geq \bar{a}$ (see Ho [29]). Furthermore, suppose for a policy D that there exists some $\hat{a} \in \mathbb{Z}^+$ such that $\boldsymbol{x} \in T$ for all $\boldsymbol{x}$ such that $a \geq \hat{a}$. Then all states $\boldsymbol{x} \in \Omega$ such that $a > \hat{a}$ are transient. Finally, observe that starting from any state $\boldsymbol{x} \in \Omega$ the process $X_{D^\star}$ associated with a user optimal policy $D^\star$ returns to the state $(0,0)$ with probability 1 regardless of the arrival rates λ and λ_T. Thus $X_{D^\star}$ has a single closed communication class, which is finite.

Let $\pi_D = \{\pi_D(\boldsymbol{x}),\, \boldsymbol{x} \in \Omega\}$ denote the stationary distribution of a process X_D if it exists. Note that for those policies D such that R is finite, the stationary distribution always exists since the number of recurrent states is finite as we have discussed above. Then the expected time spent in the system including the expected service time (or expected origin-destination (O-D) time) for a general user is given by

$$W = \sum_{\boldsymbol{x} \in \Omega_D} \pi_D(\boldsymbol{x}) \Big(y(\boldsymbol{x}) I_{\{\boldsymbol{x} \in R\}} + z_D(\boldsymbol{x}) \left(1 - I_{\{\boldsymbol{x} \in R\}}\right) \Big). \tag{2.8}$$

2.3 Existence, uniqueness and monotonicity of the user optimal policy

In this section we will use sample path and coupling arguments ([40, 65, 52, 55, 21, 35]) to show existence, uniqueness and monotonicity properties of user optimal policies for this system.

We wish to compare $y(\boldsymbol{x})$ and $z_D(\boldsymbol{x}) = \mathrm{E}\left(\omega_D(\boldsymbol{x})\right) + 1/\mu_T$ to establish the properties of a user optimal policy for this system. We will use sample path and coupling arguments to show properties of $\omega_D(\boldsymbol{x})$, the time spent in the system by a marked general user waiting for the completion of their own batch (not including the service time) before we show the properties of $z_D(\boldsymbol{x})$ in Theorem 2.3.1 and Corollary 2.3.1 respectively. For two random variables, X and Y, we write $X \overset{\text{st}}{\leq} Y$ if $P(X > x) \leq P(Y > x)$ for all $x \in \mathbb{R}$ (see Shaked and Shanthikumar [55]).

Theorem 2.3.1 *Suppose X_D is operating under a policy D that is $\xi-$level monotone for some $\xi : 1 \leq \xi \leq N$. Then $\omega_D(\boldsymbol{x}_1) \overset{st}{\leq} \omega_D(\boldsymbol{x}_2)$ for any pair of states $\boldsymbol{x}_1 = (a_1, n_1) \in \Omega$, $\boldsymbol{x}_2 = (a_2, n_2) \in \Omega$ such that $n_2 \geq N-\xi-1$ satisfying any of the following conditions:-*

I. $a_1 = a_2$, $n_1 > n_2$,

II. $a_1 < a_2$, $n_1 > n_2$, $a_2 - a_1 \leq n_1 - n_2$,

III. $a_1 > a_2$, $n_1 = n_2$,

IV. $a_1 > a_2$, $n_1 > n_2$.

Figure 2.2 gives examples of conditions I to IV respectively.

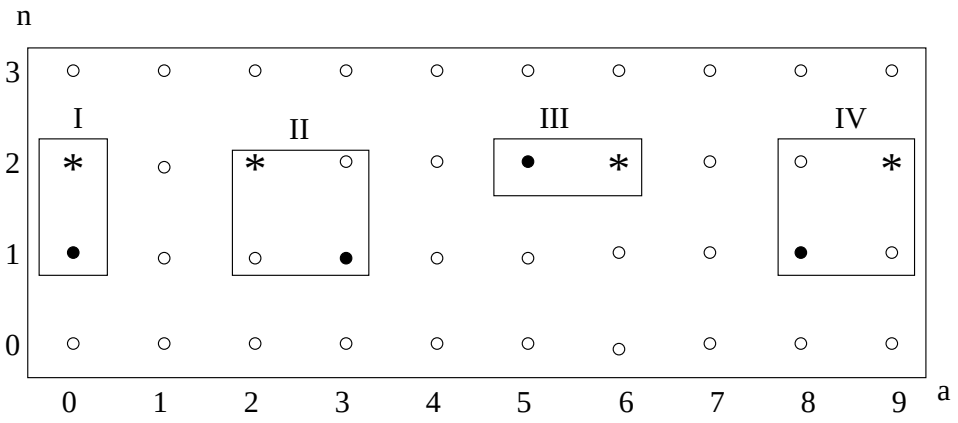


Figure 2.2: Examples of each condition *I* - *IV*. $*$ denotes $\boldsymbol{x}_1 = (a_1, n_1)$ and $\bullet$ denotes $\boldsymbol{x}_2 = (a_2, n_2)$.

Note that the above individual conditions satisfy $n_1 \geq n_2$. We will prove Theorem 2.3.1 by considering a joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$, with both $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ obeying the law of X_D and transition rates such that if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(0) = \big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfies one of the conditions *I*, *II*, *III* or *IV* then $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ will continue to do so until service commences at Q_T in the $\boldsymbol{X}^{(1)}$ process. Let $H_T^{(i)} = \{(\boldsymbol{x}_1, \boldsymbol{x}_2) : n_i = 0\}$ and $w_D^{(i)}(\boldsymbol{x}_i) = \inf_{t \geq 0}\{t : X_i(t+s) \in H_T^{(i)} | X_i(s-) = \boldsymbol{x}_i,\, X_i(s) = (a_i, (n_i + 1) \bmod N)\}$. We will show that with probability 1, $\omega_D^{(1)}(\boldsymbol{x}_1) \leq \omega_D^{(2)}(\boldsymbol{x}_2)$ and hence $\omega_D(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \omega_D(\boldsymbol{x}_2)$.

Proof. We consider first the case when $n_1 = N - 1$. Then trivially $\omega_D^{(1)}(\boldsymbol{x}_1) = 0 \leq \omega_D^{(2)}(\boldsymbol{x}_2)$ with probability 1 for all conditions *I* - *IV*, since if the arriving general commuter joins Q_T then the batch is completed.

We have left to prove the case where $n_1 < N - 1$ for $\omega_D^{(1)}(\boldsymbol{x}_1)$, that is, $n_1 \leq N - 1$ for $\tau_D^{(1)}(\boldsymbol{x}_1; H_T^{(1)})$. Since we are interested in the reaching time to $H_T^{(i)}$ only we will write $\tau_D^{(i)}(\boldsymbol{x}_1)$ instead of $\tau_D^{(i)}(\boldsymbol{x}_1; H_T^{(i)})$. Here, similarly to $\omega_D^{(i)}$, $\tau_D^{(i)}(\boldsymbol{x}_i;\, H_T^{(i)})$ and $\tau_D^{(i)}$ refer to X_i in the joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$.

Now, define the joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ with state space $\Omega \times \Omega$ and

transitions for this process as follows. Let $\left(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\right)$ be the state immediately after a transition out of $\left(\boldsymbol{x}_1,\, \boldsymbol{x}_2\right)$.

Services in Q_R occur at rate μ_R and cause transitions (2.9) - (2.11) to occur.

$$\begin{aligned} & \nearrow && \left(\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2 - e_1\right), && \text{if } a_1,\, a_2 \geq 1 && (2.9) \\ (\boldsymbol{x}_1,\, \boldsymbol{x}_2) & \longrightarrow && \left(\boldsymbol{x}_1,\, \boldsymbol{x}_2 - e_1\right), && \text{if } a_1 = 0, a_2 \geq 1 && (2.10) \\ & \searrow && (\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2), && \text{if } a_1 \geq 1,\ a_2 = 0 && (2.11) \end{aligned}$$

Transition (2.9) occurs if both $a_1 \geq 1$ and $a_2 \geq 1$ and so $(\boldsymbol{x}_1', \boldsymbol{x}_2') = \left(\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2 - e_1\right)$ continues to satisfy the condition that $(\boldsymbol{x}_1,\, \boldsymbol{x}_2)$ satisfies. Under the condition *II* transition (2.10) can occur and $\left(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\right) = \left(\boldsymbol{x}_1,\, \boldsymbol{x}_2 - e_1\right)$ will satisfy either the same condition, if $a_2 > 1$, or condition *I*, if $a_2 = 1$. Under the conditions *III* and *IV* transition (2.11) can occur. If $a_1 > 1$ and $a_2 = 0$ then $\left(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2)$ will preserve whichever of the conditions *III* or *IV* that $(x_1,\, x_2)$ satisfies. For $a_1 = 1$, if $\left(\boldsymbol{x}_1,\, \boldsymbol{x}_2\right)$ satisfies *III* then the processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ will couple and remain coupled thereafter, and if $\left(\boldsymbol{x}_1,\, \boldsymbol{x}_2\right)$ satisfies *IV* then $\left(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\right)$ will satisfy *I*. Under condition *I* no service can take place if $a_1 = a_2 = 0$.

Transitions (2.12) - (2.15) are possible if an arrival into the system occurs.

$$\nearrow \quad \big((a_1, (n_1+1)\mathrm{mod}\, N),\, (a_2, (n_2+1)\mathrm{mod}\, N)\big) \qquad (2.12)$$

$$\text{at rate } \lambda_T + \lambda I_{\{\boldsymbol{x}_1 \in T,\, \boldsymbol{x}_2 \in T\}}$$

$$\nearrow \quad \big(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2 + e_1\big) \qquad (2.13)$$

$$(\boldsymbol{x}_1,\, \boldsymbol{x}_2) \qquad \text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in R,\, \boldsymbol{x}_2 \in R\}}$$

$$\searrow \quad \big((a_1, (n_1+1)\mathrm{mod}\, N),\, \boldsymbol{x}_2 + e_1\big), \qquad (2.14)$$

$$\text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in T,\, \boldsymbol{x}_2 \in R\}}$$

$$\searrow \quad \big(\boldsymbol{x}_1 + e_1,\, (a_2, (n_2+1)\mathrm{mod}\, N)\big) \qquad (2.15)$$

$$\text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in R,\, \boldsymbol{x}_2 \in T\}}$$

Arrivals from the committed train users' stream cause transition (2.12) to occur at rate λ_T. If $n_1 < N-1$ then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 + e_2,\, \boldsymbol{x}_2 + e_2)$ will satisfy the condition that $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfies. Under the conditions *I, II* and *IV* if $n_1 = N-1$ then the batch is completed in $\boldsymbol{X}^{(1)}$ but not in $\boldsymbol{X}^{(2)}$; under the condition *III* if $n_1 = n_2 = N-1$ then both batches will complete simultaneously; in both cases $\tau_D^{(1)}(\boldsymbol{x}_1) \leq \tau_D^{(2)}(\boldsymbol{x}_2)$, with probability 1 is satisfied.

Now consider the transition when the arrival of a general user occurs. First, if on arrival $\boldsymbol{x}_1 \in T$ and $\boldsymbol{x}_2 \in T$ then a transition such as transition (2.12) occurs, which we have already discussed. The condition of $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = \big((a_1, (n_1+1)\mathrm{mod}\, N),\, (a_2, (n_2+1)\mathrm{mod}\, N)\big)$ is the same as that when an arrival from the committed train users occurs. Second, if on arrival $\boldsymbol{x}_1 \in R$ and $\boldsymbol{x}_2 \in R$ then a transition such as (2.13) occurs and $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = \big(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2 + e_1\big)$ will satisfy the same condition that $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfies. Another case is when an arriving general user finds $\boldsymbol{x}_1 \in T$ while $\boldsymbol{x}_2 \in R$ then a transition such as

(2.14) occurs. For $n_1 < N-1$, if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfies condition I or II then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = \big((a_1, (n_1+1) \mathrm{mod}\, N),\, \boldsymbol{x}_2 + e_1\big)$ will satisfy II. If $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfies condition III or IV then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ will satisfy either condition I (if $a_1 = a_2+1$) or IV (if $a_1 > a_2+1$). If $n_1 = N-1$ then the batch is completed in $\boldsymbol{X}^{(1)}$ and service commences for that batch but not in $\boldsymbol{X}^{(2)}$. Finally, consider when arriving general users find $\boldsymbol{x}_1 \in R$ while $\boldsymbol{x}_2 \in T$ then a transition such as (2.15) occurs. Since D is $\xi-$level monotone and $n_2 \geq N-\xi$ we need to consider this case for condition II only, since this case cannot hold for any of the conditions I , III or IV . Then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = \big(\boldsymbol{x}_1 + e_1,\, (a_2, (n_2+1) \mathrm{mod}\, N)\big)$ will preserve the same condition if $a_2 > a_1 + 1$ and $n_1 > n_2 + 1$; it will satisfy condition I if $a_2 = a_1 + 1$ and $n_1 > n_2 + 1$; or $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ will couple if $a_2 = a_1 + 1$ and $n_1 = n_2 + 1$.

Since all $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ for this coupled system satisfy one of the conditions I - IV, $\tau_D^{(1)}(\boldsymbol{x}_1) \leq \tau_D^{(2)}(\boldsymbol{x}_2)$ with probability 1, and so $\tau_D(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \tau_D(\boldsymbol{x}_2)$. Hence, $\omega_D(\boldsymbol{x}_1 - e_2) \overset{\text{st}}{\leq} \omega_D(\boldsymbol{x}_2 - e_2)$. ■

Corollary 2.3.1 *Suppose X_D is operating under a $\xi-$level monotone policy D, for some ξ such that $1 \leq \xi \leq N$. Then $z_D(\boldsymbol{x}_1) \leq z_D(\boldsymbol{x}_2)$ for any pair of states $\boldsymbol{x}_1,\, \boldsymbol{x}_2 \in \Omega$ satisfying any of the conditions* I - IV *in Theorem 2.3.1 such that $n_2 \geq N - \xi - 1$.*

Proof From Theorem 2.3.1 we have $\omega_D(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \omega_D(\boldsymbol{x}_2)$ which implies $E\big(\omega_D(\boldsymbol{x}_1)\big) \leq E\big(\omega_D(\boldsymbol{x}_2)\big)$. Hence, $z_D(\boldsymbol{x}_1) = E\big(\omega_D(\boldsymbol{x}_1)\big) + \frac{1}{\mu_T} \leq z_D(\boldsymbol{x}) = E\big(\omega_D(\boldsymbol{x}_2)\big) + \frac{1}{\mu_T}$. ■

Theorem 2.3.2 *Consider a process X with parameters $\Gamma = (\lambda,\, \lambda_T,\, \mu_R,\, \mu_T)$ and $N \geq 2$. Then there exists a unique user optimal policy, $D^{\star} \in \mathcal{D}^{\star}$ for this*

system. Furthermore $D^\star$ is monotone, that is, the switching curve for $D^\star$, $a^\star$, satisfies $a^\star(n+1) \leq a^\star(n)$, $0 \leq n \leq N-2$.

Proof We use proof by induction on $\xi = N - n$ such that $\xi \in \{1, 2, \ldots, N\}$. The inductive hypothesis for ξ is that the class of $\xi-$level user optimal polices, $\mathcal{D}^\star_\xi$, is non-empty and any policy $D \in \mathcal{D}^\star_\xi$ is $\xi-$level monotone and so if D_1, $D_2 \in \mathcal{D}^\star_\xi$ are two $\xi-$level user optimal policies, then for all $\boldsymbol{x} \in \Omega$ such that $n \geq N-\xi$, $\boldsymbol{x} \in R_1 \Leftrightarrow \boldsymbol{x} \in R_2$. If this is the case, we say that the decision policies, D_1 and D_2, are identical up to level ξ and that any such policy is unique up to level ξ.

Consider first $\boldsymbol{x} \in \Omega$ such that $n = N-1$, that is, $\xi = 1$. For all $a \geq 0$, $z_{D}(a,\, N-1) = \frac{1}{\mu_T}$ and since $y(a,\, N-1) = \frac{a+1}{\mu_R}$ is strictly increasing in a then $\frac{a+1}{\mu_R} \geq \frac{1}{\mu_T}$ if and only if $a \geq \frac{\mu_R}{\mu_T} - 1$. Therefore, $a^\star\,(N-1) = \lceil \frac{\mu_R}{\mu_T} - 1 \rceil$ is the unique 1-level user optimal policy switching point. Hence, the inductive hypothesis holds for $\xi = 1$.

Now suppose the inductive hypothesis hold for $\xi = m$. We will show that it also holds for $\xi = m+1$, that is, $n = N-m-1$. Let $D \in \mathcal{D}^\star_m$. By the inductive hypothesis at least one such D exists, and furthermore, D is monotone and unique up to level m. By Corollary 2.3.1 condition *III*, $z_{D}(a,\, N-m-1)$ is decreasing in a, and since $y(a,\, N-m-1)$ is strictly increasing in a, $z_{D}(a,\, N-m-1) \leq y(a,\, N-m-1)$ if and only if $a \geq \beta$, where $\beta = \min\{a : z_{D}(a, N-m-1) \leq y(a,\, N-m-1)\}$. That is, the user optimal decision in state $(a,\, N-m-1)$ is to choose Q_T if and only if $a \geq \beta$. By Corollary 2.3.1 condition *I*, $z_{D}(a,\, N-m) \leq z_{D}(a,\, N-m-1)$, while $y(a,\, N-m) = y(a,\, N-m-1) = \frac{a+1}{\mu_R}$. Therefore, $a^\star(N-m) \leq \beta$. Now suppose we have a $\xi-$level monotone policy $D' = (R',\, T') \in \mathcal{D}^\star_m$ such that for

$\boldsymbol{x} \in \Omega$ with $n \neq N-m-1$, $\boldsymbol{x} \in R'$ if and only if $\boldsymbol{x} \in R$. That is, the decision policy, D', is the same as that for D for these states. For $\boldsymbol{x} \in \Omega$ such that $n = N-m-1$, let $(a,n) \in R'$ if and only if $a < \beta$. Thus $D' \in \mathcal{D}^{\star}_{m+1}$, that is, D' is an $(m+1)-$level user optimal policy with $a^{\star}(N-m-1) = \beta \geq a^{\star}(N-m)$. Since the partition D' is uniquely determined by the partition D for $n \geq N-m$, which is identical up to level m for all $D \in \mathcal{D}^{\star}_{m}$, the uniqueness of $a^{\star}(n-m-1)$ follows. Therefore, all $(m+1)-$level user optimal policies are identical up to level $m+1$. Hence, by the principle of induction, there exists at least one $N-$level user optimal policy $D^{\star}$ and since all $N-$level user optimal decision policies are identical up to level N, the user optimal policy $D^{\star}$ is unique. ■

2.4 Monotonicity

In this section we analyze the effect of each system parameter, λ, λ_T, μ_R, μ_T on the expected transit time for an arriving general commuter seeing the system in state $\boldsymbol{x} \in \Omega$, $z_D(\boldsymbol{x})$, the user optimal policy $D^{\star}$ and the expected time in the system including the service time for a general user, W. Since we will be varying decision policies, as well as arrival and service rates, in this section we let $\omega_D(\boldsymbol{x};\Gamma)$ be the time until the batch fills for a general user who, on arrival, sees the system X_D with parameters $\Gamma = (\lambda, \lambda_T, \mu_R, \mu_T)$ in state $\boldsymbol{x}$. Also for this system X_D with parameter Γ, we let $\tau_D(\boldsymbol{x};\Gamma)$ be the first reaching time to the set $H_T = \{\boldsymbol{x} \in \Omega : n = 0\}$. As in section 2.2 $\omega_D(\boldsymbol{x} - e_2;\Gamma) \equiv \tau_D(\boldsymbol{x};\Gamma)$. Similarly, let $z_D(\boldsymbol{x};\Gamma) = \mathrm{E}\big(\omega_D(\boldsymbol{x};\Gamma)\big) + \frac{1}{\mu_T}$ be the expected time to reach the destination via Q_T for an arriving general user who sees the system X_D in state $\boldsymbol{x}$.

2.4.1 Monotonicity of $\boldsymbol{z} = \{z_D(\boldsymbol{x}), \boldsymbol{x} \in \Omega\}$ in λ, λ_T, μ_R, μ_T and D

Lemma 2.4.1 *Let X_{D_1} and X_{D_2} be two processes with common batch size N and parameters $\Gamma_1 = (\lambda_1, \mu_{R_1}, \lambda_{T_1}, \mu_{T_1})$ and $\Gamma_2 = (\lambda_2, \mu_{R_2}, \lambda_{T_2}, \mu_{T_2})$ respectively. Suppose that these processes are operating under monotone policies $D_1 = (R_1, T_1)$ and $D_2 = (R_2, T_2)$ respectively such that $T_2 \subseteq T_1$. Also suppose that $\lambda_1 \geq \lambda_2, \mu_{R_1} \leq \mu_{R_2}$, $\lambda_{T_1} \geq \lambda_{T_2}$ and $\mu_{T_1} \geq \mu_{T_2}$. Let $z_{D_i}(\boldsymbol{x}_i; \Gamma_i)$, $i = 1, 2$ be the expected time to reach the destination for a general commuter who, on arrival, sees the system X_i in state $\boldsymbol{x}_i$ and joins Q_T. Then*

$$z_{D_1}(\boldsymbol{x}_1; \Gamma_1) \leq z_{D_2}(\boldsymbol{x}_2; \Gamma_2), \tag{2.16}$$

for any pair of states $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega$ satisfying one of the conditions 0 - IV below.

0) $\boldsymbol{x}_1 = \boldsymbol{x}_2$,

I) $a_1 = a_2$, $n_1 > n_2$,

II) $a_1 < a_2$, $n_1 > n_2$, $a_2 - a_1 \leq n_1 - n_2$,

III) $a_2 < a_1$, $n_1 = n_2$,

IV) $a_2 < a_1$, $n_1 > n_2$.

Note that the conditions *I - IV* above are the same as those in Theorem 2.3.1, therefore the following proof will be very similar to that proof.

Proof. We begin by showing that for any monotone policies $D_1 \in \mathcal{D}$ and $D_2 \in \mathcal{D}$ the required inequality, $z_{D_1}(\boldsymbol{x}_1; \Gamma_1) \leq z_{D_2}(\boldsymbol{x}_2; \Gamma_2)$, holds for

$n_1 = N-1$. For $n_1 = N-1$, $\omega_{D_1}(a_1, N-1) = 0$ with probability 1 and since $n_1 = N-1 \geq n_2$ for all conditions *0 - IV* then $z_{D_1}(\boldsymbol{x}_1; \Gamma_1) = \frac{1}{\mu_{T_1}} \leq \frac{1}{\mu_{T_2}} \leq z_{D_2}(\boldsymbol{x}_2; \Gamma_2)$. Note that $z_{D_2}(\boldsymbol{x}_2; \Gamma_2) > \frac{1}{\mu_{T_2}}$ for conditions *I*, *II* and *IV*.

To show that (2.16) holds for $n \leq N-2$, we use coupling arguments similar to those in the proof of Theorem 2.3.1. Again we consider a joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$, with both $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ obeying the law of X_{D_1} and X_{D_2}, respectively. For this joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ a departure from Q_R in $\boldsymbol{X}^{(1)}$ is always accompanied by a departure from Q_R in $\boldsymbol{X}^{(2)}$, an arrival from the committed train user stream to Q_T in $\boldsymbol{X}^{(2)}$ is always accompanied by an arrival from the committed train user stream to Q_T in $\boldsymbol{X}^{(1)}$ and an arrival from the general user stream to Q_R in $\boldsymbol{X}^{(2)}$ is always accompanied by an arrival from the general user stream to Q_R in $\boldsymbol{X}^{(1)}$. The transition rates for this process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ are such that if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(0) = \big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfies one of the conditions *0 - IV* then it will continue to do so until service commences at Q_T in the $\boldsymbol{X}^{(1)}$ process. Therefore, we have $\tau_{D_1}(\boldsymbol{x}_1; \Gamma_1) \overset{\text{st}}{\leq} \tau_{D_2}(\boldsymbol{x}_2; \Gamma_2)$ for $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfying conditions *0 - IV* and such that $1 \leq n_2 \leq n_1 \leq N-1$, which implies $\omega_{D_1}(\boldsymbol{x}_1 - e_2; \Gamma_1) \overset{\text{st}}{\leq} \omega_{D_2}(\boldsymbol{x}_2 - e_2; \Gamma_2)$ and hence, $z_{D_1}(\boldsymbol{x}_1 - e_2; \Gamma_1) \leq z_{D_2}(\boldsymbol{x}_2 - e_2; \Gamma_2)$.

The possible transitions for this system are as follows.

$$
(\boldsymbol{x}_1,\, \boldsymbol{x}_2) \longrightarrow \begin{cases}
(\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2 - e_1 I_{\{a_2>0\}}) \\
\qquad \text{at rate } \mu_{R_1} I_{\{a_1>0\}} \\
(\boldsymbol{x}_1,\, \boldsymbol{x}_2 - e_1) \\
\qquad \text{at rate } \big(\mu_{R_2} - \mu_{R_1} I_{\{a_1>0\}}\big) I_{\{a_2>0\}} \\
\big((a_1, (n_1+1) \mathrm{mod} N),\, (a_2, (n_2+1) \mathrm{mod} N)\big) \\
\qquad \text{at rate } \lambda_{T_2} + \lambda_2 I_{\{\boldsymbol{x}_1 \in T_1, \boldsymbol{x}_2 \in T_2\}} \\
(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2 + e_1) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in R_1, \boldsymbol{x}_2 \in R_2\}} \\
\big((a_1, (n_1+1) \mathrm{mod} N),\, \boldsymbol{x}_2\big) \\
\qquad \text{at rate } (\lambda_{T_1} - \lambda_{T_2}) + (\lambda_1 - \lambda_2) I_{\{\boldsymbol{x}_1 \in T_1\}} \\
(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2) \\
\qquad \text{at rate } (\lambda_1 - \lambda_2) I_{\{\boldsymbol{x}_1 \in R_1\}} \\
\big((a_1, (n_1+1) \mathrm{mod} N),\, \boldsymbol{x}_2 + e_1\big) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in T_1, \boldsymbol{x}_2 \in R_2\}} \\
\big(\boldsymbol{x}_1 + e_1,\, (a_2, (n_2+1) \mathrm{mod} N)\big) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in R_1, \boldsymbol{x}_2 \in T_2\}}
\end{cases} \tag{2.17}
$$

Let $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ be the state immediately after a transition out of state $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$. As above, we will show that if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfies any of the conditions *0* - *IV* then so does $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$.

We begin by considering transitions due to service completion at Q_R. If $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big) \rightarrow \big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2 - e_1)$ then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies the same condition as $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied. If $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big) \rightarrow \big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 - e_1,\, \boldsymbol{x}_2)$, which can only occur if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied condition *III* or *IV*, then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ preserves the same condition if $a_1 > a_2 + 1$, and if $a_1 = a_2 + 1$ then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies

condition *0* if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *III* or $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *I* if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *IV*. Another possible transition due to service at Q_R in $\boldsymbol{X}^{(2)}$ only is where $(\boldsymbol{x}_1, \boldsymbol{x}_2) \to (\boldsymbol{x}_1', \boldsymbol{x}_2') = (\boldsymbol{x}_1, \boldsymbol{x}_2 - e_1)$. This transition can occur either at rate μ_{R_2} if $a_1 = 0$ but $a_2 > 0$ or at rate $\mu_{R_2} - \mu_{R_1}$ if $a_1 > 0$ and $a_2 > 0$. So if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *0* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *III*; if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *I* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *IV*; if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *II* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ preserves the same condition if $a_2 > a_1 + 1$ or $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *I* if $a_2 = a_1 + 1$. Also if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *III* or *IV* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ preserves the same condition. Note that no service can take place if $a_1 = a_2 = 0$.

Now we consider transitions due to arrivals. First, if $(\boldsymbol{x}_1, \boldsymbol{x}_2) \to (\boldsymbol{x}_1', \boldsymbol{x}_2') = \big((a_1, (n_1+1) \mathrm{mod} N), (a_2, (n_2+1) \mathrm{mod} N)\big)$ then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies the same condition that $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied if $n_1' < N - 1$. If $n_1 = N - 1$ then $\boldsymbol{x}_1' \in H_T$ and if $n_2 = N - 1$ as in conditions *0* and *III* then $\boldsymbol{x}_2' \in H_T$ also. In both cases the desired inequality (2.16) is satisfied. Similarly, if $(\boldsymbol{x}_1', \boldsymbol{x}_2') = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_1)$ then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies whichever condition $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied. If $(\boldsymbol{x}_1, \boldsymbol{x}_2) \to (\boldsymbol{x}_1', \boldsymbol{x}_2') = \big((a_1, (n_1 + 1) \mathrm{mod}\, N), \boldsymbol{x}_2\big)$ then for $n_1' \leq N - 1$ the condition of $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ depends on the condition of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. If $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied any of the conditions *I, II* or *IV* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ preserves the same condition. If $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *0* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *I* and if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *III* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *IV*. Furthermore if $n_1 = N - 1$ then $\boldsymbol{x}_1' \in H_T$. Similarly to the last transition if $(\boldsymbol{x}_1', \boldsymbol{x}_2') = \big((a_1, (n_1 + 1) \mathrm{mod} N), \boldsymbol{x}_2 + e_1\big)$ and $n_1 < N - 1$ then the condition of $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ depends on the condition of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. If $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied any of the conditions *0, I* and *II* then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condition *II*; if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfied condition *III* or *IV* with $a_1 > a_2 + 1$ then $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies condi-

tion *IV*. If $n_1 = N - 1$ then $\boldsymbol{x}_1^{'} \in H_T$. Hence, the inequality (2.16) holds for both cases. If $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2)$ and $n_1 < N - 1$ then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies *III* if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied condition *0* or *III*; if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied condition *I* or *IV* then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies condition *IV*; if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied condition *II* then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ preserves the same condition if $a_2 > a_1 + 1$ or it satisfies the condition *I* if $a_2 = a_1 + 1$. The final transition that we consider is $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big) = \big(\boldsymbol{x} + e_1,\, (a_2, (n_2 + 1) \mathrm{mod} N)\big)$ which can only occur if $\big(\boldsymbol{x}_1,\, \boldsymbol{x}_2\big)$ satisfied condition *II* due to the assumption that D_1 and D_2 are monotone with $T^{(2)} \subset T^{(1)}$. Then $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies condition *0*, if $a_2 = a_1 + 1$ and $n_1 = n_2 + 1$, condition *I*, if $a_2 = a_1 + 1$ but $n_1 > n_2 + 1$, or preserves condition *II*, if $a_2 > a_1 + 1$ and $n_1 > n_2 + 1$.

In each case, either $\boldsymbol{x}_1' \in H_T$ or $n_1' \geq n_2'$ and $\big(\boldsymbol{x}_1^{'},\, \boldsymbol{x}_2^{'}\big)$ satisfies one of the conditions *0* - *IV* . Therefore, $\tau_{D_1}(\boldsymbol{x}_1;\, \Gamma_1) \overset{\text{st}}{\leq} \tau_{D_2}(\boldsymbol{x}_2;\, \Gamma_2)$ for any $\boldsymbol{x}_1,\, \boldsymbol{x}_2 \in \Omega$ satisfying one of these conditions. Thus, $\omega_{D_1}(\boldsymbol{x}_1 - e_2;\, \Gamma_1) \overset{\text{st}}{\leq} \omega_{D_2}(\boldsymbol{x}_2 - e_2;\, \Gamma_2)$ for any $\boldsymbol{x}_1,\, \boldsymbol{x}_2 \in \Omega$. Finally, since $\mu_{T_1} \geq \mu_{T_2}$, we have

$$z_{D_1}\big(\boldsymbol{x}_1;\, \Gamma_1\big) = \mathrm{E}\big(\omega_{D_1}(\boldsymbol{x}_1;\, \Gamma_1)\big) + \frac{1}{\mu_{T_1}} \leq \mathrm{E}\big(\omega_{D_2}(\boldsymbol{x}_2;\, \Gamma_2)\big) + \frac{1}{\mu_{T_2}} = z_{D_2}\big(\boldsymbol{x}_2;\, \Gamma_2\big)$$

as required. ■

We will show the effect of the batch size, N, on $z_{D}(\boldsymbol{x}; \Gamma)$ in the following lemma.

Lemma 2.4.2 *Let* $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{X}^{(2)}$ *be two processes with state space* Ω_1 *and* Ω_2 *respectively. Suppose that the arrival and service rates are the same for both systems, denoted by* $\Gamma(\lambda,\, \lambda_T,\, \mu_R,\, \mu_T)$*, but the batch size parameters* N_1 *and* N_2 *are different. In particular we assume* $N_1 < N_2$*. Also suppose that*

$\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are operating under monotone decision policies $D_1 = (R_1, T_1)$ and $D_2 = (R_2, T_2)$ with $\Omega_1 \subset \Omega_2$. Furthermore, we suppose that $(a, n) \in R_1$ if and only if $(a, n + (N_2 - N_1)) \in R_2$ for $0 \leq n \leq N_1 - 1$ and $a \geq 0$, that is the decisions of D_2 for $(a, n + (N_2 - N_1))$ are identical to D_1 for $a \geq 0$ and $0 \leq n \leq N_1 - 1$. Let $z_{D_i}(\boldsymbol{x}_i)$, $i = 1, 2$ be the expected time to reach the destination via Q_T for a general commuter who arrives at the system $\boldsymbol{X}^{(i)}$ when it is in state $\boldsymbol{x} \in \Omega_i$. Then for $(a, n) \in \Omega_1$,

$$z_{D_1}(a, n) = z_{D_2}(a, n + (N_2 - N_1)) \tag{2.18}$$

Proof. Observe for $n = N_1 - 1$ and all $a \geq 0$, $z_{D_1}(a, N_1 - 1) = \frac{1}{\mu_T} = z_{D_2}(a, (N_1 - 1) + (N_2 - N_1)) = z_{D_2}(a, N_2 - 1)$. Generally $z_D(a, n)$ is determined by the arrival rates, service rates and decision policy for $\boldsymbol{x}'$ such that $n' > n$, and the batch size, where the dependence on the batch size is through the remaining free capacity in the batch. ■

2.4.2 Monotonicity of the user optimal policy $D^\star$ in λ, λ_T, μ_R and μ_T

In this section we show that the user optimal policy changes monotonically with changes in the parameters $\lambda, \lambda_T, \mu_R$ and μ_T in the following theorem.

Theorem 2.4.1 *Let $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ be systems with parameters $\Gamma_1 = (\lambda_1, \mu_{R_1}, \lambda_{T_1}, \mu_{T_1})$ and $\Gamma_2 = (\lambda_2, \mu_{R_2}, \lambda_{T_2}, \mu_{T_2})$ respectively, and common batch size N. Suppose that $\lambda_1 \geq \lambda_2$, $\mu_{R_1} \leq \mu_{R_2}$, $\lambda_{T_1} \geq \lambda_{T_2}$ and $\mu_{T_1} \geq \mu_{T_2}$. Let $D_1^\star$ and $D_2^\star$ be the user optimal policies for $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ respectively. Then*

$a_1^\star(n) \leq a_2^\star(n)$ *for all* $n \in \{0, 1, \ldots, N-1\}$.

The following proof is similar in spirit to the proof of Theorem 2.3.2.

Proof We use a proof by induction $\xi = N - n$, where $\xi = 1, 2, \ldots, N$. We begin with $\xi = 1$, that is, $n = N - 1$. Let $y_i(\boldsymbol{x}; \Gamma_i)$ and $z_{D_i}(\boldsymbol{x}; \Gamma_i)$ denote the expected transit times via Q_R and Q_T respectively for a general user who on arrival finds $\boldsymbol{X}^{(i)}$ in state $\boldsymbol{x} = (a, n)$. Since $z_{D_1^\star}(a, N-1; \Gamma_1) = \frac{1}{\mu_{T_1}} \leq \frac{1}{\mu_{T_2}} = z_{D_2^\star}(a, N-1; \Gamma_2)$ and $y_2(a, N-1; \Gamma_2) = \frac{a+1}{\mu_{R_2}} \leq \frac{a+1}{\mu_{R_1}} = y_1(a, N-1; \Gamma_1)$, we have $a_1^\star(N-1) \leq a_2^\star(N-1)$.

Now, suppose that the inductive hypothesis holds for $\xi = m$ that is, $a_1^\star(n) \leq a_2^\star(n)$ for all n such that $N - m \leq n \leq N - 1$. We will show that it also holds for $\xi = m + 1$, that is, $n = N - m - 1$. Let $D \in \mathcal{D}_m^\star$ be an $m-$level user optimal monotone policy for $\boldsymbol{X}^{(1)}$. Note that $y_1(\boldsymbol{x}; \Gamma_1) = \frac{a+1}{\mu_{R_1}}$ is independent of both D and n but increases in a. Let $\beta = \min\{a : z_D(a, N-m-1; \Gamma_1) \leq y_1(a, N-m-1; \Gamma_1)\}$ then by the proof of Theorem 2.3.2 $\beta = a_1^\star(N-m-1)$. If we let $a_D^\star(n) = \min\{a : z_D(a, n; \Gamma_1) \leq y_1(a, n; \Gamma_1)\}$ for $N - m \leq n \leq N - 1$ then by the inductive hypothesis $a_D^\star(n) \leq a_2^\star(n)$ for $N - m \leq n \leq N - 1$. Now use the fact that $z_D(\boldsymbol{x}; \Gamma_1)$ depends on D only for states $\boldsymbol{x}_1 \in \Omega$ such that $n_1 > n$ and since $a_D^\star(n) \leq a_2^\star(n)$ for all n such that $N-m \leq n \leq N-1$, we have $z_D(a, N-m-1; \Gamma_1) \geq z_D(a, N-m; \Gamma_1)$ and $z_{D_2^\star}(a, N-m-1; \Gamma_2) \geq z_{D_2^\star}(a, N-m; \Gamma_2)$ for all $a \geq 0$ by Corollary 2.3.1 (condition I). Then by Lemma 2.4.1, $z_D(a, N-m-1; \Gamma_1) \leq z_{D_2^\star}(a, N-m-1; \Gamma_2)$ for all $a \geq 0$. By Corollary 2.3.1 (condition III), $z_D(a, N-m-1; \Gamma_1)$ and $z_{D_2^\star}(a, N-m-1; \Gamma_2)$ are decreasing in a, and since $y(a, N-m-1; \Gamma_1) = \frac{a+1}{\mu_{R_1}} \geq \frac{a+1}{\mu_{R_2}} = y_2(a, N-m-1; \Gamma_2)$, we have $a_D^\star(N-m-1) \leq a_2^\star(N-m-1)$ and therefore, $a_1^\star(N-m-1) \leq a_2^\star(N-m-1)$. Hence, by the principle of induction, the theorem holds. ■

For any batch size, N, and fixed system parameter values $\lambda > 0, \lambda_T > 0, \mu_R > 0$ and $\mu_T > 0$, the user optimal policy depends only on the amount of free space left in a batch (see Lemma 2.4.3).

Lemma 2.4.3 *Suppose* $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{X}^{(2)}$ *denote two systems with the same parameters* $\Gamma = (\lambda, \mu_R, \lambda_T, \mu_T)$ *but different batch sizes* N_1 *and* N_2*. We assume* $N_1 < N_2$*. Let* Ω_1 *and* Ω_2 *denote the state space for* $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{X}^{(2)}$ *respectively and* $\Omega_1 \subset \Omega_2$*. For* $(a, n) \in \Omega_1$, $(a, n) \in R_1^\star$ *if and only if* $(a,\, n+N_2-N_1) \in R_2^\star$ *for* $(a,\, n + N_2 - N_1) \in \Omega_2$ *or equivalently* $a_1^\star(n) = a_2^\star(n + N_2 - N_1)$.

Proof Observe that $z_{D_1}(a, N_1 - 1) = 1/\mu_T = z_{D_2}(a,\, (N_1 - 1) + (N_2 - N_1)) = z_{D_2}(a,\, N_2 - 1)$ since $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ have common parameters λ, λ_T, μ_R and μ_T. Also, since $y(\boldsymbol{x})$ are independent of the decision policies D_1 and D_2, $y_1(a, n) = y_2(a, n+(N_2-N_1)) = \frac{a+1}{\mu_R}$ for all $a \geq 0$. Thus, we have $a_1^\star(N_1-1) = a_2^\star(N_2 - 1)$. Furthermore, since $z_{D_i}(\boldsymbol{x}_i)$ is determined by decision policy for $\boldsymbol{x}'$ such that $n' > n$ and batch size, as in the proof of Lemma 2.4.2, it follows that $z_{D_1^\star}(a, n) = z_{D_2^\star}(a,\, n + (N_2 - N_1))$ for all $(a, n) \in \Omega_1$. Finally, $z_{D_i}(\boldsymbol{x}_i)$ are decreasing in a, while $y_i(\boldsymbol{x}_i)$ are increasing in a. Hence, $a_1^\star(n) = a_2^\star(n+N_2-N_1)$. ■

Figure 2.3 gives an example of Lemma 2.4.3 with $\lambda = \lambda_T = 1$, $\mu_R = \mu_T = 2$ and $N_1 = 2$ and $N_2 = 4$. Let (a, n) and $(a, n + 2)$ denote the states of the systems with batch sizes $N_1 = 2$ and $N_2 = 4$, respectively. Also let $a_1^\star(n) = \min\{a : z_{D_1}((a, n); \Gamma) \leq y_{D_1}((a, n); \Gamma)\}$ and $a_2^\star(n + 2) = \min\{a : z_{D_2}((a,\, n + 2); \Gamma) \leq y_{D_2}((a,\, n + 2); \Gamma)\}$. From this example we observe that $a_1^\star(0) = a_2^\star(2) = 1$ and $a_1^\star(1) = a_2^\star(3) = 0$.

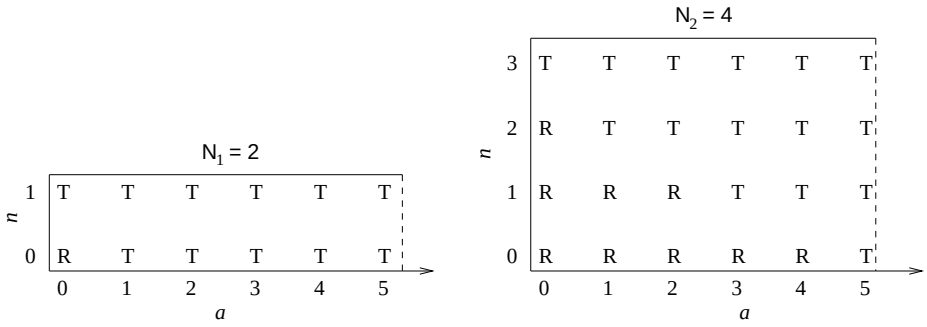


Figure 2.3: Example of two systems with batch sizes $N_1 = 2$ and $N_2 = 4$ while the other parameters are fixed as $\lambda = \lambda_T = 1$, $\mu_R = \mu_T = 2$ for both.

Figure 2.4 plots the switching curve for $N = 10$, 20, 40, 60, 80 and 100 with $\lambda = \lambda_T = \mu_T = 1$ and $\mu_R = 3$. We see in this numerical example that the switching curve is approximately linear even in large N.

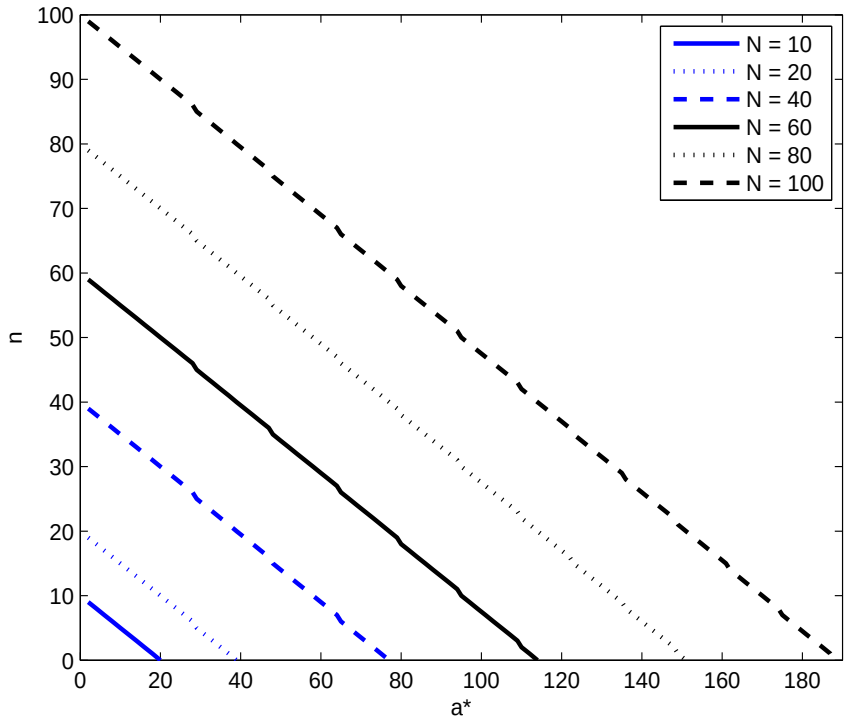


Figure 2.4: Plot of the switching curve for batch sizes $N = 10$, 20, 40, 60, 80 and 100 while the other parameters are fixed as $\lambda = \lambda_T = \mu_T = 1$, $\mu_R = 3$.

2.4.3 Numerical examples

From section 2.2 (equation 2.8), $W = \sum_{\boldsymbol{x}\in\Omega_{D^\star}} \pi_{D^\star}(\boldsymbol{x}) \left(y(\boldsymbol{x})\, I_{\{\boldsymbol{x}\in R^\star\}} + z_{D^\star}(\boldsymbol{x})\, I_{\{\boldsymbol{x}\in T^\star\}}\right)$, is the overall expected time spent in the system (including service time) by a general user, where $D^\star$, and $\pi_{D^\star}$ denote the user optimal policy and the stationary distribution of the process $X_{D^\star}$. Note that we only consider the recurrent states of $D^\star$ for the calculating of W. In this subsection we give numerical examples which indicate that W does not change monotonically with the change of λ and μ_R. Furthermore, we give numerical examples in support of the conjecture that W decreases monotonically with increasing λ_T and μ_T. In general it is not easy to show the relationship of W and the parameters analytically since each time we change the value of one parameter we have to recalculate $z_{D^\star}(\boldsymbol{x};\Gamma)$ and $y(\boldsymbol{x};\Gamma)$, if μ_R changes, as well as $D^\star$ and $\pi_{D^\star}(\boldsymbol{x})$. From subsection 2.4.2, $D^\star$ changes monotonically with changes in any of the parameters λ, μ_R, λ_T and μ_T.

The following numerical examples were produced using the Matlab code in Appendix A.1. This code was tested using the simple example in [15], where $\lambda = \lambda_T = \mu_R = \mu_T = 1$ and $N = 2$. We also checked the code by examining some special cases, such as when the confirmed train users arrival rate λ_T is near zero and when it is very large. If λ_T is near zero then we expect general users to almost always choose Q_R and if λ_T is very large then we expect general users to almost always choose Q_T.

We begin with the two extreme cases where general users either always choose Q_R, that is, $R = \Omega$ or general users always choose Q_T, that is, $T = \Omega$. We consider first $R = \Omega$. This policy is not the user optimal policy if $\lambda_T > 0$,

which we consider in this analysis. In this case W is just that for a single server queue Q_R that is, $W = 1/(\mu_R - \lambda)$ provided that $\lambda < \mu_R$. Note that if $\lambda > \mu_R$ then the system is unstable under this policy ($R = \Omega$). Also note for this policy that W increases in λ but decreases in μ_R, for $\lambda < \mu_R$. This is not always the case with other policies. The second case we consider is when $T = \Omega$, and under this policy $W = 1/\mu_T + (N-1)/(2(\lambda+\lambda_T))$. The system is always stable under this policy and in this case W is monotonically decreasing in any of the parameters λ, λ_T and μ_T although W is not monotone in general.

Figure 2.5 plots the expected transit time, W, under the user optimal policy, $D^\star$. The left-hand plot shows W versus the general stream arrival rate, λ, and the right-hand plot is of W versus the service rate at Q_R, μ_R. For the left plot we set $\lambda_T = 0.1$, $\mu_R = \mu_T = 1$, $N = 3$ and allow λ to vary between 0.01 and 3 and for the right plot we set $\lambda = 1$, $\lambda_T = 0.1$, $\mu_T = 1$, $N = 3$ and allow μ_R to vary between 0.01 and 3 as well. For both plots W is calculated at increments of 0.01 We see in these examples that W is monotone neither in λ nor in μ_R, although the user optimal policy, $D^\star$, is monotone in both of these parameters as shown in Theorem 2.4.1.

In the left-hand plot in Figure 2.5, where λ is varying, when $\lambda < 0.4$ with $\lambda_T = 0.1, \mu_R = \mu_T = 1$ and $N = 3$, the user optimal policy $D^\star = (R^\star, T^\star)$ has
$R^\star = \{(0,0),(0,1),(1,0),(1,1),(1,2),(2,0),(3,0),(4,0),(5,0)\}$
and
$T^\star = \{(0,2),(2,1),(2,2),(3,1),(3,2),(4,1),(4,2),(5,1),(5,2),(6,0),(6,1),(6,2)\}$
and the expected delay in the system, W, increases as λ increases with $W = 1.4891$ when $\lambda = 0.399$. When $\lambda = 0.4$ the user optimal policy $D_1^\star$ has $R_1^\star = R_1^\star \setminus (1,2)$ and $T_1^\star = T_1^\star \cup \{(1,2)\}$, that is, $T_1^\star$ contains one state more

than when $\lambda = 0.399$ and now $W = 1.4601 < 1.4891$. For $\lambda \in (0.4,\ 0.561]$ the user optimal policy $D_2^\star$ has $R_2^\star = \{(0,0),(0,1),(1,0),(1,1),(2,0),(3,0),(4,0)\}$ and W increases from 1.4601 to 1.6930, and when $\lambda = 0.562$ $D^\star$ has $R_3 = \{(0,0),(0,1),(1,0),(1,1),(2,0),(3,0)$ and $W = 1.6488 < 1.6930$, which explain the sharp decrease in W at the point where the new user optimal policy is introduced. In this example it can be seen that as λ increases, $R^\star$ decreases monotonically as expected while W fluctuates when λ is sufficiently small and continues to decrease when $\lambda > 1$.

In the right-hand plot in Figure 2.5, where μ_R is varying, when $\mu_R < 0.3549$, $R^\star = \phi$ and $W = 1.9091$. At $\mu_R = 0.3549$, $R_1^\star = \{(0,0)\}$ and $T_1^\star = \{(0,1),(0,2),(1,0),(1,1),(1,2)\}$, and $W = 2.0807 > 1.9091$, which explains the sharp increase in W seen in the plot. This is what is known as the Down-Thomson effect but it is much reduced in comparison to that under probabilistic routing. W continues to decrease when μ_R increases until a new policy is introduced. In this example, it can be seen that as μ_R increases, $R^\star$ increases as expected but W is not monotone in μ_R.

Figure 2.6 plots W versus λ_T and W versus μ_T for $N = 3, 10$ and 20. For the first plot, we set $\lambda = 1$, $\mu_R = \mu_T = 1$ and allow λ_T to vary between 0.01 and 3. For the second plot we set $\lambda = \lambda_T = 1$, $\mu_R = 1$ and vary μ_T between 0.01 and 3 as well. We see that W is monotonically decreasing in both λ_T and μ_T.

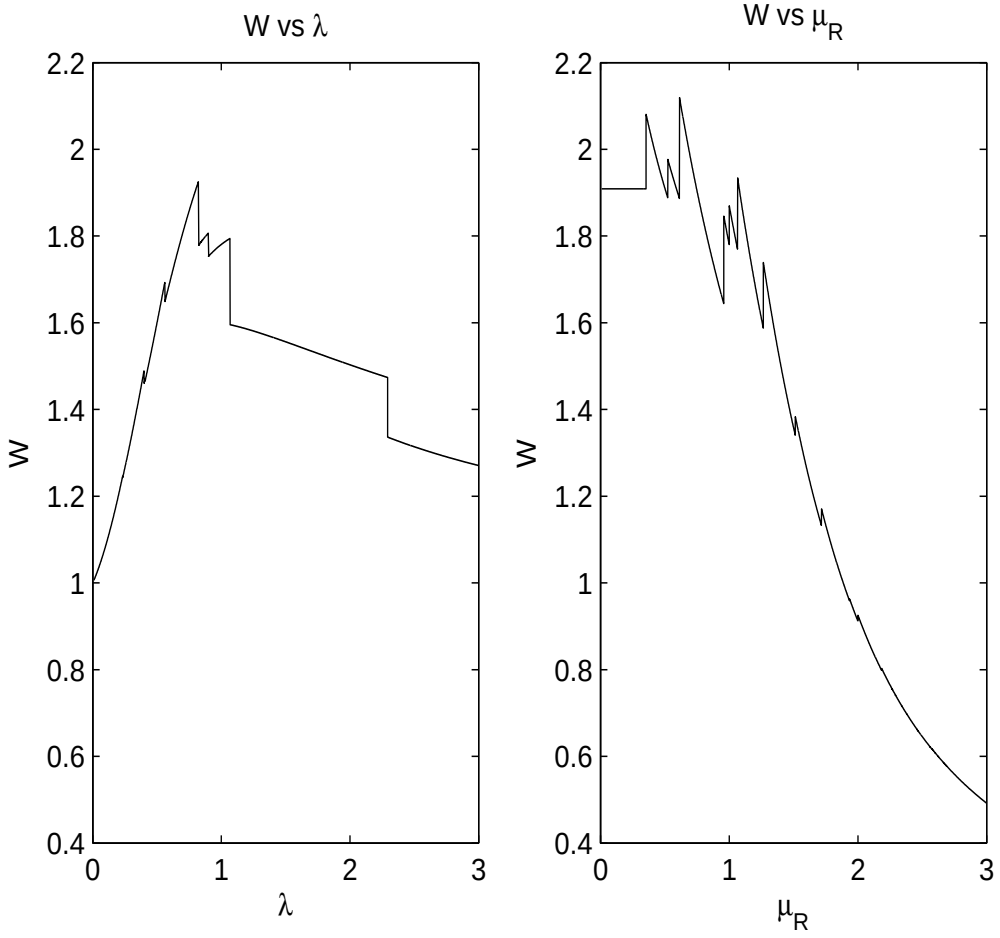


Figure 2.5: Left: The expected transit time, W, when $\lambda_T = 0.1, \mu_R = \mu_T = 1, N = 3$ and $0.01 \leq \lambda \leq 3$. Right: W when $\lambda = 1, \lambda_T = 0.1, \mu_T = 1, N = 3$ and $0.01 \leq \mu_R \leq 3$.

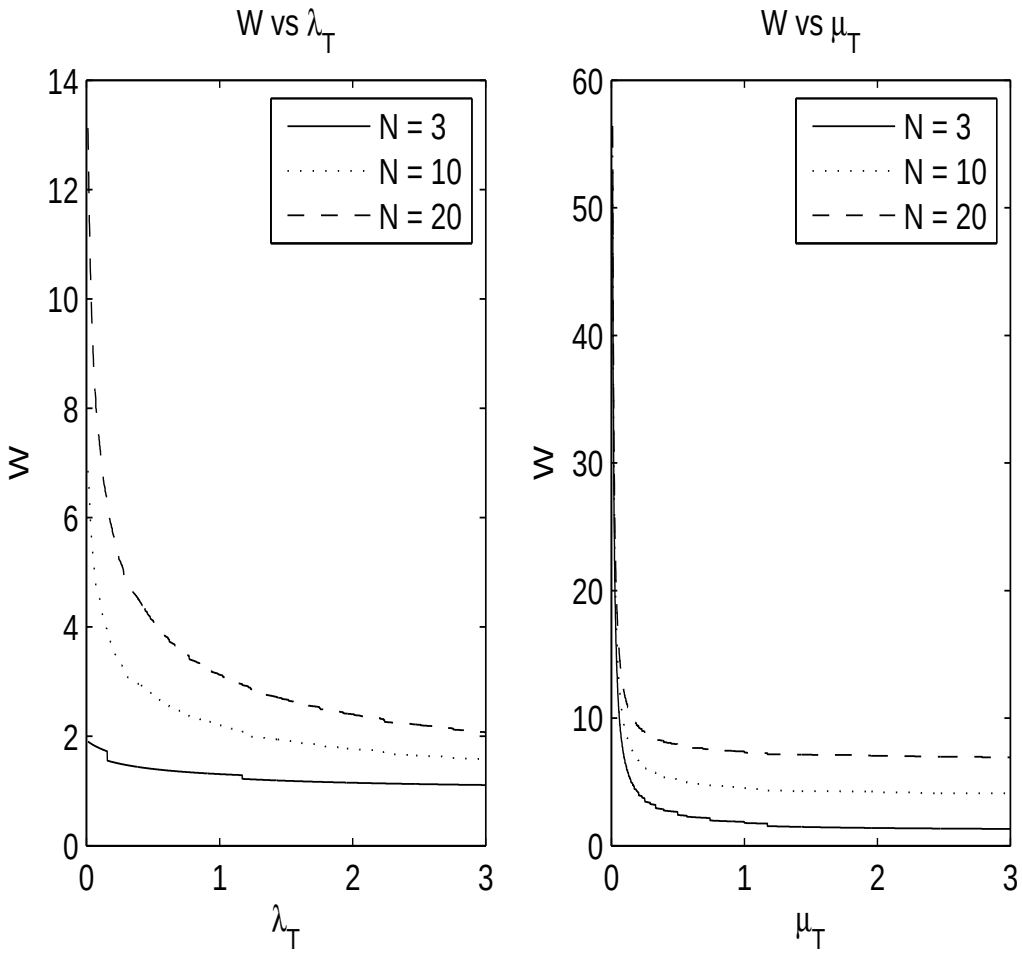


Figure 2.6: Left: The expected transit time, W, when $\lambda = \mu_R = \mu_T = 1$, and $0.01 \leq \lambda_T \leq 3$. Right: W when $\lambda_T = 0.1, \lambda = \mu_R = 1$ and $0.01 \leq \mu_T \leq 3$ for $N = 3, 10$ and 20.

2.5 Discussion and conclusion

We have given a proof of existence, uniqueness and monotonicity properties of the user optimal policy (or the user equilibrium) $D^\star$, under state-dependent routing in a simple queueing network using coupling arguments. Furthermore, we have shown that $D^\star$ is monotone in the parameters λ, λ_T, μ_R and μ_T, and have given examples showing that the expected time spent in the system by a general user, W, is not monotone in λ and μ_R.

The Downs-Thomson effect also exists in this routing case as we see in Figure 2.5, but in our numerical examples it is not as pronounced as that in the probabilistic routing case seen in the numerical example in [2]. We revisit the numerical example studied in [2], with $\lambda = 1, \lambda_T = 0.1, \mu_R = 1, N = 3$ while μ_R varies between 0 and 3, and plot the mean transit time, W, under the user optimal policy in Figure 2.7 for both probabilistic and state-dependent routing. The most interesting feature in this plot is the considerable reduction of the Downs-Thomson effect observed under state-dependent routing in comparison with probabilistic routing. That is, there is no sudden and sharp increase in W as μ_R increases under the state-dependent routing. We also provide similar plots of W versus λ and W versus λ_T in Figures 2.8 and 2.9, respectively. From these plots we see that the poor performance of the user optimal policy under probabilistic routing is substantially reduced under state-dependent routing. From Figures 2.7 and 2.8 we also note that state-dependent routing may give a user optimal policy where W is slightly greater than that under probabilistic routing. This needs to be investigated further for this system.

The following numerical examples were produced using the Matlab code in

Appendix A.1.

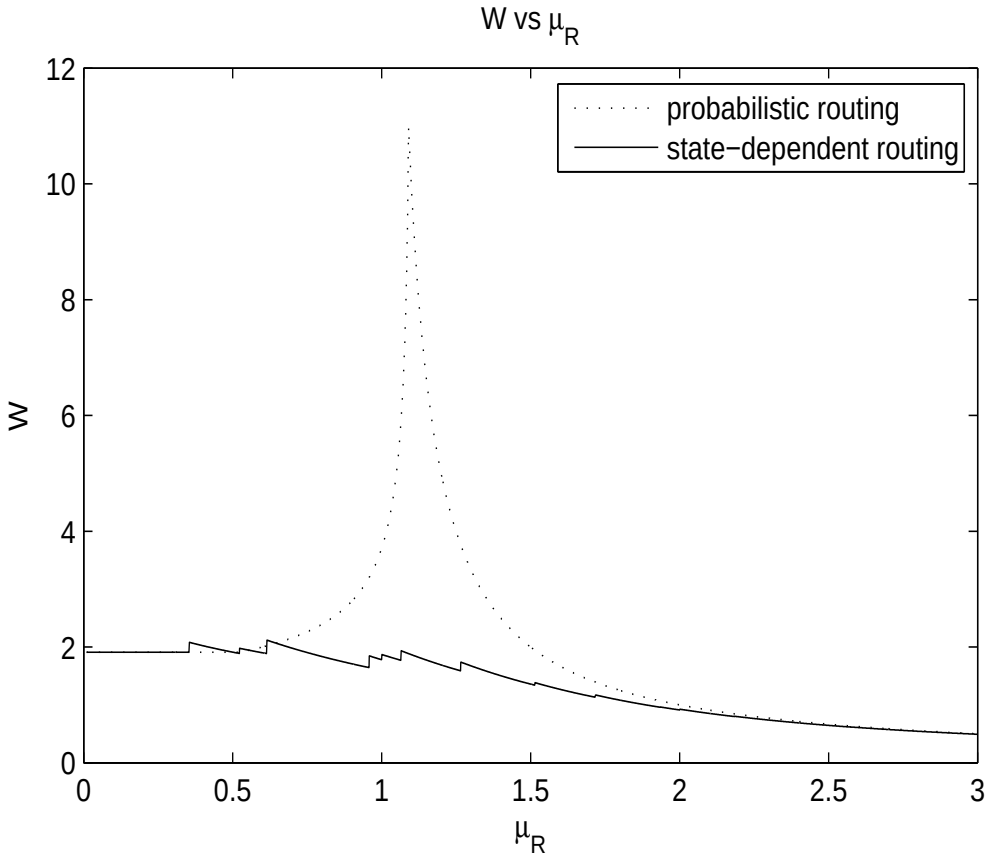


Figure 2.7: The expected transit time, W, when $\lambda = \mu_T = 1, \lambda_T = 0.1, N = 3$ and $0.01 \leq \mu_R \leq 3$ under both probabilistic and state-dependent routing.

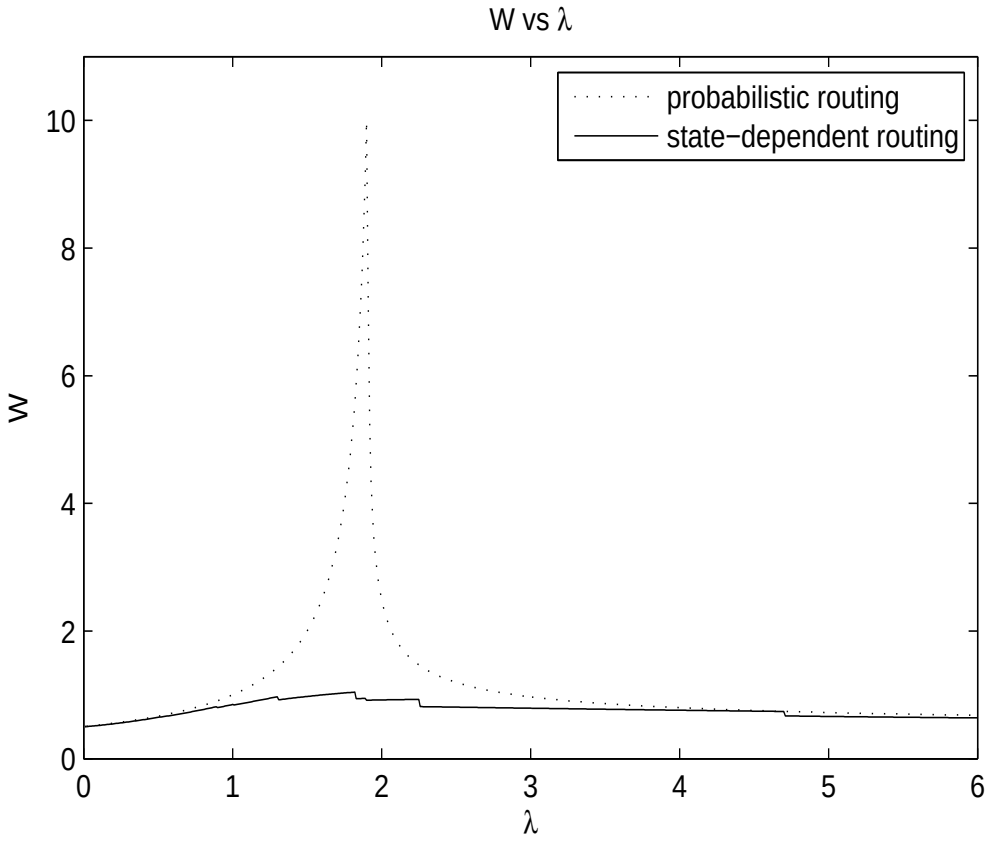


Figure 2.8: The expected transit time, W, when $\mu_R = \mu_T = 2, \lambda_T = 0.1, N = 3$ and $0.01 \leq \lambda \leq 6$ under probabilistic and state-dependent routing.

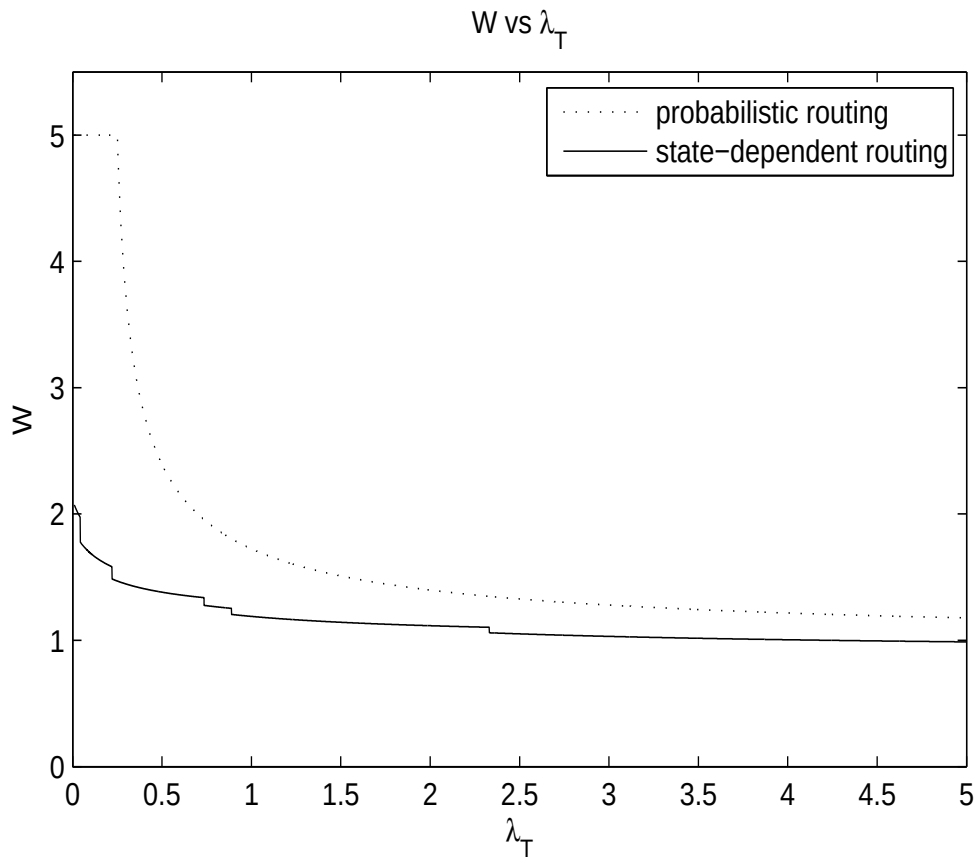


Figure 2.9: The expected transit time, W, when $\lambda = 1, \mu_R = 1.2, \mu_T = 1, N = 3$ and $0.01 \leq \lambda_T \leq 5$ under both probabilistic and state-dependent routing.

Chapter 3

Probabilistic routing for $\cdot|G|1$ and $\cdot|G^{(N)}|\infty$ queues

3.1 Introduction

In this chapter we consider the same queueing network that we studied in the last chapter but the single server service queue Q_R is modelled as an $M|G|1$ queue instead of an $M|M|1$ queue while the batch service queue Q_T is an $M|G^{(N)}|\infty$ queue with batch size $N \geq 2$ and service rate μ_T as before. That is, we consider generally distributed service times for the single server queue with mean $1/\mu_R$, since in practice not all service times are exponentially distributed. As defined before there are two streams of arrivals into this system, a stream of confirmed train users to Q_T, at rate λ_T, and a stream of general users arriving at rate λ who can join Q_R and Q_T. The delay in Q_R *increases* as the load increases while that in Q_T *decreases* as the load increases. We assume for this system

that general users know the arrival rates and the service rates only and they choose the queue which will minimize their own expected delay independently of the current state of the system, that is, we are considering probabilistic routing. We assume that arriving general users choose Q_R with probability p_R and Q_T with probability $1 - p_R$. Then we seek a routing strategy p_R such that each general user has no incentive to deviate from it if all other users use it. Such a strategy policy is known as a user equilibrium, user optimal policy or Wardrop equilibrium. In the remainder of this chapter we will often refer to it as a user equilibrium. We also assume for this system that all inter-arrival times, routing decisions and service times are independent of one another.

In this chapter we study the existence, uniqueness and stability properties of the user equilibrium for a network consisting of an $M|G|1$ queue and $M|G^{(N)}|\infty$ queue. Calvert [15] considered the same network with Q_R as an $M|M|1$ queue instead of $M|G|1$ queue and provided a single numerical example of the existence of the user equilibrium and the Downs-Thomson paradox. As we mentioned in Chapter 1, Afimeimounga, Solomon and Ziedins [2] consider the same network model as Calvert [15] and completely analyzed the existence, uniqueness and stability properties of user equilibria for that model. They showed that multiple user equilibria exist only if $N = 2$ and a unique stable user equilibrium always exists if $N \geq 3$. This is not the case for the system model we consider in this chapter. We will provide numerical examples of the existence of multiple user equilibria for $N \geq 2$ in section 3.3. The formal definition of the user equilibria for this system is given in [2] p. 325. We also show that the stable user equilibria we obtain are not all evolutionarily stable strategies (ESS). Furthermore, we show that the system possesses the avoid the crowd (ATC) property if the user equilibrium is stable and it possesses the

follow the crowd (FTC) property if the user equilibrium is unstable.

The outline of this chapter is as follows. We provide the definitions and some preliminary results in section 3.2 and then the analysis of the user equilibria in section 3.3. We examine the relationship of the user equilibria with ESS and social optima and check when the system has the ATC property or FTC property in section 3.4. Finally we give a conclusion and discussion in section 3.5.

3.2 Definitions and Preliminaries

For the model described in section 3.1, the state space is given by $\Omega = \{(a,\, n) \in \mathbb{Z}^+ \times \{0, 1, \ldots, N-1\}\}$, where a is the total number of commuters in Q_R and n is the number waiting in Q_T for service, since we assume the train leaves immediately after the arrival of Nth commuter and so we cannot see N commuters in Q_T.

Let $W_T(p_R)$ and $W_R(p_R)$ be the expected time taken in equilibrium to reach the destination by a general user joining Q_T and Q_R respectively. Also let $\Lambda_T = \lambda_T + (1 - p_R)\,\lambda$. Then from Lemma 2.1 in [2],

$$W_T(p_R) = \frac{1}{\mu_T} + \frac{N-1}{2\,\Lambda_T},$$

where $1/\mu_T$ is the expected travel time from the source to the destination and $(N-1)/(2\,\Lambda_T)$ is the expected time that a user spends in Q_T before departure. Note for this queue that the only possible transitions are to move from state n to $(n+1) \bmod N$, which happens at a uniform rate Λ_T. So in

equilibrium Q_T is in state i, $0 \leq i \leq N-1$ with probability $1/N$. Therefore, an arrival is equally likely to see Q_T in any state i, and the expected time they have to wait until their service commences given they see Q_T in state i is $(N-(i+1))/\Lambda_T$. Hence, the expected time that a user spends waiting for service in Q_T is $\frac{1}{N}\sum_{i=0}^{N-1}\frac{N-(i+1)}{\Lambda_T} = \frac{N-1}{2\,\Lambda_T}$ (see [2] for more detail).

The expression for W_R is obtained by using the Pollaczek-Khinchin formula (3.1) and Little's Law. Let L_R denote the expected number of commuters in Q_R including any commuter in service. Then

$$\begin{aligned} L_R &= \rho + \frac{\rho^2\,(1+c^2)}{2\,(1-\rho)},\ \text{if } \rho < 1 \qquad (3.1)\\ &= \frac{\lambda\,p_R}{\mu_R} + \frac{\left(\frac{\lambda\,p_R}{\mu_R}\right)^2(1+c^2)}{2\left(1-\frac{\lambda\,p_R}{\mu_R}\right)} = \lambda\,p_R\left(\frac{1}{\mu_R} + \frac{\lambda\,p_R\,(1+c^2)}{2\mu_R\,(\mu_R-\lambda\,p_R)}\right) \end{aligned}$$

where $\rho = \frac{\lambda\,p_R}{\mu_R}$ is the utilization rate, $c^2 = Var\,(S)\,/E\,(S)^2$ and S is a random variable having the distribution of a service time in Q_R. Then by Little's Law

$$\begin{aligned} W_R(p_R) &= L_R\,/\,\lambda\,p_R \qquad (3.2)\\ &= \frac{1}{\mu_R} + \frac{\lambda\,p_R\,(1+c^2)}{2\mu_R\,(\mu_R-\lambda\,p_R)},\quad \text{provided that } \rho < 1 \qquad (3.3) \end{aligned}$$

Since the $M|M|1$ queue, which is a special case of the $M|G|1$ queue, was already studied in [2] we keep our variable names as similar as possible to those used there. If the coefficient of variation is 1, that is, $c^2 = 1$, then (3.3) simplifies to $1/(\mu_R - \lambda\,p_R)$ as seen in [2].

As in [2], we let

$$W(p_R) = p_R\, W_R(p_R) + (1 - p_R)\, W_T(p_R) \tag{3.4}$$

be the expected steady state transit time for all general users. A user equilibrium or Wardrop equilibrium $p_R^{EQ} \in [0,\, 1]$ for general users requires the transit times on all routes in use to be equal and less than those which would be experienced by a single user on any unused route. As in [2] Definition 2.2 we define three cases.

(A) $p_R^{\mathrm{EQ}} = 0$ if $W_T\,(0) \leq W_R\,(0)$,

(B) $p_R^{EQ} = 1$ if $W_T\,(1) \geq W_R\,(1)$ and

(C) $p_R^{EQ} \in (0,\, 1)$ if $W_T\left(p_R^{EQ}\right) = W_R\left(p_R^{EQ}\right)$.

We will refer to cases A and B as pure user equilibria and to case C as a mixed user equilibrium. Furthermore, a user equilibrium is called a stable user equilibrium $p_R^{\star} \in [0,\, 1]$ if

(a) $W_R(p) > W_T(p)$ for $p \in (p_R^{\mathrm{EQ}},\, \min(p_R^{\mathrm{EQ}} + \epsilon, 1))$ and

(b) $W_T(p) > W_R(p)$ for $p \in (\max(p_R^{\mathrm{EQ}} - \epsilon, 0),\, p_R^{\mathrm{EQ}})$ with $\epsilon > 0$ (see [2] Definition 2.3).

From these definitions, if there exists a stable user equilibrium $p_R^{\star} \in (0,\, 1)$ and the current strategy is $p \in (p_R^{\star},\, p_R^{\star} + \epsilon)$ (so that $W_R(p) > W_T(p)$) then an individual general user will do better by *decreasing* their likelihood of joining Q_R, whereas if the current strategy is $p \in (p_R^{\star},\, p_R^{\star} - \epsilon)$ (so that $W_R(p) < W_T(p)$)

then an individual general user will do better by *increasing* their likelihood of joining Q_R.

For $M|M|1$ and $M|G^{(N)}|\infty$ queues, (i.e. $c^2 = 1$) and $N = 2$ four cases may occur (see [2]). In the first case a unique stable user equilibrium exists at $p_R^\star = 0$ or $p_R^\star = 1$. The second case is when a mixed stable user equilibrium (i.e. $p_R^\star \in (0, 1)$) exists and the third case is when an unstable mixed user equilibrium occurs (i.e. $p_R^{\mathrm{EQ}} \in (0, 1)$) with two stable user equilibria occurring at $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$. The last case is when two mixed user equilibria, $p_R^{EQ,1}$ and $p_R^{EQ,2}$, occur in the interval (0, 1) with two stable equilibria at $p_R^{\star,1} = p_R^{\mathrm{EQ},1}$ and $p_R^{\star,2} = 1$. Note that $p_R^{\mathrm{EQ},2}$ is an unstable user equilibrium. However, when $N \geq 3$ there always exists a unique and stable user equilibrium $p_R^\star \in [0, 1]$. This is not the case when the service time is generally distributed as we will see in the next section.

3.3 User equilibria (optimal policies)

Although the Pollaczek-Khinchin formula gives the expected delay in Q_R for the system we study here we have not obtained the values of the system parameters where multiple user equilibria exist as we did in [2], due to the increased complexity of the formulae. However, we will provide some numerical examples of the existence of multiple user equilibria for this system when $N \geq 3$ and the service time is generally distributed, in contrast with the case when the service time is exponentially distributed, with $c^2 = 1$ and $N \geq 3$, when the user equilibrium is unique and stable.

As in Remark 2 of [2], without loss of generality, one of the parameters

λ, λ_T, μ_R, or μ_T can be fixed, and so we fix $\lambda = 1$. Then W_T and W_R are rewritten as follows;

$$W_T(p_R) = \frac{1}{\mu_T} + \frac{N-1}{2\,(\lambda_T + 1 - p_R)} \tag{3.5}$$

$$W_R(p_R) = \frac{1}{\mu_R} + \frac{(1+c^2)}{2\mu_R}\,\frac{p_R}{\mu_R - p_R}. \tag{3.6}$$

From (3.5) and (3.6), W_T and W_R both increase in p_R if we fix the other variable values (λ_T, μ_R, μ_T, c and N). For stability of this system, we require that $p_R < \mu_R$. Since our definitions of user equilibrium and stable user equilibrium are in terms of the comparison between $W_T(p_R)$ and $W_R(p_R)$ for $p_R \in [0,\, 1]$, we therefore consider the difference of W_T and W_R for this model, $W_T - W_R$. Then

$$\begin{aligned}
\left[W_T - W_R\right](p_R) &= \frac{1}{\mu_T} + \frac{N-1}{2\,(\lambda_T + 1 - p_R)} - \frac{1}{\mu_R} - \frac{(1+c^2)\,p_R}{2\,\mu_R(\mu_R - p_R)} \\
&= \frac{Q(p_R)}{\theta} = \frac{\alpha\,p_R^2 - \beta\,p_R + \gamma}{\theta}
\end{aligned} \tag{3.7}$$

where

$$\alpha = \mu_R + \frac{c^2-1}{2}\mu_T, \text{ with } \alpha > 0 \text{ if } c > \sqrt{1 - 2\,\frac{\mu_R}{\mu_T}} \tag{3.8}$$

$$\beta = \mu_R^2 + \big(1 + \lambda_T + \frac{N-3}{2}\,\mu_T\big)\mu_R + \frac{1}{2}(1+\lambda_T)(c^2-1)\mu_T \tag{3.9}$$

$$\gamma = \big(1 + \lambda_T + \frac{N-1}{2}\,\mu_T\big)\,\mu_R^2 - (1+\lambda_T)\,\mu_T\mu_R \tag{3.10}$$

$$\theta = \mu_T\,(1 + \lambda_T - p_R)\,(\mu_R - p_R)\,\mu_R. \tag{3.11}$$

From (3.7), $Q(p_R)$ is a quadratic in p_R with $\lim_{p_R \to \pm\infty} Q(p_R) = +\infty$ if $\alpha > 0$ and $\lim_{p_R \to \pm\infty} Q(p_R) = -\infty$ if $\alpha < 0$. If p_R is a zero of $Q(\cdot)$ such that $p_R \in (0,\, 1)$ then $W_T(p_R) = W_R(p_R)$ and a mixed user equilibrium exists. We begin by showing that it is possible for this system to have multiple user equilibria. However, to simplify matters we consider just the case $N = 3$ in the following argument. If $N = 3$ then

$$\begin{aligned} [W_T - W_R]\,(p_R) &= \frac{1}{\mu_T} + \frac{1}{\lambda_T + 1 - p_R} - \frac{1}{\mu_R} - \frac{(1+c^2)\,p_R}{2\,\mu_R(\mu_R - p_R)} \\ &= \frac{Q(p_R)}{\theta} = \frac{\alpha\,p_R^2 - \beta\,p_R + \gamma}{\theta} \end{aligned} \tag{3.12}$$

where α, γ, and θ are as in (3.8), (3.10) and (3.11) while β in (3.9) becomes

$$\beta = \mu_R^2 + \big(1 + \lambda_T\big)\mu_R + \frac{1}{2}(1+\lambda_T)(c^2-1)\mu_T = \mu_R^2 + (1+\lambda_T)\,\alpha. \tag{3.13}$$

$Q(p_R)$ has no real root if and only if

$$\begin{aligned} &\beta^2 - 4\,\alpha\,\gamma < 0 \qquad (3.14) \\ \Leftrightarrow\quad &\big(\mu_R^2 + (1+\lambda_T)\,\alpha\big)^2 - 4\,\big((1+\lambda_T+\mu_T)\mu_R - (1+\lambda_T)\mu_T\big)\,\alpha\,\mu_R < 0 \end{aligned}$$

If $Q(p_R)$ does have real roots, denote them by

$$p_\pm = \frac{\beta \pm \sqrt{\beta^2 - 4\,\alpha\,\gamma}}{2\,\alpha}$$

Now

$$Q(0) = \gamma \text{ and } \gamma < 0 \text{ iff } \mu_R < \mu_0,$$

where

$$\mu_0 \equiv \Big(\frac{1+\lambda_T}{1+\lambda_T+\mu_T}\Big)\,\mu_T. \tag{3.15}$$

Also

$$Q(1) = \alpha - \beta + \gamma \text{ and } Q(1) < 0 \text{ iff } \mu_{1_-} < \mu_R < \mu_{1_+},$$

where

$$\mu_{1_\mp} \equiv \frac{1}{2}\Big(1 + \frac{\lambda_T\,\mu_T}{\lambda_T+\mu_T} \pm \sqrt{\big(1+\frac{\lambda_T\,\mu_T}{\lambda_T+\mu_T}\big)^2 + 2\,\frac{\lambda_T\,\mu_T}{\lambda_T+\mu_T}\,(c^2-1)}\,\Big) \tag{3.16}$$

Now let δ be the value of p_R that minimizes or maximizes $Q(p_R)$. Then solving $\frac{d}{dp_R}\big(Q(p_R)\big) = 2\,\alpha\,p_R - \beta = 0$ gives

$$\delta = \frac{\beta}{2\,\alpha} = \frac{(1+\lambda_T)\,\alpha + \mu_R^2}{2\,\alpha} = \frac{1}{2}\big(1 + \lambda_T + \frac{\mu_R^2}{\alpha}\big) \tag{3.17}$$

We use the above information to find examples of $[W_T - W_R](\cdot)$ possessing two real roots in the interval $[0, 1]$ when $N = 3$. We provide examples for $\alpha = \mu_R + \frac{c^2-1}{2}\mu_T > 0$ in Example 3.3.1 and Example 3.3.2 and that for $\alpha < 0$ in Example 3.3.3. From Examples 3.3.1 and 3.3.3 multiple equilibria exist for this system when $N = 3$. We also provide examples of the existence of multiple user equilibria for $N > 3$.

Example 3.3.1 *$N = 3$, $\mu_T = 3$, $\lambda = 1$, $\lambda_T = 0.1$ and $c^2 = 2.25$. Then $\mu_0 = 0.8049$ and $\mu_{1_+} = 1.1494$.*

Figure 3.1 plots the transit time via Q_R (W_R) and the transit time via Q_T (W_T) on the same axis for Example 3.3.1 with their differences ($W_T - W_R$)

side-by-side with it.

Since $\mu_0 = 0.8049$ and $\mu_{1_+} = 1.1494$, we choose $\mu_R = 1.12$ and $\mu_R = 1.16$. For $\mu_R = 1.12$ there exists a unique stable user equilibrium $p_R^\star = p_R^{EQ} = 0.4535$ and for $\mu_R = 1.16$ there exist two mixed user equilibria, $p_R^{EQ,1} = 0.5742$ and $p_R^{EQ,2} = 0.9692$. Two stable user equilibria occur at $p^{\star,1} = p_R^{EQ,1}$ and $p^{\star,2} = 1$.

Example 3.3.2 shows that this system can have a stable unique user equilibrium at the boundary.

Example 3.3.2 *Let* $N = 3$, $\mu_T = 3$, $\lambda = \lambda_T = 1$ *and* $c^2 = 2.25$. *For these parameter values,* $\mu_0 = 1.2$ *and* $\mu_{1_+} = 1.986$.

Figure 3.2 plots W_R and W_T on the same axis for Example 3.3.2 with $W_T - W_R$ side-by-side with it.

Since $\mu_0 = 1.2$ and $\mu_{1_+} = 1.986$, we choose $\mu_R = 1.1$ and a unique stable user equilibrium occurs at $p_R^\star = 0$. If $\mu_R = 2$ then an unique stable user equilibrium occurs at $p_R^\star = 1$.

The next example is for $\alpha = \mu_R + \frac{c^2-1}{2}\mu_T < 0$, where $W_T(p_R) - W_T(p_R)$ has a maximum at $\delta \in (0, 1)$ rather than a minimum.

Example 3.3.3 *Let* $N = 3$, $\mu_T = 10$, $\lambda = \lambda_T = 1$, $c^2 = 0.01$ *then* $\mu_0 = 1.6667 > \mu_{1_+} \equiv 1.6336$.

Figure 3.3 plots W_R and W_T on the same axis for Example 3.3.3 with $W_T - W_R$ side-by-side with it.

If $\mu_R = 1.64$ then there exists a mixed unique unstable user equilibrium

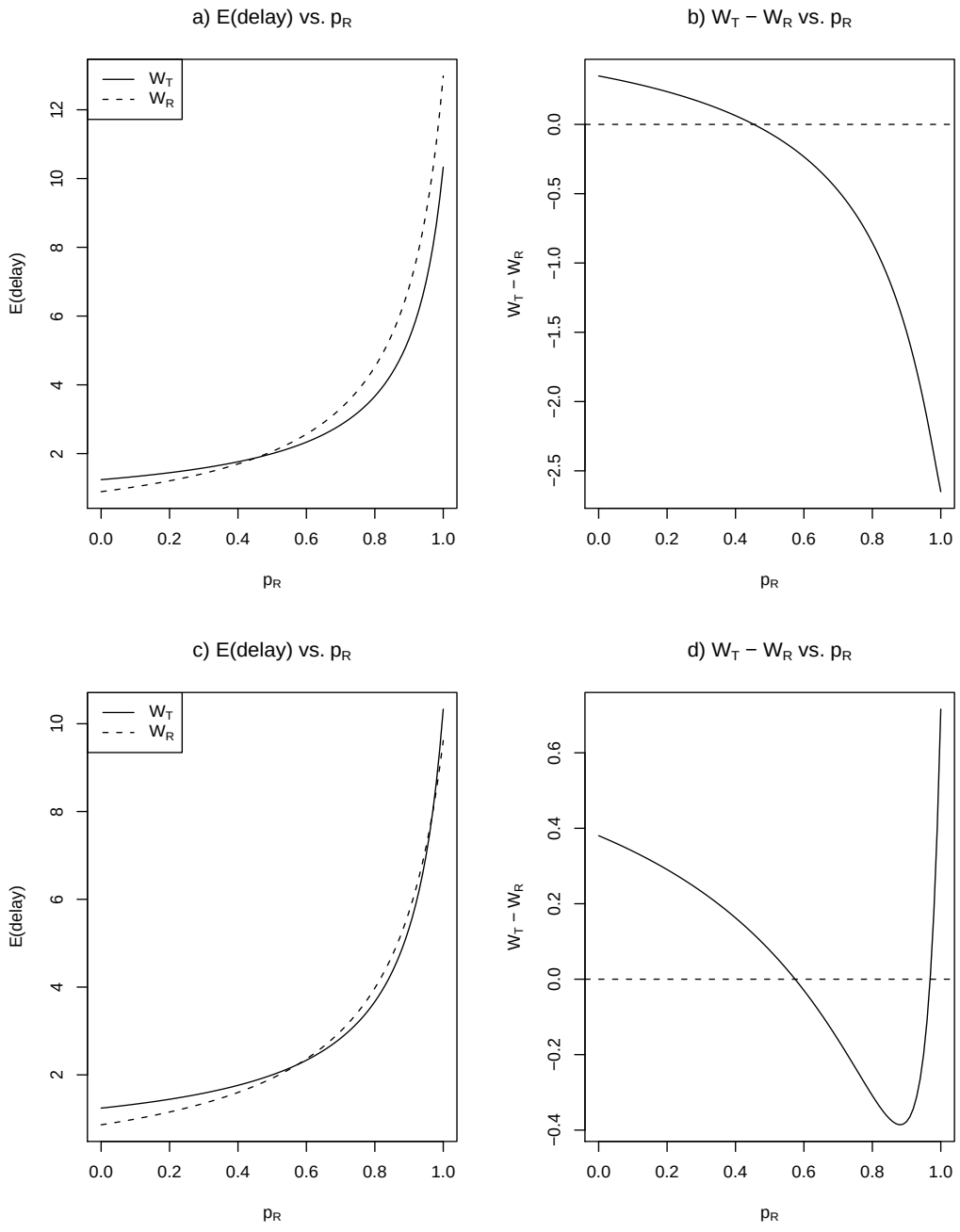


Figure 3.1: The first column contains plots of W_R and W_T on the same axis and the second column corresponds to $W_T - W_R$ when $N = 3$, $\lambda = 1, \lambda_T = 0.1, \mu_T = 3, c^2 = 2.25$, with $\mu_R = 1.12$ for plots (a) and (b), and $\mu_R = 1.16$ for plots (c) and (d). For (a) and (b), a unique stable equilibrium occurs at $p_R = 0.4535$ and for (c) and (d), stable equilibria occur at $p_R = 0.5742$ and $p_R = 1$ with an unstable equilibrium at $p_R = 0.9692$.

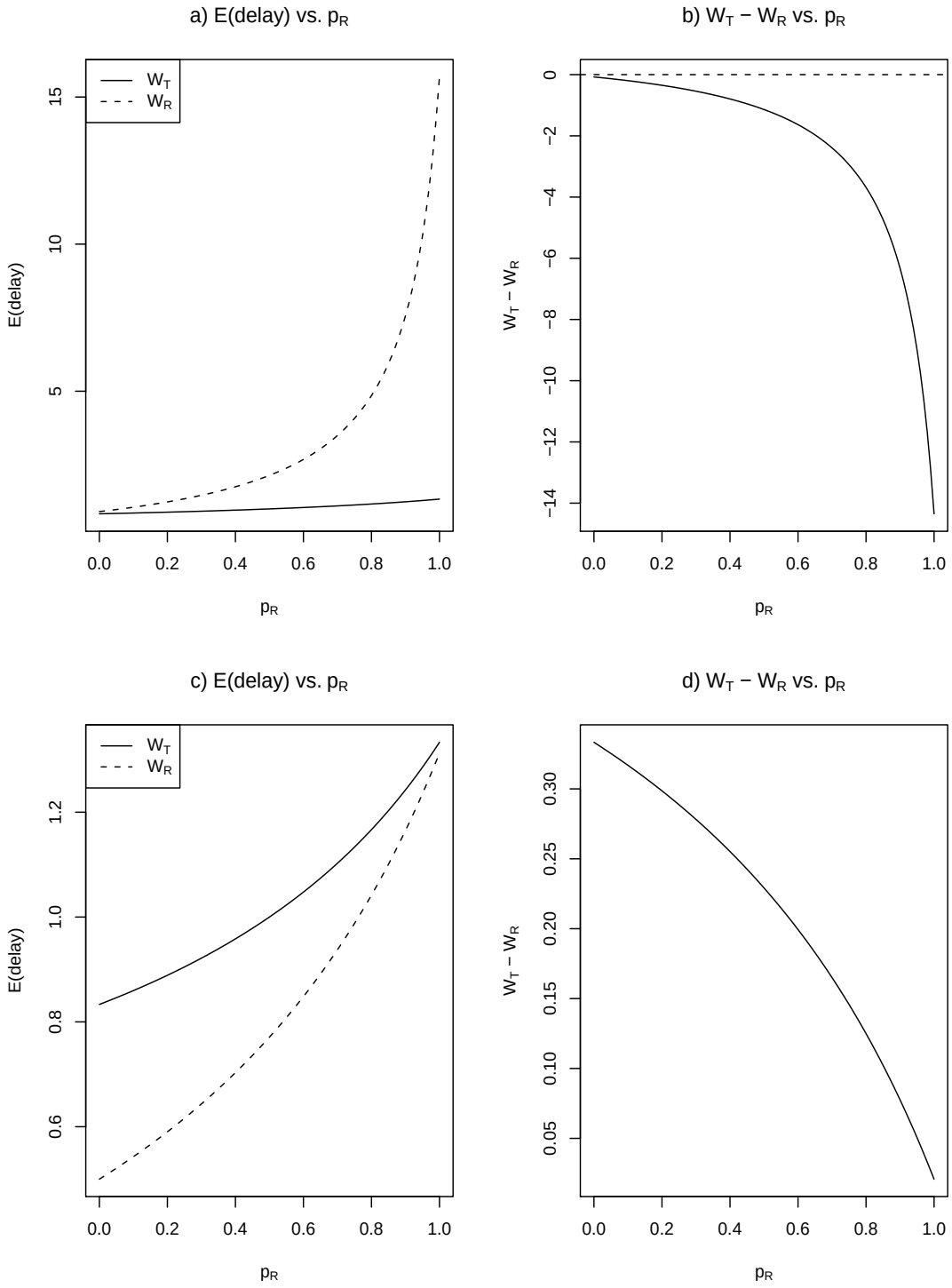


Figure 3.2: Plots of W_R and W_T on the same axis (first column) and $W_T - W_R$ (second column) when $N = 3$, $\mu_T = 3$, $\lambda = \lambda_T = 1$ and $c^2 = 2.25$ with $\mu_R = 1.1$ for plots (a) and (b) and $\mu_R = 2$ for plots (c) and (d). For plots (a) and (b), a unique stable equilibrium occurs at $p_R = 0$, and that for plots (c) and (d) occurs at $p_R = 1$.

$p_R^{EQ} = 0.1533$ with two stable user equilibria at $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$. If $\mu_R = 1.62 < \mu_{1_+}$, there exist two mixed user equilibria $p_R^{EQ,1} = 0.2982$ and $p_R^{EQ,2} = 0.9137$. Two stable user equilibria occur at $p_R^{\star,1} = 0$ and $p_R^{\star,2} = p_R^{EQ,2}$.

The last example shows that this system can have two roots also when $N > 3$.

Example 3.3.4 *Let* $\lambda = 1, \mu_R = 1.8, \mu_T = 5,\ \lambda_T = 0.1$ *and* $N = 4, 10, 40$ *and* 100 *with corresponding* $c^2 = 25, 64, 289$ *and* 729.

Figure 3.4 contains plots of $W_T - W_R$ for differing batch sizes, N, as in Example 3.3.4. For plot (a) $p^{\star,1} = p^{EQ,1} \approx 0.3217$, plot (b) $p^{\star,1} = p^{EQ,1} \approx 0.5307$, plot (c) $p^{\star,1} = p^{EQ,1} \approx 0.5380$ and plot (d) $p^{\star,1} = p^{EQ,1} \approx 0.5561$, and for all plots $p^{\star,2} = 1$.

From Examples 3.3.2 - 3.3.4, we see that this system may have a unique stable user equilibrium $p_R^\star = 0$, $p_R^\star = 1$ or $p_R^\star \in (0, 1)$ and multiple user equilibria within the interval [0, 1] for some parameter values when $N = 3$. There are three cases in which multiple equilibria occur: (A) $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$ with an unstable mixed user equilibrium $p_R^{EQ} \in (0, 1)$, (B) $p_R^{\star,1} = p_R^{EQ,1} \in (0, 1)$ and $p_R^{\star,2} = 1$ with an unstable mixed user equilibrium $p_R^{EQ,2} \in (0, 1)$ and (C) $p_R^{\star,1} = 0$ and $p_R^{\star,2} = p_R^{EQ,2} \in (0, 1)$ with an unstable mixed user equilibrium $p_R^{EQ,1} \in (0, 1)$, where $0 < p_R^{EQ,1} < p_R^{EQ,2} < 1$. From Example 3.3.4 this system also has multiple user equilibria when $N = 4$, 10, 40 and 100 and large enough coefficient of variance, c^2, while other parameter values are fixed. Note that if $c^2 = 1$ for W_R, case (B) does not exist for $N \geq 3$ and case (C) does not exist for any $N \geq 2$. We will examine the relationship of the possible user equilibria we have found for this system with other well known concepts of user equilibria

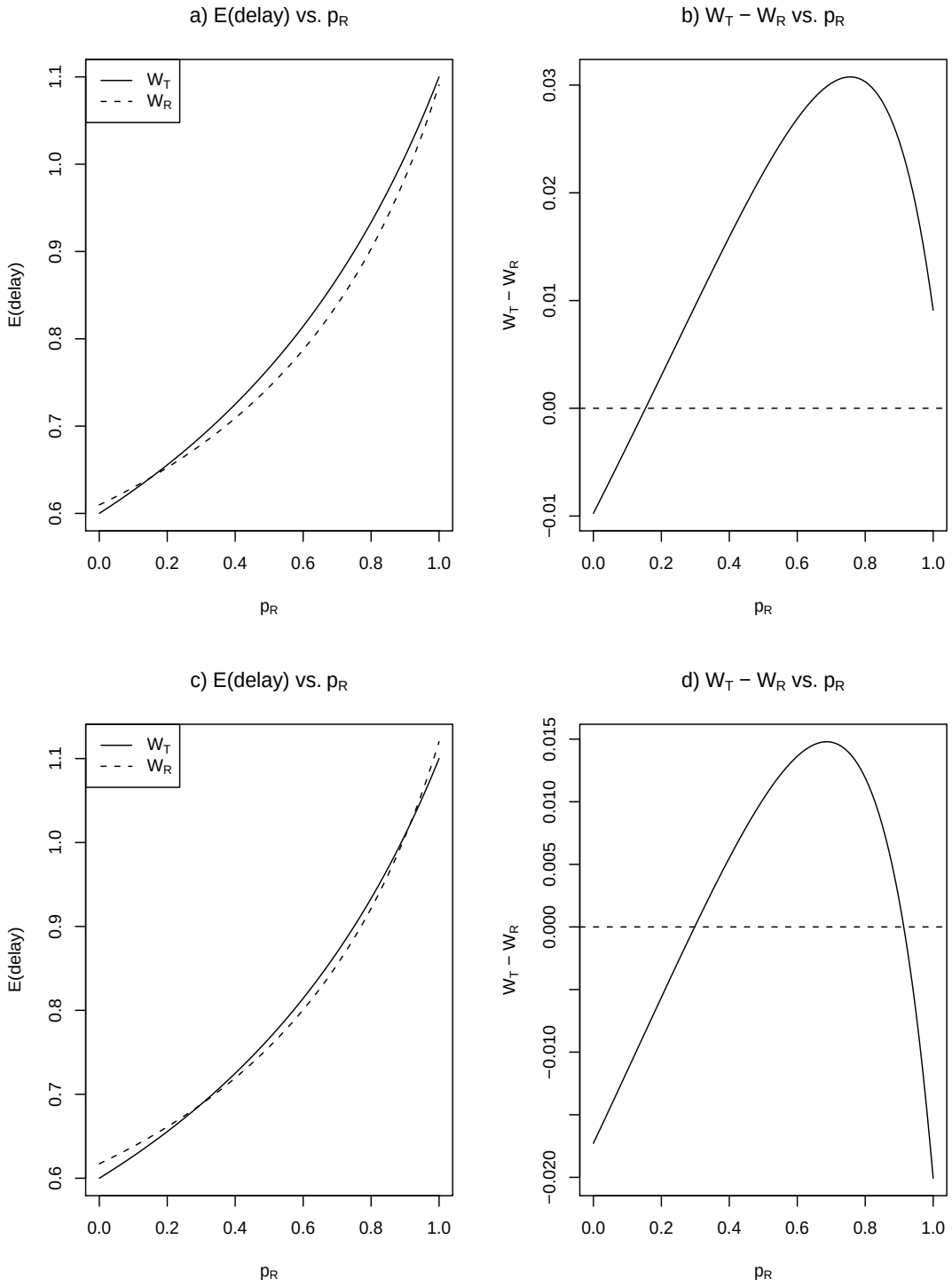


Figure 3.3: Plots of W_R and W_T on the same axis (first column) and $W_T - W_R$ (second column) when $N = 3$, $\lambda = \lambda_T = 1, \mu_T = 10$ and $c^2 = 0.01$, with $\mu_R = 1.64$ for plots (a) and (b), and $\mu_R = 1.62$ for plots (c) and (d). For (a) and (b), stable equilibria occur at $p_R = 0$ and $p_R = 1$ and $p_R = 0.1533$ is an unstable equilibrium. For (c) and (d), stable equilibria occur at $p_R = 0$ and $p_R = 0.9137$ and $p_R = 0.2982$ is an unstable equilibrium.

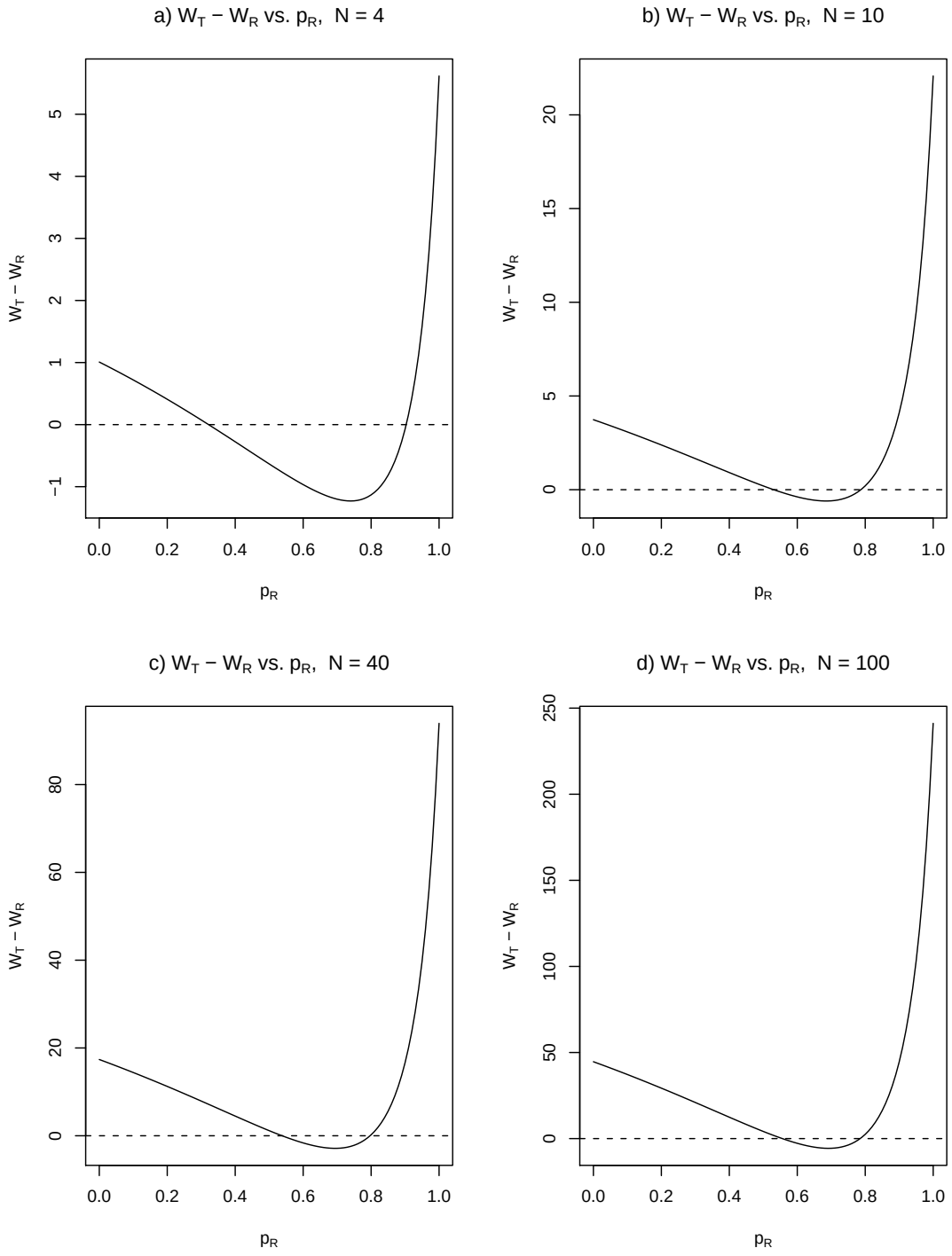


Figure 3.4: Plots of $W_T - W_R$ for $N = 4, 10, 40$ and 100 with $c = 25, 64, 289$ and 729 respectively when other parameters are fixed with $\lambda = 1, \mu_R = 1.8$, $\lambda_T = 0.1$ and $\mu_T = 5$. The smallest stable equilibria for plots (a), (b), (c) and (d) occur at $p_R = 0.3217$, $p_R = 0.5307$, $p_R = 0.5380$ and $p_R = 0.5561$, respectively. The largest stable equilibria for these plots occur at $p_R = 1$.

in the next section.

We summarize the user equilibria from the above numerical examples into 5 cases. (There may be other possible user equilibria for the system we consider here).

1. A unique stable user equilibrium occurs at one of the boundaries (i) $p_R^* = 0$ or (ii) $p_R^* = 1$ (as in Example 3.3.2, Figure 3.2, (a) and (c), respectively).

2. An unstable mixed user equilibrium, $p_R^{EQ} \in (0, 1)$, exists with two stable user equilibria $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$ (as in Example 3.3.3, Figure 3.3, (a)).

3. A unique stable mixed user equilibrium, $p_R^\star = p_R^{EQ} \in (0,\ 1)$, exists (as in Example 3.3.1, Figure 3.1, (a)).

4. Two mixed user equilibria $p_R^{EQ,1} \in (0,\ 1)$ and $p_R^{EQ,2} \in (0,\ 1)$ exist and the two stable user equilibria occur at $p_R^{\star,1} = p_R^{EQ,1}$ and $p_R^{\star,2} = 1$ (as in Example 3.3.1, Figure 3.1, (c) and Example 3.3.4, Figure 3.4, (a) - (d)).

5. Two mixed user equilibria $p_R^{EQ,1} \in (0,\ 1)$ and $p_R^{EQ,2} \in (0,\ 1)$ exist and the two stable user equilibria occur at $p_R^{\star,1} = 0$ and $p_R^{\star,2} = p_R^{EQ,2}$ (as in Example 3.3.3, Figure 3.3, (c).

Note that if $c^2 = 1$ for W_R then cases 1 to 4 exist if $N = 2$ but only cases 1 and 3 exist if $N \geq 3$. Case 5 only exists if $c^2 \neq 1$.

3.4 User equilibria, avoid the crowd or follow the crowd and evolutionary stable strategies

In this section we examine the relationship of the possible user equilibria for the system model we study here with other well known forms of equilibria such as the avoid the crowd (ATC) or follow the crowd (FTC) and the evolutionary stable strategy (ESS). Roughly speaking ATC refers to the case where an individual arriving general user acts differently from other general users and FTC refers to the case where an individual general user does what other general users do. A user equilibrium is said to be an ESS if it is the strategy used by all other general users then a marked general commuter will do better by using the same strategy. In term of Definition 3.4.4 a marked general user chooses Q_R with probability p^{EQ} if all other general users choose Q_R with probability p^{EQ}. Then p^{EQ} is an ESS.

Before proceeding, we need to extend Hassin and Haviv's definitions of ATC, FTC and ESS in terms of the system we study.

We begin by defining the strategies for our system. In equilibrium (with all parameters fixed), we consider a marked general arriving commuter who uses the road with probability p, independently of all other general arriving commuters who all, independently, travel by road with probability p_R. These probabilities p and p_R are called strategies in Hassin and Haviv's terminology (see [27]). Let

$$\mathcal{T}(p,\, p_R) = pW_R(p_R) + (1-p)\, W_T(p_R)\,, \tag{3.18}$$

be the expected transit time for the marked general commuter. The strategy p may be regarded as a *response* by our marked general commuter to the strategy p_R used by other general commuters. A *best response* by our marked general commuter is a response p that minimizes $\mathcal{T}$. Such a minimum if it is achieved may not be uniquely achieved. For example, if $p_R = p_R^\star \in (0,\, 1)$ then any choice of $p \in [0,\, 1]$ minimizes $\mathcal{T}$. We then define

$$p_1(p_R) = \left\{ \hat{p} : \mathcal{T}\left(\hat{p},\, p_R\right) = \min_{0 \le p \le 1} \mathcal{T}\left(p,\, p_R\right) \right\}$$

Note that

$$p_1(p_R) = \begin{cases} 1 & \text{if } W_R\left(p_R\right) < W_T\left(p_R\right) \\ [0,\, 1] & \text{if } W_R\left(p_R\right) = W_T\left(p_R\right) \\ 0 & \text{if } W_R\left(p_R\right) > W_T\left(p_R\right). \end{cases} \tag{3.19}$$

Definition 3.4.1 *Let f be a real set-valued function. We say that f is an increasing function if for $s < t$, $f(s) \le f(t)$, that is for all $x \in f(s)$ and $y \in f(t)$, $x \le y$. Decreasing functions are similarly defined.*

Formally, for this system, we define the ATC and FTC properties globally and locally as follows:-

Definition 3.4.2 *This particular system has the global avoid the crowd (ATC) property if p_1 is a decreasing function of $p_R \in (0, 1)$ and it has the global follow the crowd (FTC) property if p_1 is an increasing function of $p_R \in (0, 1)$.*

Definition 3.4.2 agrees with Hassin and Haviv [27] when $p_1(\cdot)$ is a singleton. However, their uniqueness requirement for $p_1(\cdot)$ must be relaxed in order for us to consider the ATC and FTC properties for our system. In the ATC

definition, we do not require $p_1(\cdot)$ to be a *strictly* decreasing function of p_R. A similar relaxation applies to the definition of FTC.

Definition 3.4.3 gives a local version of the ATC/FTC properties. The relaxation of the uniqueness requirement for $p_1(\cdot)$ is also applied for this definition.

Definition 3.4.3 *This system has the ATC property at strategy $s \in (0,1)$ if there exists $\epsilon > 0$ such that $p_1(\cdot)$ is a decreasing function of p_R in the interval $[s-\epsilon, s+\epsilon]$, and has the FTC property at $s \in (0,1)$ if $p_1(\cdot)$ is an increasing function of p_R in the interval $[s-\epsilon, s+\epsilon]$.*

Note that from Definitions 3.4.2 and 3.4.3 if the system with an user equilibrium, p^{eq}, has the property of ATC or FTC globally then the system has the property of ATC or FTC locally at p^{eq} as well.

The formal analog of the definition of an ESS in Hassin and Haviv [27] p.5 in terms of the system we study here is as follows.

Definition 3.4.4 *A user equilibrium p^{EQ} is said to be an ESS if for strategy $p \neq p^{EQ}$ either*

i. $\mathcal{T}\left(p^{EQ}, p^{EQ}\right) < \mathcal{T}\left(p, p^{EQ}\right)$

or

ii. $\mathcal{T}\left(p^{EQ}, p\right) < \mathcal{T}\left(p, p\right)$ *if* $\mathcal{T}\left(p^{EQ}, p^{EQ}\right) = \mathcal{T}\left(p, p^{EQ}\right)$

That is, for any $p \neq p^{EQ}$ which is a best response against p^{EQ}, p^{EQ} is better than p as a response to p itself.

Now we are ready to check each case of the user equilibria in the beginning of this section for whether they possess ATC or FTC properties in subsection 3.4.1. We will then check whether the possible user equilibria are evolutionary stable strategies (ESS) in subsection 3.4.2.

3.4.1 Avoid the crowd or follow it

A best response by our marked general commuter $p_1(p_R)$ to the strategy p_R is 0 if $W_R(p_R) > W_T(p_R)$, 1 if $W_R(p_R) > W_T(p_R)$ and [0, 1] if $W_R(p_R) = W_T(p_R)$ (see (3.19)). By Definition 3.4.2, this system possesses the global ATC property if $p_1(\cdot)$ is decreasing for all $p_R \in [0,\ 1]$ and it possesses the global FTC property if $p_1(\cdot)$ is increasing for all $p_R \in [0,\ 1]$. By Definition 3.4.3, X possesses the local ATC property if for $s \in (0,\ 1),\ p_1(\cdot)$ is decreasing in p_R for $[s-\epsilon,\ s+\epsilon]$ and it possesses the local FTC property if $p_1(\cdot)$ is increasing in p_R for $[s-\epsilon,\ s+\epsilon]$, where $\epsilon > 0$.

We begin with the global ATC or FTC properties of the system we study here. If this system in equilibrium has a unique user equilibrium $p_R^\star = 0$ or $p_R^\star = 1$ as in case 1 then it satisfies the condition for both the global ATC and FTC properties, since $p_1(p_R)$ is neither strictly decreasing nor strictly increasing for all $p_R \in [0,\ 1]$. If this system has $p_R^{EQ} \in (0, 1)$ with $p_R^{\star,1} = 0$, $p_R^{\star,2} = 1$ as in case 2 then it possesses the global FTC property since $p_1(\cdot)$ is non-decreasing in all $p_R \in [0, 1]$. If this system has a unique mixed user equilibrium $p_R^\star = p_R^{EQ} \in (0, 1)$ which is stable, then it possesses the ATC property, since $p_1(\cdot)$ is non-increasing for all $p_R \in [0, 1]$. This system possesses neither global ATC nor global FTC properties if it has two mixed user equilibria as in cases 4 and 5. For case 4, $p_1(\cdot) = 1$ for $p_R \in [0,\ p_R^{\star,1}]$ and $p_R \in [p_R^{EQ,2},\ 1]$, $p_1(\cdot) = 0$ for

$p_R \in [\, p_R^{\star,1},\, p_R^{EQ,2}\,]$ and $p_1(\cdot) = [0,\, 1]$ for $p_R = p_R^{\star,1}$ and $p_R = p_R^{EQ,2}$. Therefore, $p_1(\cdot)$ is decreasing for $p_R \in [0,\, p_R^{EQ.\,2}]$ and $p_1(\cdot)$ is increasing for $p_R \in [p_R^{\star,1},\, 1]$. A similar argument will show for case 5 that $p_1(\cdot)$ is increasing for $p_R \in [0,\, p_R^{\star,2}]$ and $p_1(\cdot)$ is decreasing for $p_R \in [p_R^{EQ,1},\, 1]$. See Figure 3.5 for plots of $p_1(\cdot)$ versus p_R for cases 1 - 5 of user equilibria. We summarize this argument in Lemma 3.4.1.

Lemma 3.4.1 *Suppose X is a system with fixed parameter $\Gamma = (\lambda, \lambda_T, \mu_R, \mu_T, c)$ and $N \geq 2$. Then, in equilibrium, X possesses*

i. Both the global ATC and FTC properties if there exists a unique pure stable user equilibrium $p_R^\star = 0$ or $p_R^\star = 1$.

ii. The global ATC property if there exists a unique stable mixed user equilibrium $p_R^\star \in (0,\, 1)$.

iii. The global FTC property if there exists a single mixed user equilibrium. That is, $p_R^{EQ} \in (0, 1)$ and the two pure stable user equilibria occur at $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$.

Now we determine whether this system model possesses local FTC or ATC properties. In the following discussion we will consider the mixed user equilibria only as in cases 2, 3, 4 and 5. By similar arguments to those for the global property if the system has a unique mixed user equilibrium, as in cases 2 and 3, then it possesses the local FTC property at the unstable mixed equilibrium, p_R^{EQ}, and the local ATC property at the stable mixed equilibrium, $p_R^\star$. It remains only to consider cases 4 and 5, where two mixed user equilibria exist. Neither of these last cases have the global FTC or ATC properties.

We consider first the unstable mixed user equilibrium $p^{eq} \in (0,\, 1)$. Note that if this process has $p^{eq} \in (0,\, 1)$ then $W_R > W_T$ for $p < p^{eq}$ such that $p \in [p^{eq} - \epsilon,\, p^{eq})$ and $W_R < W_T$ for $p > p^{eq}$ such that $p \in (p^{eq},\, p^{eq} + \epsilon]$, for $\epsilon > 0$. Then $p_1(p) = 0$ for $p \in [p^{eq} - \epsilon,\, p^{eq})$, $p_1(p) = [0, 1]$ for $p = p^{eq}$ and $p_1(p) = 1$ for $p \in (p^{eq},\, p^{eq} + \epsilon]$. Therefore, by Definition 3.4.3 this particular system model possesses the local FTC property at the unstable mixed user equilibrium. Now we consider the stable mixed user equilibrium, $p^\star \in (0, 1)$, and we use a similar argument to the one above for this case. If $p^\star \in (0, 1)$ exists for our system model then $W_R < W_T$ for $p < p^\star$ and $W_R > W_T$ for $p > p^\star$. Then $p_1(p) = 1$ for $p \in [p^{eq} - \epsilon,\, p^{eq})$, $p_1(p) = [0, 1]$ for $p = p^{eq}$ and $p_1(p) = 0$ for $p \in (p^{eq},\, p^{eq} + \epsilon]$. Hence, this particular system model possesses the local ATC property by Definition 3.4.3 at the stable mixed user equilibrium. We summarize this argument in the following lemma:-

Lemma 3.4.2 *Let X be a system with fixed parameter $\Gamma = (\lambda, \lambda_T,\, \mu_R, \mu_T, c)$ and $N \geq 2$. Then in equilibrium X possesses the local FTC property at the unstable mixed user equilibrium, $p^{eq} \in (0, 1)$, and the local ATC property at the stable mixed equilibrium, $p^\star \in (0, 1)$.*

3.4.2 Evolutionary Stable Strategy (ESS)

We examine each user equilibrium case from the beginning of section 3.3 according to Definition 3.4.4, that is, the strategy p^{EQ} is an evolutionary stable strategy (ESS) if for all $p \neq p^{EQ}$ such that $p \in [0, 1]$ either $\mathcal{T}(p,\, p^{EQ}) > \mathcal{T}(p^{EQ},\, p^{EQ})$ or if $\mathcal{T}(p,\, p^{EQ}) = \mathcal{T}(p^{EQ},\, p^{EQ})$ then $\mathcal{T}(p,\, p) > \mathcal{T}(p^{EQ},\, p)$, where $\mathcal{T}(p,\, p^{EQ}) = p\, W_R(p^{EQ}) + (1 - p)\, W_T(p^{EQ})$ is the expected transit time

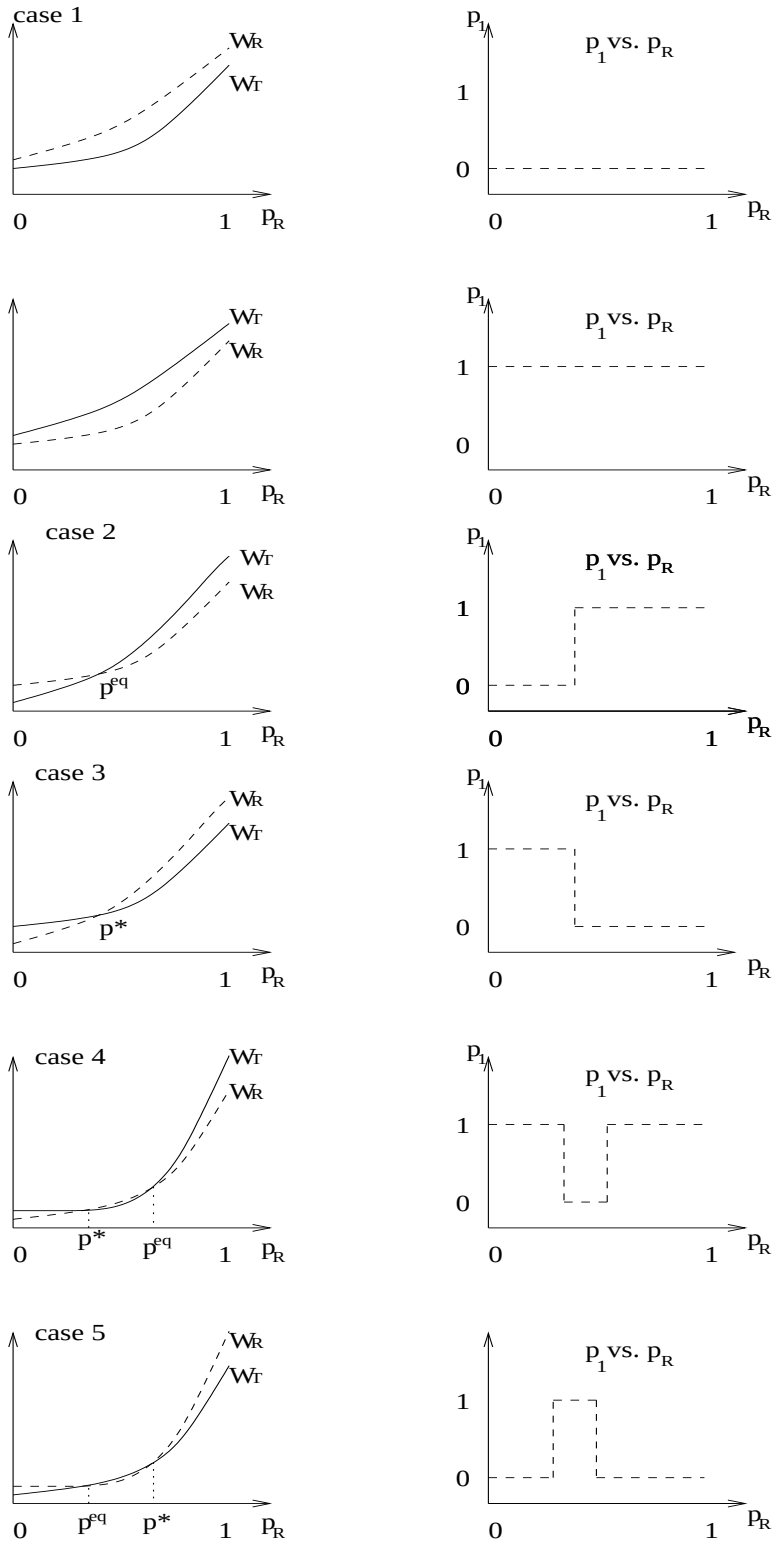


Figure 3.5: The plots in the first column plot W_T (- - -) and W_R (——-) on the same axis and those in the second column plot the best response (p_1) versus p_R for each user equilibrium case. Case 1 possesses both ATC and FTC properties, Cases 2 and 3 possess the FTC and ATC properties, respectively. Cases 4 and 5 have the local ATC property at $p^\star \in (0,1)$ and the local FTC property at $p^{EQ} \in (0,1)$.

for a marked general commuter who chooses Q_R with probability p while the other general commuters choose Q_R with probability p^{EQ} [see (3.18)].

Consider first the stable user equilibria, $p_R^\star = 0$ or $p_R^\star = 1$ as in cases 1 and 2 of the possible user equilibria in section 3.3. For $p^\star = 0$,

$$\begin{aligned} \mathcal{T}(p,0) - \mathcal{T}(0,0) &= pW_R(0) + (1-p)W_T(0) - W_T(0) \\ &= p\left[W_R(0) - W_T(0)\right] \end{aligned} \tag{3.20}$$

Since $W_R(0) > W_T(0)$ then (3.20) is greater than 0 for any $p \neq 0$, therefore $p^\star = 0$ is an ESS. Similarly, $T(p,1) - T(1,1) = (p-1)\left[W_R(1) - W_T(1)\right] > 0$ for any $p \neq 1$ since $W_R(1) < W_T(1)$. Hence, $p_R^\star = 1$ is an ESS.

The second case we consider is an unstable mixed user equilibrium $p^{EQ} \in (0,1)$ as in case 2 of the possible user equilibria in section 3.3. For this case, $\mathcal{T}\left(p,\, p^{EQ}\right) - T\left(p^{EQ},\, p^{EQ}\right) = 0$ since $W_R(p^{EQ}) - W_T(p^{EQ}) = 0$ for all $p \neq p^{EQ}$ as shown next.

$$\begin{aligned} \mathcal{T}\left(p, p^{EQ}\right) - \mathcal{T}\left(p^{EQ}, p^{EQ}\right) &= pW_R(p^{EQ}) + (1-p)W_T(p^{EQ}) \\ &\quad - \left[p^{EQ}\, W_R(p^{EQ}) + (1-p^{EQ})\, W_T(p^{EQ})\right] \\ &= (p - p^{EQ})\left[W_R(p^{EQ}) - W_T(p^{EQ})\right] \\ &= 0 \end{aligned}$$

Then by Definition 3.4.4 *ii* we need to show that $\mathcal{T}\left(p^{EQ},\, p\right) < \mathcal{T}\left(p,\, p\right)$ for all $p \neq p^{EQ}$. That is, if $\mathcal{T}\left(p,p\right) - \mathcal{T}\left(p^{EQ},p\right) = \left(p - p^{EQ}\right)\left[W_{R}\left(p\right) - W_{T}\left(p\right)\right] > 0$ for all $p \neq p^{EQ}$ then p^{EQ} is an ESS. However, if $p \in [0,\, p^{EQ})$ then $(p - p^{EQ})[W_{R}(p) - W_{T}(p)] < 0$ since $W_{R}\left(p\right) > W_{T}\left(p\right)$. Similarly, $\left(p - p^{EQ}\right)\left[W_{R}\left(p\right) - W_{T}\left(p\right)\right] < 0$ for $p \in (p^{EQ},\, 1]$ since $W_{R}\left(p\right) > W_{T}\left(p\right)$. Thus, p^{EQ} is not an ESS.

The third case we consider is a unique mixed user equilibrium that is stable, $p^{\star} \in (0,1)$, as in case 3 of the possible user equilibria in section 3.3. Similar to the second case above we need to show that $\mathcal{T}\left(p,\, p\right) - \mathcal{T}\left(p_{R}^{\star},\, p\right) > 0$ for all $p \neq p_{R}^{\star}$ since $\mathcal{T}\left(p,\, p_{R}^{\star}\right) - T\left(p_{R}^{\star},\, p_{R}^{\star}\right) = 0$. First, suppose $p \in [0,\, p_{R}^{\star})$ then $(p - p_{R}^{\star})[W_{R}(p) - W_{T}(p)] > 0$ since $W_{R}\left(p\right) < W_{T}\left(p\right)$ for all $p \in [0,\, p_{R}^{\star})$. Now, suppose $p \in (p_{R}^{\star},\, 1]$ then $\left(p - p_{R}^{\star}\right)\left[W_{R}\left(p\right) - W_{T}\left(p\right)\right] > 0$ since $W_{R}\left(p\right) > W_{T}\left(p\right)$ for all $p \in (p_{R}^{\star},\, 1]$. Thus, $p_{R}^{\star}$ is an ESS.

The last case we consider is when two mixed user equilibria exist with one of them being stable ($p^{\star,m}$) and the other unstable ($p^{EQ,m}$), where m is either 1 or 2. Another stable pure user equilibrium occurs either at $p^{\star,m} = 0$ as in case 5 or $p^{\star,m} = 1$ as in case 4. For $p^{\star,1} = 0$ as in case 5, $\mathcal{T}\left(p,0\right) - \mathcal{T}\left(0,0\right) = p\left[W_{R}(0) - W_{T}(0)\right] > 0$ for all $p \in (0,\, 1]$ since $W_{R}(0) > W_{T}(0)$. Similarly, $\mathcal{T}\left(p,1\right) - \mathcal{T}\left(1,1\right) = \left(p - 1\right)\left[W_{R}(1) - W_{T}(1)\right] > 0$ for all $p \in (0,1)$, since $W_{R}(1) < W_{T}(1)$. Therefore, both pure user equilibria, $p^{\star,m} = 0$ and $p^{\star,m} = 1$, are ESS. For an unstable or stable mixed equilibrium, (i.e. $p^{EQ,m} \in (0,1)$ and $p^{\star,m} \in (0,1)$, respectively) since $\mathcal{T}\left(p,\, p^{EQ,m}\right) - \mathcal{T}\left(p^{EQ,m},\, p^{EQ,m}\right) = 0$, and $\mathcal{T}\left(p,\, p^{\star,m}\right) - \mathcal{T}\left(p^{\star,m},\, p^{\star,m}\right) = 0$ we need to show that for all $p \neq p^{EQ,m}$, $\mathcal{T}\left(p,\, p\right) - \mathcal{T}\left(p^{EQ,m},\, p\right) = \left(p - p^{EQ,m}\right)\left[W_{R}\left(p\right) - W_{T}\left(p\right)\right] > 0$ for $p^{EQ,m}$ to be an ESS, and for all $p \neq p^{\star,m}$ $\mathcal{T}\left(p,\, p\right) - \mathcal{T}\left(p^{\star,m},\, p\right) = \left(p - p^{\star,m}\right)\left[W_{R}\left(p\right) - W_{T}\left(p\right)\right] >$

0 for $p^{\star,m}$ to be an ESS. Since $W_R > W_T$ if $p < p^{EQ,m}$ and $W_R < W_T$ if $p > p^{EQ,m}$ both the unstable mixed equilibrium, $p^{EQ,m} \in (0,1)$, and the stable mixed equilibrium,$p^{\star,m} \in (0,1)$, are not ESS.

The following lemma summarizes the above arguments:-

Lemma 3.4.3 *Let X be a system with parameter $\Gamma = (\lambda,\, \mu_R, c^2, \lambda_T,\, \mu_T)$ and $N \geq 2$. Suppose X has a stable user equilibrium, $p^\star$. Then $p^\star$ is an evolutionary stable strategy (ESS) if it is a pure user equilibrium (i.e. $p^\star = 0$ or $p^\star = 1$) or a unique mixed user equilibrium (i.e. $p^\star \in (0,1)$).*

3.4.3 Social optimum

The social optimum minimizes the expected transit time of all commuters, while at the user equilibrium each commuter minimizes his own expected transit time. In this subsection we prove that the value of p_R at which the social optimum occurs is less than or equal to the value of p_R at which the stable user equilibrium for the system we study.

The social optimum, $\hat{p}$, is the value of p that minimizes the expected steady state transit time for all commuters, $W(p)$, on the interval [0, 1], where

$$W(p) = pW_R(p) + (1-p)W_T(p) \tag{3.21}$$

$$= p\left(\frac{1}{\mu_R} + \frac{(1+c^2)\,p}{2\mu_R(\mu_R - p)}\right) + (1-p)\left(\frac{1}{\mu_T} + \frac{N-1}{\lambda_T + 1 - p}\right), \quad \lambda = 1$$

$$= \left(-\beta_1\, p^3 + \beta_2\, p^2 - \beta_3\, p + \beta_4\right) / \left((\mu_R - p)(1 - p + \lambda_T)\mu_T\mu_R\right)$$

$$\begin{aligned}
\text{where} \qquad \beta_1 &= \mu_R - \frac{1-c^2}{2}\mu_T \\
\beta_2 &= \mu_R^2 + (2 + \lambda_T + \frac{N-3}{2}\mu_T)\mu_R - (1+\lambda_T)(\frac{1-c^2}{2}\mu_T) \\
\beta_3 &= (2 + \lambda_T + \frac{N-1}{2}\mu_T)\mu_R^2 + (1 + \lambda_T + \frac{N-3}{2}\mu_T - \lambda_T\mu_T)\mu_R \\
\beta_4 &= (1 + \lambda_T + \frac{N-1}{2}\mu_T)\mu_R^2
\end{aligned}$$

Then

$$\begin{aligned}
\frac{dW}{dp} &= \Big(-\beta_1 p^4 + 2\beta_1(1+\lambda_T+\mu_R)p^3 \\
&\quad -\big[3(1+\lambda_T)\beta_1\mu_R + (1+\lambda_T+\mu_R)\beta_2 - \beta_3\big]p^2 \\
&\quad +2\big[(1+\lambda_T)\mu_R\beta_2 - \beta_4\big]p + (1+\lambda_T+\mu_R)\beta_4 \\
&\quad -(1+\lambda_T)\mu_R\beta_3\Big) \,/\, \beta_5
\end{aligned} \tag{3.22}$$

where $\beta_5 = \big[(\mu_R - p)(1+\lambda_T - p)\big]^2\mu_T\mu_R$.

For the $\cdot|M|1$ and $\cdot|M^{(N)}|\infty$ queues (i.e. $c^2 = 1$), if $N \geq 3$ then the social optimum is always less than or equal to the unique stable user equilibrium (i.e. $\hat{p} \leq p_R^\star$) with the inequality being strict if $p_R^\star \in (0,1)$ (see Theorem 3.5 in [2]). Note for this system model that the user equilibrium is always unique and stable if $N \geq 3$. However, if $N = 2$ then multiple equilibria may exist. From [2] also if a single unstable mixed user equilibrium $p_R^{EQ} \in (0,1)$ exists with $p_R^{\star,1} = 0$ and $p_R^{\star,2} = 1$ then $\hat{p} = p_R^{\star,1} = 0$. The remaining case to analyze is when two mixed user equilibria exist, one stable and one unstable. This case also exists if $c^2 \neq 1$, where $p_R^{\star,1} = p_R^{EQ,1} \in (0,1)$ and $p_R^{\star,2} = 1$ From the numerical examples in section 3.3 there is also another possible case with $p_R^{\star,1} = 0$ and $p_R^{\star,2} = p_R^{EQ,2} \in (0,1)$. If $p_R^{\star,1} = 0$ then $\hat{p} = 0$ also, since $W(0) = W_T(0) = 1/\mu_T + (N-1)/(2\,\lambda_T + \lambda)$. We are left with the case in

which $p_R^{\star,1} = p_R^{EQ,1} \in (0,1)$ to check and we claim that $\hat{p} < p_R^{\star,1}$. The idea for the following result is due to an anonymous referee.

Consider any p' such that $0 \leq p' < p_R^{\star,1}$. Then, since $W_R(p_R)$ and $W_T(p_R)$ are both increasing in p_R,

$$\begin{aligned} W(p_R^{\star,1}) &= p_R^{\star,1} W_R(p_R^{\star,1}) + (1 - p_R^{\star,1}) W_T(p_R^{\star,1}) && (3.23) \\ &\geq p_R^{\star,1} W_R(p_R^{\star,1}) + (1 - p_R^{\star,1}) W_T(p') && (3.24) \\ &\geq p' W_R(p_R^{\star,1}) + (1 - p') W_T(p'), \\ &\qquad \text{since } W_R(p_R^{\star,1}) > W_T(p') && (3.25) \\ &\geq p' W_R(p') + (1 - p') W_T(p') = W(p'), && (3.26) \end{aligned}$$

where (3.25) follows from (3.24) since

$$\begin{aligned} & p_R^{\star,1} W_R(p_R^{\star,1}) + (1 - p_R^{\star,1}) W_T(p') \geq p' W_R(p_R^{\star,1}) + (1 - p') W_T(p') \\ \Leftrightarrow \quad & (p_R^{\star,1} - p') W_R(p_R^{\star,1}) \geq (p_R^{\star,1} - p') W_T(p') \\ \Leftrightarrow \quad & W_R(p_R^{\star,1}) \geq W_T(p'), \end{aligned}$$

which holds since $W_R(p_R^{\star,1}) = W_T(p_R^{\star,1})$. Similarly, consider any p' such that $p_R^{\star,1} < p' \leq 1$. Then, since W_R and W_T are both increasing in p_R,

$$\begin{aligned} W(p_R^{\star,1}) &= p_R^{\star,1} W_R(p_R^{\star,1}) + (1 - p_R^{\star,1}) W_T(p_R^{\star,1}) && (3.27) \\ &\leq p_R^{\star,1} W_R(p') + (1 - p_R^{\star,1}) W_T(p_R^{\star,1}) && (3.28) \\ &\leq p' W_R(p') + (1 - p') W_T(p_R^{\star,1}), \\ &\qquad \text{since } W_R(p_R^{\star,1}) < W_T(p') && (3.29) \\ &\leq p' W_R(p') + (1 - p') W_T(p') = W(p'). && (3.30) \end{aligned}$$

Now (3.29) follows from (3.28) since

$$\begin{aligned} & p_R^{\star,1}\, W_R(p') + (1 - p_R^{\star,1})\, W_T(p_R^{\star,1}) \;\leq\; p'\, W_R(p') + (1 - p')\, W_T(p_R^{\star,1}) \\ \Leftrightarrow\;\; & (p' - p_R^{\star,1})\, W_T(p_R^{\star,1}) \leq (p' - p_R^{\star,1})\, W_R(p') \\ \Leftrightarrow\;\; & W_T(p_R^{\star,1}) \;\leq\; W_R(p') \end{aligned}$$

again, since $W_R(p_R^{\star,1}) = W_T(p_R^{\star,1})$.

For all $p' < p_R^{\star,1}$, $W(p') < W(p_R^{\star,1})$ and $W(p') > W(p_R^{\star,1})$, for all $p' > p_R^{\star,1}$, therefore, $\hat{p} < p_R^{\star,1}$ as we claim.

We summarise the relationship of the social optimum and the user equilibrium for the system we study here in Lemma 3.4.4.

Lemma 3.4.4 *Let X be a system with parameter $\Gamma = (\lambda, \mu_R, c^2, \lambda_T, \mu_T)$ and $N \geq 2$. If X has an unique stable user equilibrium, $p^\star$, then $\hat{p} \leq p^\star$, with the inequality being strict if $p^\star \in (0, 1)$. If X has multiple user equilibria, with the smallest stable user equilibrium occurring at $p^{\star,1}$, then $\hat{p} \leq p^{\star,1}$.*

We use (3.21) and (3.22) to obtain the social optima for the following numerical examples. The first example revisits one of the examples discussed in [2], where $N = 2$, $\lambda = 1, \mu_R = 1.3, \lambda_T = 0.1$ and $\mu_T = 3$. For this example the coefficient of variance, c^2 is equal to 1. The two stable user equilibria occur at $p_R^{\star,1} = p_R^{\mathrm{EQ},1} = 0.1$ and $p_R^{\star,2} = 1$ with the unstable equilibrium occurring at $p_R^{\mathrm{EQ},2} = 0.8$. The social optimum occurs at $\hat{p} = 0$. The second example revisits Example 3.3.3 where $N = 3$, $\lambda = 1, \mu_R = 1.62, \lambda_T = 1, \mu_T = 10$ and $c^2 = 0.01$. Then $p_R^{\star,1} = 0$, and it is also socially optimum for this case. The second stable user equilibrium, $p_R^{\star,2} = p_R^{\mathrm{EQ},2} = 0.9137$, and $p_R^{\mathrm{EQ},1} = 0.2982$

is an unstable mixed equilibrium. The remaining examples revisit Example 3.3.4 where $\lambda = 1$, $\mu_R = 1.8$, $\lambda_T = 0.1$ and $\mu_T = 5$ for all examples but N and c are varied as in i - iv. For all of these examples the second stable user equilibrium, $p_R^{\star,2} = 1$, and the other user equilibria and the social optimum for each numerical example are as follows.

i. If $N = 4$ and $c^2 = 25$ then the smaller stable user equilibrium $p_R^{\star,1} = p^{EQ,1} = 0.3217$ and the social optimum $\hat{p} = 0$ with the unstable equilibrium, $p^{EQ,2} = 0.9035$,

ii. If $N = 10$ and $c^2 = 64$ then $p_R^{\star,1} = 0.5307$ and $\hat{p} = 0.0008$ with $p_R^{EQ,2} = 0.7874$,

iii. If $N = 40$ and $c^2 = 289$ then $p_R^{\star,1} = 0.5380$ and $\hat{p} = 0.0143$ with $p_R^{EQ,2} = 0.7971$,

iv. If $N = 100$ and $c^2 = 729$ then $p_R^{\star,1} = 0.5561$ and $\hat{p} = 0.0169$ with $p_R^{EQ,2} = 0.7853$.

3.5 Discussion and Conclusion

From the numerical examples in section 3.3 multiple user equilibria can exist for $N \geq 2$ if the network we study here consists of $\cdot|G|1$ and $\cdot|G^{(N)}|\infty$ queues. This behaviour does not exist for $N \geq 3$ if the single server service queue is restricted to be an $\cdot|M|1$ queue (see [2]).

If the system we study in this chapter has a pure user equilibrium that is unique and stable (i.e. $p^\star = 0$ or $p^\star = 1$) then by definition it possesses both

the global ATC property and the global FTC property. Furthermore, these pure user equilibria are ESS. If this system has a mixed user equilibrium that is unique and stable, $p^\star \in (0, 1)$, then it possesses the global ATC property and the local ATC property also. In addition, $p^\star$ is ESS. This system possesses the global FTC property and the local FTC property also if it has a mixed user equilibrium which is unstable, $p^{EQ} \in (0, 1)$, with the two stable user equilibria occurring at 0 and 1, i.e. $p_R^{\star,1} = 0$ and $p^{\star,2} = 1$. Moreover, p^{EQ} is not ESS while $p_R^{\star,1} = 0$ and $p^{\star,2} = 1$ are both ESS. If this system has multiple user equilibria with two mixed user equilibria then it possesses neither the global ATC property nor the global FTC property but it possesses the local ATC property at its stable mixed user equilibrium and the local FTC property at its unstable mixed user equilibrium. Note that we consider the mixed user equilibria only for the local ATC and local FTC properties. None of the last two equilibria is ESS.

We showed that the value of p at which the social optimum occurs is less than or equal to the smallest value of p at which a stable user equilibrium occurs.

Note that all the numerical examples for this chapter were produced using the R code in the Appendix A.2. Similar to those examples in Chapter 2 the code was checked using the examples in [2, 15].

Chapter 4

State-dependent routing for $\cdot|M|1$ versus $\cdot|M^{(N)}|1$ queues

4.1 Introduction

We consider the same network with two parallel queues that we analyzed in the last two chapters but in this chapter the batch service queue, Q_T, has one server instead of an infinite number of servers, that is Q_T is modelled as a $\cdot|M^{(N)}|1$ queue, with batch size N, while Q_R is a $\cdot|M|1$ queue, as in Chapter 2. We are interested in analyzing the properties of the user optimal policy for this system, analogous to those we studied in Chapter 2 for the system with a $\cdot|G^{(N)}|\infty$ queue. As we mentioned in Chapter 1 it is more realistic to consider a finite number of servers for Q_T but due to the lack of a compact expression for the expected delay via Q_T for an arriving general user given the current state of the system we choose to work with a single server. However, a similar

analysis can be performed if Q_T has multiple servers (finite) while Q_R is still modelled as a $\cdot|M|1$ queue.

As we described in the previous chapters, both queues model possible routes from the same source to the same destination. The service times at Q_R and Q_T are both exponentially distributed with rates μ_R and μ_T respectively. There are two Poisson arrival processes into the system. The first, at rate λ_T, is the arrival process of committed train users who will always join Q_T. The second, at rate λ, is the stream of general users who can join Q_R or Q_T. The system has state-dependent routing since arriving general users base their choice of route on the current state of the system including the number of customers in each queue as well as the arrival rates to the network and the service rates at each queue. We also assume that whenever a general user joins a queue they cannot switch queues. All inter-arrival times and service times are independent of one another. Note that in this chapter we need the assumption that service times at Q_T are exponentially distributed, and can no longer assume a general distribution.

Similarly to Chapter 2 we will show for this system that a user optimal policy exists, is unique and possesses various monotonicity properties. Note that we will not analyse the social (system) optimal policy for this model.

The outline of this chapter is as follows. Section 4.2 gives the definitions of the relevant variables and some preliminary results. In section 4.3 we show that the expected transit times, conditional on finding the system in a given state and taking Q_T, have a structural property that we have named slice-monotonicity. In section 4.4 we show that a unique user optimal policy exists and is slice monotone. Section 4.5 concludes with a discussion of the results

obtained in this chapter.

4.2 Definitions and preliminaries

In Chapters 2 and 3, since we only needed to calculate the expected transit times for a marked commuter via both routes, we were able to consider a reduced description of the state of the system. In particular, for Q_T, since we assumed an infinite number of servers, we could just consider the expected time until the marked commuter's batch was completed, since then service would begin immediately, and the expected service time is known. Here that is not the case, since there is only a single server, so the marked commuter's batch may also have to wait for service once the batch is complete. However the basic approach remains the same. We observe the process until the marked commuter's batch is complete as before, and then observe the number of batches still remaining in Q_T. If there are i batches in the queue ahead of the marked commuter's batch (including the batch in service) at the time the marked commuter's batch is completely full, then the expected time until the marked commuter's batch completes service is $(i+1)/\mu_T$.

Let $\tilde{X}_1(t) \in \mathbb{Z}^+$ and $\tilde{X}_2(t) \in \mathbb{Z}^+$ denote the number of commuters in Q_R and Q_T respectively, including any commuters in service. Given $\tilde{X}_2(t) = n$, the number of commuters in Q_T, we can also deduce the number of complete batches in the queue, i, and the number of commuters, l, in the incomplete batch that a marked general user joins, that is, n can be written as $n = iN + l$, where $i \in \mathbb{Z}^+$ and $l \in \{0, 1, \ldots, N-1\}$. The state space for the process $\{\tilde{X}(t) = (\tilde{X}_1(t), \tilde{X}_2(t)), t \geq 0\}$ for this system is given by $\Omega = \{\, (a, n) \in$

$\mathbb{Z}^+ \times \mathbb{Z}^+\}$. The definition of a decision policy is similar to that given earlier in Chapter 2.

Definition 4.2.1 *A decision policy $D = (R, T)$ for the process $\tilde{X}$ is a partition of Ω into R and T. If an arriving general user finds the system in state $\boldsymbol{x} \in R$ or $\boldsymbol{x} \in T$ then they join Q_R or Q_T respectively. Let $\mathcal{D}$ be the class of all decision policies for this system. We will write $\tilde{X}_D$ to denote the process $\tilde{X}$ operating under decision policy $D \in \mathcal{D}$.*

The process $\tilde{X}_D = \{\tilde{X}_D(t),\, t \geq 0\}$ is a Markov process with transition rates as follows. Let $e_1 = (1, 0)$ and $e_2 = (0, 1)$. For all $a \geq 0$ and $n \geq 0$

$$\boldsymbol{x} \longrightarrow \begin{cases} \boldsymbol{x} + e_1 & \text{at rate } \lambda I_{\{\boldsymbol{x} \in R\}} \\ \boldsymbol{x} + e_2 & \text{at rate } \lambda_T + \lambda I_{\{\boldsymbol{x} \in T\}} \\ \boldsymbol{x} - e_1 & \text{at rate } \mu_R \text{ if } a \geq 1 \\ \boldsymbol{x} - N\, e_2 & \text{at rate } \mu_T \text{ if } n \geq N \end{cases} \tag{4.1}$$

where $I_{\{\boldsymbol{x} \in R\}}$ and $I_{\{\boldsymbol{x} \in T\}}$ are indicator functions as defined in (2.1).

The expression for the expected transit times for a general user who arrives to find the system in state $\boldsymbol{x} \in \Omega$ are similar to those in Chapter 2 and are as follows.

Since, given the state $\boldsymbol{x} = (a,\, iN + l)$, the expected transit time via Q_R, $y(\boldsymbol{x})$, is independent of the number of servers available in Q_T and the decision policy $D \in \mathcal{D}$, its expression is the same as before. That is, if the system is in state $\boldsymbol{x}$ and an arrival from the stream of general users joins Q_R then that

arrival reaches its destination after $a+1$ services are completed at Q_R and

$$y(\boldsymbol{x}) = \frac{a+1}{\mu_R}.$$

The expressions for the expected transit time via Q_T, $z_D(\boldsymbol{x})$, are as follows.

If $l = N-1$ then the arrival of the next general user to Q_T completes the batch containing that user and if there are i batches ahead, then

$$z_D(a,\, iN+N-1) = \frac{i}{\mu_T} + \frac{1}{\mu_T} = \frac{i+1}{\mu_T}, \qquad \text{for } a \geq 0 \tag{4.2}$$

If $l < N-1$ then the equations for $z_D(\boldsymbol{x})$ are found by conditioning on the first transition after the general user has joined Q_T, where $\boldsymbol{x} = (a,\, n)$ with $n = iN + l$ and $l \in \{0, 1, \ldots, N-2\}$. Let $\eta = \lambda + \lambda_T$ and $\Lambda = \eta + \mu_R$. We express the equations for $z_D(\boldsymbol{x})$ when $i = 0$ (that is, the server is free) and when $i > 0$ (that is, the server is busy) separately.

The expressions for the $z_D(\boldsymbol{x})$ when $n < N-1$ that is, $l = 0, 1, 2, \ldots, N-2$ with $i = 0$ are very similar to (2.4) and (2.5) in Chapter 2 as follows.
If $\boldsymbol{x} + e_2 \in T$ then

$$z_D(\boldsymbol{x}) = \begin{cases} \frac{1}{\Lambda}\Big(1 + \eta\, z_D(\boldsymbol{x}+e_2) + \mu_R\, z_D(\boldsymbol{x}-e_1)\Big), & \text{if } a \geq 1 \\ \\ \frac{1}{\eta} + z_D(\boldsymbol{x}+e_2), & \text{if } a = 0 \end{cases} \tag{4.3}$$

If $\boldsymbol{x} + e_2 \in R$ then

$$z_{D}(\boldsymbol{x}) = \begin{cases} \frac{1}{\Lambda}\Big(1 + \lambda_{T}\, z_{D}(\boldsymbol{x} + e_2) + \lambda\, z_{D}(\boldsymbol{x} + e_1) \\ \qquad + \mu_{R}\, z_{D}(\boldsymbol{x} - e_1)\Big), & \text{if } a \geq 1 \\ \\ \frac{1}{\eta}\big(1 + \lambda_{T}\, z_{D}(\boldsymbol{x} + e_2) + \lambda\, z_{D}(\boldsymbol{x} + e_1)\big), & \text{if } a = 0 \end{cases} \tag{4.4}$$

When $n \geq N$, that is, $i > 0$ and $l = 0, 1, 2, \ldots, N-2$, the expressions for the $z_{D}(\boldsymbol{x})$ are as follows.

If $\boldsymbol{x} + e_2 \in T$ then

$$z_{D}(\boldsymbol{x}) = \begin{cases} \frac{1}{\Lambda+\mu_{T}}\big(1 + \eta\, z_{D}(\boldsymbol{x} + e_2) + \mu_{R}\, z_{D}(\boldsymbol{x} - e_1) \\ \qquad + \mu_{T}\, z_{D}(\boldsymbol{x} - Ne_2)\big), & \text{if } a \geq 1 \\ \\ \frac{1}{\eta+\mu_{T}}\big(1 + \eta\, z_{D}(\boldsymbol{x} + e_2) + \mu_{T}\, z_{D}(\boldsymbol{x} - Ne_2)\big), & \text{if } a = 0 \end{cases} \tag{4.5}$$

If $\boldsymbol{x} + e_2 \in R$ then

$$z_{D}(\boldsymbol{x}) = \begin{cases} \frac{1}{\Lambda+\mu_{T}}\big(1 + \lambda_{T}\, z_{D}(\boldsymbol{x} + e_2) + \lambda\, z_{D}(\boldsymbol{x} + e_1) \\ \qquad + \mu_{R}\, z_{D}(\boldsymbol{x} - e_1) + \mu_{T}\, z_{D}(\boldsymbol{x} - Ne_2)\big), & \text{if } a \geq 1 \\ \frac{1}{\eta+\mu_{T}}\big(1 + \lambda_{T}\, z_{D}(\boldsymbol{x} + e_2) + \lambda\, z_{D}(\boldsymbol{x} + e_1) \\ \qquad + \mu_{T}\, z_{D}(\boldsymbol{x} - Ne_2)\big), & \text{if } a = 0 \end{cases} \tag{4.6}$$

We shall refer to the set of states $\Omega_k = \{(a, n) : a \geq 0,\, n \in \{kN, kN + 1, kN + 2, \ldots, (k+1)N - 1\}\}$ as the k^{th} slice, for $k \in \mathbb{Z}^{+}$. The following definitions are analogous to those in Chapter 2. The first definition relates to user optimality.

Definition 4.2.2

1. *A decision policy $D = (R, T) \in \mathcal{D}$ is a k^{th} slice ξ-level user optimal policy for some ξ such that $1 \leq \xi \leq N$ and $k \in \mathbb{Z}^+$ if*

$$\boldsymbol{x} \in R \Leftrightarrow y(\boldsymbol{x}) < z_D(\boldsymbol{x}) \tag{4.7}$$

for all $\boldsymbol{x} \in \Omega_i$, $0 \leq i < k$ and for all $\boldsymbol{x} = (a, kN + l) \in \Omega_k$ such that $N - \xi \leq l \leq N - 1$ and $a \geq 0$. Let $\mathcal{D}^{\star}_{k\,\xi}$ denote the class of all k^{th} slice $\xi-$level user optimal policies.

2. *We say D is a k^{th} slice user optimal policy if it is a k^{th} slice $N-$level user optimal policy. Let $\mathcal{D}^{\star}_k = \mathcal{D}^{\star}_{kN}$ be the class of all k^{th} slice user optimal policies.*

3. *We say D is a user optimal policy for this system if it is k^{th} slice user optimal for all $k \geq 0$. Let $\mathcal{D}^{\star}$ be the class of all user optimal policies.*

As in Chapter 2, we will show that the user optimal policy possesses various monotonicity properties. Here, however, the monotonicity properties are not quite as straightforward. The following definition of slice monotonicity is the definition needed for this system.

Definition 4.2.3

1. *A policy $D = (R, T) \in \mathcal{D}$ is k^{th} slice $\xi-$level monotone for some ξ such that $1 \leq \xi \leq N$ and $k \in \mathbb{Z}^+$ if both of the following properties hold:-*

   (A) $\boldsymbol{x} \in T \Rightarrow \boldsymbol{x} + e_1 \in T$, *for all* $\boldsymbol{x} \in \Omega_i$ *such that* $0 \leq i < k$ *and* $\boldsymbol{x} = (a, kN + l) \in \Omega_k$ *such that* $N - \xi \leq l \leq N - 1$ *and* $a \geq 0$.

(B) $\boldsymbol{x} \in T \Rightarrow \boldsymbol{x} + e_2 \in T$, *for all* $\boldsymbol{x} = (a, iN + l) \in \Omega_i$, $0 \leq i < k$ *such that* $a \geq 0$ *and* $0 \leq l \leq N - 2$ *and all* $\boldsymbol{x} = (a, kN + l) \in \Omega_k$ *such that* $a \geq 0$ *and* $N - \xi \leq l \leq N - 2$.

2. *If* D *is a* k^{th} *slice* $N-$*level monotone policy we say* D *is* k^{th} *slice monotone. Let* $\mathcal{D}_{k,mon}$ *be the class of all* k^{th} *slice monotone policies.*

3. *If* D *is* k^{th} *slice monotone for all* $k \geq 0$, *we say* D *is slice monotone. Let* $\mathcal{D}_{mon}$ *be the class of all slice monotone policies.*

By using arguments similar to those for the system with the $\cdot|G^{(N)}|\infty$ queue, we can show that for any $D^\star \in \mathcal{D}^\star$ the set of states $R \cap \Omega_k$ is finite for any k such that $k \geq 0$.

We need to define switching curves as we did in Chapter 2, but here we define switching curves for each level. We let $a_i^\star = \{a_i^\star(n);\, iN \leq n \leq (i + 1)N - 1\}$ be the switching curve for level i and define it in a similar fashion to Definition 2.2.5.

Definition 4.2.4 *For* $D = (R, T) \in \mathcal{D}$, $i \geq 0$ *and* n *such that* $iN \leq n \leq (i + 1)N - 1$, *let* $a_i^\star(n) = \inf\{a : (a, n) \in T\}$. *If* $\inf\{a : (a, n) \in T\} = \phi$ *then* $a_i^\star(n) = +\infty$.

Figure 4.1 gives an example of part of a decision policy D for a system with batch size $N = 3$, which is 0^{th} and 1^{st} slice monotone. For this example the switching curves are defined by $a_0^\star = (5, 3, 2)$ and $a_1^\star = (7, 5, 4)$.

As we mentioned at the beginning of this section, for the coupling arguments that follow, we will consider a process which stops once the marked

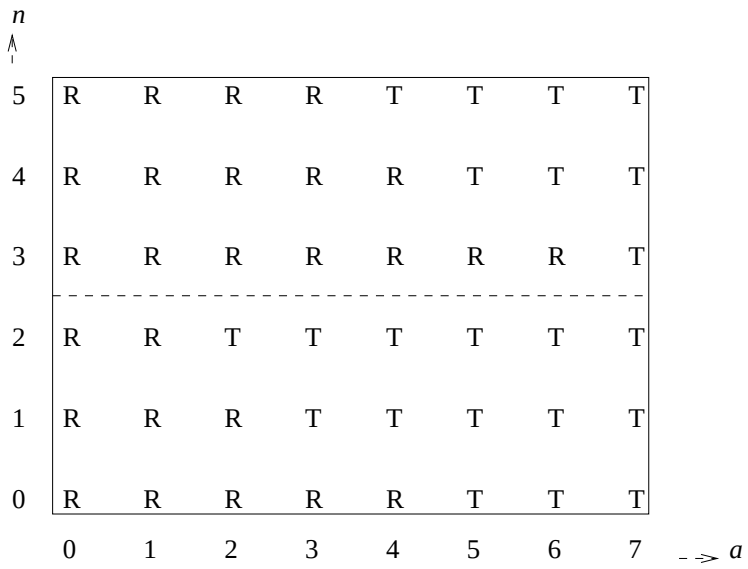


Figure 4.1: Example of part of a 0^{th} and 1^{st} slice monotone policy for a system with a one server batch service queue.

commuter's batch is completed. Let $X = (X_1, X_2)$ denote this associated process, where X_1 and X_2 are the number of commuters seen by an arriving marked general commuter at Q_R and Q_T respectively. We may write X_D to denote the process X operating under a decision policy D, although where it is clear, we may also omit the subscript D. We also need the following variables for the stochastic comparison and coupling arguments. First, let

$$\tau_D(\boldsymbol{x}, H; s) = \inf_{t \geq 0} \{t : \ X_D(t+s) \in H | X_D(s) = \boldsymbol{x}\}$$

be the reaching time to the set of states H after time s given that the process X_D is in state $\boldsymbol{x}$ at time s, for $\boldsymbol{x} \in \Omega$ and $H \subset \Omega$. Let $m_D(\boldsymbol{x}, H, s) = \mathrm{E}\big(\tau_D(\boldsymbol{x}, H; s)\big)$. We will be considering the special case with $H = H_N$ where $H_N = \{\boldsymbol{x} = (a, n) \in \Omega : \ n = kN \text{ for some } k \in \mathbb{Z}^+\}$ is the set of states with no incomplete batches. Also let $\varpi_D(\boldsymbol{x}\,; s)$ denote the time spent in the system until their own batch is complete by a marked general commuter who joins Q_T at time s when the system is in state $\boldsymbol{x}$ (and operating under decision policy

D). Then

$$\begin{aligned}\varpi_D(\boldsymbol{x}\,;s) &= \inf_{t\geq 0}\{t:\, X(t+s)\in H_N | X(s-)=\boldsymbol{x},\, X(s)=\boldsymbol{x}+e_2\}\\ &\sim \tau_D\big(\boldsymbol{x}+e_2, H_N; s\big)\end{aligned}$$

by the Markov property, where $X(s^-) = \lim_{t\Rightarrow s} X(t)$. By time-homogeneity, the distributions of $\tau_D\big(\boldsymbol{x}+e_2, H_N; s\big)$ and $\varpi_D(\boldsymbol{x}\,;s)$ do not depend on s and so we will omit this and write $\tau_D\big(\boldsymbol{x}+e_2, H_N\big)$, $m_D(\boldsymbol{x}+e_2, H_N)$ and $\varpi_D(\boldsymbol{x})$. Since we are primarily interested in the first reaching times to the set H_N, we will also make a further simplification and write $\tau_{D_k}\big(\boldsymbol{x}+e_2\big)$ and $m_{D_k}(\boldsymbol{x}+e_2)$ for $\tau_{D_k}\big(\boldsymbol{x}+e_2, H_N\big)$ and $m_{D_k}(\boldsymbol{x}+e_2, H_N)$.

The total expected time spent in the system by a marked general commuter who sees the system in state $\boldsymbol{x}$ and joins Q_T including their service time, $z_D(\boldsymbol{x})$, is then the expected time to complete their own batch plus the expected time waiting for the service completions of any batches ahead of their own batch plus the expected service time of their own batch and so

$$z_D(\boldsymbol{x}) = E\big(\varpi_D(\boldsymbol{x})\big) + C + 1/\mu_T, \tag{4.8}$$

where C denotes the expected time until service is completed on any batches ahead of the marked commuter's own batch. For this system if $\boldsymbol{x} = (a, N-1)$, that is the server is free and the batch lacks only one commuter to be full when the marked commuter arrives, then $C = 0$ and service on the marked commuter's batch starts immediately after he/she arrives, so that $z_D(a, N-1) = 1/\mu_T$. Furthermore, if $\boldsymbol{x} = (a, n)$ is such that $n = kN + N - 1$ for some $k \geq 1$ then $z_D(\boldsymbol{x}) = (k+1)/\mu_T$ since $E\big(\varpi_D(\boldsymbol{x})\big) = 0$ and $C = k/N$.

4.3 Existence, uniqueness and slice monotonicity of the user optimal policy

Since we will be comparing $y(\boldsymbol{x})$ and $z_D(\boldsymbol{x})$ in order to establish the properties of the user optimal policy for this system we begin with a structural property for $z_D(\boldsymbol{x})$ in the following lemma.

Lemma 4.3.1 *Let the process X be operating under a k^{th} slice ξ-level monotone policy $D_{k\xi} \in \mathcal{D}_{k\xi}$ for some $k \geq 0$ and ξ such that $1 \leq \xi \leq N$. Let $\boldsymbol{x}_j = (a_j,\, n_j)$ with $n_j = k_j N + l_j$ for some $k_j \geq 0$ and $0 \leq l_j \leq N-1$, $j = 1, 2$. Then $z^{(1)}_{D_{k\xi}}(\boldsymbol{x}_1) \leq z^{(2)}_{D_{k\xi}}(\boldsymbol{x}_2)$ for any pair of states $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega_0 \cup \Omega_1 \cup \cdots \cup \Omega_k$ satisfying $k_1 = k_2 = k$, $l_2 \geq N - \xi - 1$ and any of the following conditions:-*

I $a_1 = a_2,\ l_1 > l_2,$

II $a_1 < a_2, l_1 > l_2;\ a_2 - a_1 \leq l_1 - l_2,$

III $a_1 > a_2,\ l_1 = l_2,$

IV $a_1 > a_2,\ l_1 > l_2.$

Proof. For the proof of this lemma, similarly to the proofs in Chapter 2, we consider a joint process, $(\boldsymbol{X}^{(1)},\, \boldsymbol{X}^{(2)})$, such that both $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ obey the law of $X_{D_{k\xi}}$, where $D_{k\xi} \in \mathcal{D}_{k\xi}$ is a policy that is k^{th} slice ξ-level monotone. We use a proof by forwards induction on the transitions of the process and observe the joint process until the completion time in the process $\boldsymbol{X}^{(1)}$ of the batch that the marked general commuter has joined. As in Chapter 2, we will match the transitions made by both processes in such a way that both move together whenever possible. In particular, the service of a batch at Q_T in $\boldsymbol{X}^{(1)}$ will always be accompanied by the service of a batch at Q_T in $\boldsymbol{X}^{(2)}$. Then since the service of a batch in Q_T can only take place if the batch is complete,

we will show below that $\varpi_{D_{k\xi}}(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \varpi_{D_{k\xi}}(\boldsymbol{x}_2)$. If $\varpi_{D_{k\xi}}(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \varpi_{D_{k\xi}}(\boldsymbol{x}_2)$ then $z_{D_{k\xi}}(\boldsymbol{x}_1) \leq z_{D_{k\xi}}(\boldsymbol{x}_2)$, since the batch containing the marked customer in $\boldsymbol{X}^{(2)}$ has either completed simultaneously with completion of the marked customer's batch in $\boldsymbol{X}^{(1)}$, or it has yet to complete. In either case, the batches ahead of the marked customer in both systems are served simultaneously, until the batch containing the marked customer in $\boldsymbol{X}^{(1)}$ commences service. Since there is positive probability that the batch containing the marked general commuter at Q_T in $\boldsymbol{X}^{(2)}$ is not yet complete, the result follows.

We show $\varpi_{D_{k\xi}}(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \varpi_{D_{k\xi}}(\boldsymbol{x}_2)$ by using a coupling argument. We couple the two processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ in such a way that they make the same transition together whenever possible until $\boldsymbol{X}^{(1)}$ hits the set H_N. Define processes $\{A(t); t \geq 0\}$, $\{L^{(j)}(t); t \geq 0\}$ and $\{K^{(j)}(t); t \geq 0\}$ such that $\mathbf{X}^{(j)}(t) = (A^{(j)}(t), K^{(j)}(t)N + L^{(j)}(t))$. Then the inductive hypothesis is that for a k^{th} slice ξ-level monotone policy $D_{k\xi}$, if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(t) = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies one of the conditions I - IV (with $l_2 \geq N - \xi - 1$) then at the next transition, occurring, say, at time t', $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(t')$ also satisfies one of the conditions I-IV with probability 1. It then follows that if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(0) = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies one of the conditions I - IV, then $L^{(1)}(t) \geq L^{(2)}(t)$ with probability 1 until $\tau^{(1)}_{D_{k\xi}}(\boldsymbol{x}_1 + e_2)$ and therefore, $\varpi^{(1)}_{D_{k\xi}}(\boldsymbol{x}_1) \leq \varpi^{(2)}_{D_{k\xi}}(\boldsymbol{x}_2)$ w.p.1, which then implies $\varpi_{D_{k\xi}}(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \varpi_{D_{k\xi}}(\boldsymbol{x}_2)$.

We observe first that if $\boldsymbol{x}_1, \boldsymbol{x}_2$ satisfy one of the conditions I - IV, with $l_1 = N - 1$, then $\varpi^{(1)}_{D_k}(\boldsymbol{x}_1) = 0$ with probability 1, and we are done. So consider now the cases where $0 \leq l_1 \leq N - 2$.

Consider first the case when $k = 0$, that is, consider a policy that is 0^{th} slice ξ-level monotone with $\boldsymbol{x}_1$, $\boldsymbol{x}_2 \in \Omega_0$. For this case, $n_1, n_2 \in \{0, 1, \ldots, N-1\}$

and service can start immediately as soon this batch completes since the server is free. Then the waiting time for service for an arriving marked general commuter who sees the system in state (a, n) and joins Q_T is just the time until their batch is complete, which is the same as the time spent waiting for service by a marked general user who arrives and sees the system in (a, n) and joins Q_T for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues (see Chapter 2). Therefore, that proof implies that here we have $\tau^{(1)}_{D0\xi}(\boldsymbol{x}_1) \leq \tau^{(2)}_{D0\xi}(\boldsymbol{x}_2)$ w.p.1 for $\boldsymbol{x}_1, \boldsymbol{x}_2$ satisfying one of the conditions $I - IV$ and such that $n_2 \geq N - \xi - 1$, that is, $\varpi^{(1)}_{D0\xi}(\boldsymbol{x}_1) \leq \varpi^{(2)}_{D0\xi}(\boldsymbol{x}_2)$ w.p.1 for $\boldsymbol{x}_1, \boldsymbol{x}_2$ such that $n_2 \geq N - \xi - 1$ and therefore, $\varpi_{D0\xi}(\boldsymbol{x}_1) \overset{\text{st}}{\leq} \varpi_{D0\xi}(\boldsymbol{x}_2)$ for $\boldsymbol{x}_1, \boldsymbol{x}_2$ such that $n_2 \geq N - \xi - 1$. Hence, $z_{D0\xi}(\boldsymbol{x}_1) \leq z_{D0\xi}(\boldsymbol{x}_2)$ for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega_0$ satisfying any one of the conditions I - IV and such that $n_2 \geq N - \xi - 1$.

Now consider the case where $k > 0$ with $D \in \mathcal{D}_{k\xi}$. Define the joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ with transition rates as follows:

$$(\boldsymbol{x}_1, \boldsymbol{x}_2) \rightarrow \begin{cases} (\boldsymbol{x}_1 - e_1 I_{\{a_1>0\}}, \boldsymbol{x}_2 - e_1 I_{\{a_2>0\}}) \\ \qquad \text{at rate } \mu_R \\ (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_2) \\ \qquad \text{at rate } \lambda_T + \lambda I_{\{\boldsymbol{x}_1 \in T, \boldsymbol{x}_2 \in T\}} \\ (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_1) \\ \qquad \text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in R, \boldsymbol{x}_2 \in R\}} \\ (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_1) \\ \qquad \text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in T, \boldsymbol{x}_2 \in R\}} \\ (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_2) \\ \qquad \text{at rate } \lambda I_{\{\boldsymbol{x}_1 \in R, \boldsymbol{x}_2 \in T\}} \\ \big((\boldsymbol{x}_1 - Ne_2) I_{\{n_1 \geq N\}}, (\boldsymbol{x}_2 - Ne_2) I_{\{n_2 \geq N\}} \big) \\ \qquad \text{at rate } \mu_T \end{cases} \tag{4.9}$$

Since we will be stopping the process as soon as the marked customer's batch fills, and we will show below that both $L^{(1)}(t) \geq L^{(2)}(t)$ and $K^{(1)}(t) \leq K^{(2)}(t)$ up until that time, we have not specified transitions that might be needed beyond that point.

We need to show that if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(0) = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ for any pair of states $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfying one of the conditions I - IV then $L^{(1)}(t) \geq L^{(2)}(t)$ and $K^{(1)}(t) \leq K^{(2)}(t)$ w.p.1 until $\boldsymbol{X}^{(1)}$ hits H_N, that is, the batch that a marked general commuter joins is complete. As in Chapter 2 we do so by showing that any state that the joint process visits up to that time will either satisfy one of the conditions $I - IV$ or the two processes are in the same state and have coupled. Let $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ denote the state of the system immediately after a

transition out of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. We will treat each transition separately and show that if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies any of the conditions I - IV then either $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ satisfies one of the conditions $I - IV$ or $\boldsymbol{x}_1' = \boldsymbol{x}_2'$ and the processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ have coupled and remain coupled thereafter.

First we discuss the process when a service completion occurs at Q_R. Services at Q_R occur at rate μ_R provided that either $a_1 > 0$ or $a_2 > 0$ or both $a_1 > 0$ and $a_2 > 0$. If both $a_1 > 0$ and $a_2 > 0$ then $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - e_1, \boldsymbol{x}_2 - e_1)$ and so $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies whichever condition $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied. Additionally, the following transitions can occur. Under condition I, if $a_1 = a_2 = 0$ then no service completion in Q_R can occur. Under condition II, $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1, \boldsymbol{x}_2 - e_1)$ if $a_2 > a_1 = 0$ and so $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ either satisfies condition I if $a_2 = 1$ or condition II is preserved if $a_2 > 1$. Under conditions III and IV, $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - e_1, \boldsymbol{x}_2)$, if $a_1 > a_2 = 0$. Then the sub-processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ couple, if $a_1 = 1$ and $n_1 = n_2$ or $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition I, if $a_1 = 1$ and $n_1 > n_2$. Under both conditions III and IV $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ retains the condition of $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$, if $a_1 > 1$.

Now consider the transition when an arrival from the committed train users' stream occurs at Q_T. These transitions occur at rate λ_T. If $l_2 \leq l_1 < N - 1$, then $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_2)$ and so $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied. Under conditions I, II and IV (where $l_1 > l_2$) if $l_1' = N$ then the process stops, since the batch is completed in $\boldsymbol{X}^{(1)}$ and awaits service. The process also stops if $l_1' = l_2' = N$, which is possible under condition III. That is, both the batches in processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are completed simultaneously and await service (if $k_1 = 0$ or $k_1 = k_2 = 0$ then service will commence immediately).

Now consider the transition when an arrival from the general users' stream occurs. In this case the transition depends on the decisions of the states $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We consider first the cases where both processes follow the same decision. If both $\boldsymbol{x}_1 \in T$ and $\boldsymbol{x}_2 \in T$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_2)$ which we have already discussed above when the transition was due to arrivals from the committed train users' stream. If both $\boldsymbol{x}_1 \in R$ and $\boldsymbol{x}_2 \in R$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_1)$, and $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ preserves whichever condition $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied. Now consider the cases where the two processes follow different decisions. If $\boldsymbol{x}_1 \in T$ and $\boldsymbol{x}_2 \in R$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_1)$. If $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied conditions I or II then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies condition II. If $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied conditions III or IV then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies condition I if $a_1 = a_2 + 1$, or condition IV if $a_1 > a_2 + 1$. Finally we consider the case where $\boldsymbol{x}_1 \in R$ and $\boldsymbol{x}_2 \in T$. Since D is k^{th} slice ξ-level monotone we need consider this case for condition II only and $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_2)$. Then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the same condition if $a_2 > a_1 + 1$ and $l_1 > l_2 + 1$, the condition I if $a_2 = a_1 + 1$ and $l_1 > l_2 + 1$, or the two processes $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ couple and remain coupled thereafter if $a_2 = a_1 + 1$ and $l_1 = l_2 + 1$ (i.e. $\boldsymbol{x}_1' = \boldsymbol{x}_2'$).

Finally we consider the transitions when a service in Q_T occurs. These services occur at rate μ_T provided that both $n_1 > N$ and $n_2 > N$ with $l_2 \leq l_1$ and $l_1 > 0$ (if $n_1 = N$ or $l_1 = 0$ then the joint process is stopped). If both $n_1 > N$ and $n_2 > N$ then $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \rightarrow \left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - Ne_2, \boldsymbol{x}_2 - Ne_2)$ and so the condition satisfied by $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ continues to hold for $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$.

Since all possible transitions lead to a $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ that either satisfies one of the conditions I - IV or has $\boldsymbol{x}_1' = \boldsymbol{x}_2'$, $\tau_{D_{k\xi}}^{(1)}(\boldsymbol{x}_1) \leq \tau_{D_{k\xi}}^{(2)}(\boldsymbol{x}_2)$ w.p.1, which

then implies $\varpi^{(1)}_{D_{k\xi}}(\boldsymbol{x}_1 - e_2) \leq \varpi^{(2)}_{D_{k\xi}}(\boldsymbol{x}_2 - e_2)$ w.p.1. Therefore, $\varpi_{D_{k\xi}}(\boldsymbol{x}_1 - e_2) \overset{\text{st}}{\leq} \varpi_{D_{k\xi}}(\boldsymbol{x}_2 - e_2)$. As we have argued above, this is sufficient to ensure $z_{D_{k\xi}}(\boldsymbol{x}_1 - e_2) \leq z_{D_{k\xi}}(\boldsymbol{x}_2 - e_2)$ and we are done. ■

Corollary 4.3.1 *Suppose a process X_D is operating under a slice monotone policy $D \in \mathcal{D}$. Then for $\boldsymbol{x} \in \Omega$ such that $n > N$,*

$$z_D(\boldsymbol{x}) \geq z_D(\boldsymbol{x} - Ne_2), \quad \textit{where } e_2 = (0,1). \tag{4.10}$$

Proof: By Lemma 4.3.1 $z_D(\boldsymbol{x})$ is monotone for each slice $i \in \{0, 1, \ldots, k\}$. For a state $\big(a, (i+1)N-1\big)$, $a \geq 0$, the inequality (4.10) holds since $z_D(a, (i+1)N-1) = (i+1)/\mu_T$. Then a very similar argument to the one given above shows the result ■

Theorem 4.3.1 *Consider a process X with parameters $\Gamma = (\lambda, \mu_R, \lambda_T, \mu_T)$ and $N \geq 2$. Then there exists a unique user optimal policy $D^\star \in \mathcal{D}^\star$, for this system. Furthermore, $D^\star$ is slice-monotone, that is, the switching curve, $a_i^\star$, for slice i satisfies $a_i^\star(n+1) \leq a_i^\star(n)$, $iN \leq n \leq (i+1)N - 2$, for all $i \geq 0$.*

Proof. Let $n = kN + l$ where $k \geq 0$ and $0 \leq l \leq N-1$. We use a proof by induction on $\xi = N - l$ such that $\xi \in \{1, 2, \ldots, N\}$ and $k \geq 0$. The inductive hypothesis for ξ and k is that the class of k^{th} slice $\xi-$level user optimal policies, $D^\star_{k\,\xi}$, is non-empty and any policy $D \in \mathcal{D}^\star_{k\xi}$ is k^{th} slice ξ-level monotone. Furthermore, if $D_1 = (R_1, T_1)$, $D_2 = (R_2, T_2)$ are two k^{th} slice ξ-level user optimal policies then $\boldsymbol{x} \in R_1 \Leftrightarrow \boldsymbol{x} \in R_2$ for all $\boldsymbol{x} \in \Omega_0 \cup \Omega_1 \cup \cdots \cup \Omega_{k-1}$ and $\boldsymbol{x} \in \Omega_k$ such that $l \geq N - \xi$. If this is the case then we say that the k^{th} slice ξ-level user optimal policies are identical up to level ξ in slice k.

We consider first the case where $k = 0$, that is $n \in \{0, 1, \ldots, N-1\}$. The equations for the $z_{D_0}(\boldsymbol{x})$ for this case as in (4.3) and (4.4) are the same as equations (2.3) - (2.5) in Chapter 2. In addition the equations for the $y(\boldsymbol{x})$ are the same for both systems and so the proof for this case follows in a similar fashion to the proof of Theorem 2.3.2 in Chapter 2.

Consider first $\boldsymbol{x} \in \Omega_0$ such that $n = N-1$, that is, $\xi = 1$. For all $a \geq 0$, $z(a, N-1) = 1/\mu_T$ and $y(a, N-1) = (a+1)/\mu_R$ is strictly increasing in a. Then $(a+1)/\mu_R \geq 1/\mu_T$ if and only if $a > \frac{\mu_R}{\mu_T} - 1$ and so $a_0^\star(N-1) = \lceil \frac{\mu_R}{\mu_T} - 1 \rceil$. Hence, the inductive hypothesis holds for $\xi = 1$.

Suppose the inductive hypothesis holds for $\xi = m$. We will show that it holds for $\xi = m+1$, that is, $n = N-m-1$. By the inductive hypothesis there exists at least one policy $D \in \mathcal{D}_{0m}^\star$ which is 0^{th} slice $m-$level monotone and unique up to level m in slice 0. Now, let $\beta_0 = \min\{a : z(a, N-m-1) \leq y(a, N-m-1)\}$. By Lemma 4.3.1 condition *III* $z(a, N-m-1)$ is decreasing in a, while $y(a, N-m-1)$ is strictly increasing in a. Therefore, $z(a, N-m-1) \leq y(a, N-m-1)$ if and only if $a \geq \beta_0$, that is, the user optimal decision in state $(a, N-m-1)$ is to choose Q_T if and only if $a \geq \beta_0$. By Lemma 4.3.1 condition *I* $z(a, N-m) \leq z(a, N-m-1)$ and therefore, $a_0^\star(N-m) \leq \beta_0$ since $y(a, N-m) = y(a, N-m-1) = \frac{a+1}{\mu_R}$. Now construct a partition $D' = (R', T')$ of Ω in the following manner. For $\boldsymbol{x} \in \Omega$ such that $n \neq N-m-1$ let $\boldsymbol{x} \in R'$ if and only if $\boldsymbol{x} \in R$, that is, the decision policy remains unchanged for these states. For $\boldsymbol{x} \in \Omega$ such that $n = N-m-1$, let $\boldsymbol{x} \in R$ if and only if $a < \beta_0$. Then D' is 0^{th} slice $m+1$-level monotone, with $a_0^\star(N-m-1) = \beta_0 \geq a_0^\star(N-m)$, and we have shown the existence of at least one 0^{th} slice $m+1$-level user optimal policy. Furthermore, since β_0 is uniquely

determined by the partition D and D is unique up level m of the 0^{th} slice, it follows that any other 0^{th} slice $m+1$ user optimal policies are identical up to level $m+1$ of the 0^{th} slice with this one.

The argument for slices $k > 0$ follows in a very similar fashion.

Consider first $\boldsymbol{x} \in \Omega_k$ such that $l = N-1$, that is, $\xi = 1$. For all $a \geq 0$, $z(a, kN+N-1) = (k+1)/\mu_T$ and $y(a, N-1) = (a+1)/\mu_R$ is strictly increasing in a. Then $(a+1)/\mu_R \geq (k+1)/\mu_T$ if and only if $a > \frac{\mu_R(k+1)}{\mu_T} - 1$ and so $a_k^\star(N-1) = \lceil \frac{\mu_R(k+1)}{\mu_T} - 1 \rceil$.

Now suppose the hypothesis holds for k and $\xi = m$. We will show that it holds for k and $\xi = m+1$. By hypothesis there exists at least one policy $D \in \mathcal{D}_{km}^\star$ which is k^{th} slice $m-$level monotone and unique up to level m in slice k. Let $\beta = \min\{a : \ z(a, kN+N-m-1) \leq y(a, kN+N-m-1)\}$. By Lemma 4.3.1 condition *III* $z(a, kN+N-m-1)$ is decreasing in a, while $y(a, kN+N-m-1)$ is strictly increasing in a. Therefore, $z(a, kN+N-m-1) \leq y(a, kN+N-m-1)$ if and only if $a \geq \beta$, that is, the user optimal decision in state $(a, kN+N-m-1)$ is to choose Q_T if and only if $a \geq \beta$. By Lemma 4.3.1 condition *I* $z(a, kN+N-m) \leq z(a, kN+N-m-1)$ and therefore, $a_k^\star(kN+N-m) \leq \beta$ since $y(a, kN+N-m) = y(a, kN+N-m-1) = \frac{a+1}{\mu_R}$. Now construct a partition $D' = (R', T')$ of Ω in a similar manner to the construction above. For $\boldsymbol{x} \in \Omega$ such that $n \neq kN+N-m-1$ let $\boldsymbol{x} \in R'$ if and only if $\boldsymbol{x} \in R$, that is, the decision policy remains unchanged for these states. For $\boldsymbol{x} \in \Omega$ such that $n = kN+N-m-1$, let $\boldsymbol{x} \in R$ if and only if $a < \beta$. Then D' is k^{th} slice $m+1$-level monotone, with $a_k^\star(N-m-1) = \beta \geq a_k^\star(N-m)$, and we have shown the existence of at least one k^{th} slice $m+1$-level user optimal policy. Furthermore, since β is uniquely determined by the partition D and D

is unique up level m of the k^{th} slice, it follows that any other k^{th} slice $m+1$ user optimal policies are identical up to level $m+1$ of slice k with this one. Hence, by the principle of induction, the theorem follows. ■

4.4 Monotonicity

In this section we show some further monotonicity results for this system. In section 4.4.1 we show that the expected transit times via Q_T, $z_{D_k}(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega_k$, are monotone with respect to changes in λ, μ_R, λ_T, μ_T and also as the policy D changes monotonically. We will then use these results to show that a user optimal policy $D^\star \in \mathcal{D}^\star$ changes monotonically with respect to parameter changes in subsection 4.4.2. As in Chapter 2, since we will be varying both the decision policies and system parameters in this section, we let $\varpi_D(\boldsymbol{x};\Gamma)$ be the time until the batch that a marked general commuter joins is complete who, on arrival, sees the system X_D with parameters $\Gamma = (\lambda, \lambda_T, \mu_R, \mu_T)$, in state $\boldsymbol{x}$. Similarly, let $y_D(\boldsymbol{x};\Gamma)$ and $z_D(\boldsymbol{x};\Gamma)$ be the expected transit times for an arriving general user who sees the system in state $\boldsymbol{x}$ and joins Q_R or Q_T respectively.

4.4.1 Monotonicity of $\boldsymbol{z} = \{z(\boldsymbol{x});\, \boldsymbol{x} \in \Omega\}$ in λ, μ_R, λ_T, μ_T and D

Lemma 4.4.1 *Let $X_{D^{(1)}}$ and $X_{D^{(2)}}$ be processes with parameters $\Gamma_1 = (\lambda_1, \lambda_{T_1}, \mu_{R_1}, \mu_{T_1})$ and $\Gamma_2 = (\lambda_2, \lambda_{T_2}, \mu_{R_2}, \mu_{T_2})$ respectively, and common batch size, N. Furthermore, suppose that $X_{D^{(1)}}$ and $X_{D^{(2)}}$ are operating under*

slice monotone decision policies $D^{(1)} = (R^{(1)}, T^{(1)})$ *and* $D^{(2)} = (R^{(2)}, T^{(2)})$ *such that* $T^{(2)} \subseteq T^{(1)}$, *and that* $\lambda_1 \geq \lambda_2$, $\lambda_{T_1} \geq \lambda_{T_2}$, $\mu_{R_1} \leq \mu_{R_2}$ *and* $\mu_{T_1} \geq \mu_{T_2}$. *Then*

$$z_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) \leq z_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)}) \tag{4.11}$$

for any pair of states $\boldsymbol{x}_1 = (a_1, k_1 N + l_1)$ *and* $\boldsymbol{x}_2 = (a_2, k_2 N + l_2)$ *satisfying one of the conditions* 0 - IV *below, with* $0 \leq k_1 \leq k_2$.

0) $\boldsymbol{x}_1 = \boldsymbol{x}_2$,

I) $a_1 = a_2$, $l_1 > l_2$,

II) $a_1 < a_2$, $l_1 > l_2$; $a_2 - a_1 \leq l_1 - l_2$,

III) $a_2 < a_1$, $l_1 = l_2$,

IV) $a_2 < a_1$, $l_1 > l_2$.

Proof This proof is very similar in spirit to the proof of Lemma 4.3.1. We use a coupling argument on a joint process $\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right)$ to show that

$$\varpi_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) \overset{\text{st}}{\leq} \varpi_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)}) \tag{4.12}$$

for any $\boldsymbol{x}_1$, $\boldsymbol{x}_2 \in \Omega$ such that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy one of the conditions *0 - IV* with $k_1 \leq k_2$. The joint process $\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right)$ will now be such that $\boldsymbol{X}^{(1)}$ follows the law of $X_{D^{(1)}}$ and $\boldsymbol{X}^{(2)}$ follows the law of $X_{D^{(2)}}$. As above, we use a proof by forwards induction on the transitions of the process and observe the joint process until the completion time in the process $\boldsymbol{X}^{(1)}$ of the batch that the marked general commuter has joined. We will match the transitions made by both processes in such a way that both move together whenever possible. In particular, the service of a batch at Q_T in $\boldsymbol{X}^{(1)}$ will always be accompanied by the service of a batch at Q_T in $\boldsymbol{X}^{(2)}$. If $\varpi_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) \leq \varpi_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)})$ w.p.1 then, as above, $z_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) \leq z_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)})$ also holds. The argument is as

follows. The batch containing the marked customer in $\boldsymbol{X}^{(2)}$ either completes simultaneously with completion of the marked customer's batch in $\boldsymbol{X}^{(1)}$, or it has yet to complete. In either case, since in $\boldsymbol{X}^{(2)}$ the number of batches ahead of the marked customer's batch is no less than in $\boldsymbol{X}^{(1)}$, and the Q_T in both systems are served simultaneously whenever possible, with the only exceptions being if a batch in Q_T completes service in $\boldsymbol{X}^{(1)}$ without a corresponding service completion in $\boldsymbol{X}^{(2)}$, once the marked customer's batch is completed in $\boldsymbol{X}^{(1)}$, the time until service commences in $\boldsymbol{X}^{(1)}$ is no greater than in $\boldsymbol{X}^{(2)}$, and, since expected service times in $\boldsymbol{X}^{(1)}$ are also less than or equal to those in $\boldsymbol{X}^{(2)}$, the result follows.

As for Lemma 4.3.1, when $k_1 = k_2 = 0$ the proof is the same as that for the corresponding Lemma in Chapter 2, which is 2.4.1, and so we do not repeat it here.

Now suppose $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ satisfying one of the conditions *I-IV* are such that $k_2 > 0$. We begin by considering those cases where $\boldsymbol{x}_1$ is such that $n_1 = k_1 N + N - 1$. Then $\varpi_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) = 0 \leq \varpi_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)})$ w.p.1 (if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy the conditions *0* and *III*, that is $l_1 = l_2 = N - 1$, then $\varpi_{D^{(1)}}(\boldsymbol{x}_1; \Gamma^{(1)}) = \varpi_{D^{(2)}}(\boldsymbol{x}_2; \Gamma^{(2)}) = 0$ w.p.1).

To show that the inequality (4.12) holds under conditions *0* - *IV* for $l_1 < N-1$ we use coupling arguments. Consider the joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ with

transition rates

$$
(\boldsymbol{x}_1,\, \boldsymbol{x}_2) \to \begin{cases}
(\boldsymbol{x}_1 - e_1 I_{\{a_1>0\}},\, \boldsymbol{x}_2 - e_1 I_{\{a_2>0\}}) \\
\qquad \text{at rate } \mu_{R_1}\left(1 - I_{\{a_1=0, a_2=0\}}\right) \\
(\boldsymbol{x}_1,\, \boldsymbol{x}_2 - e_1 I_{\{a_2>0\}}) \\
\qquad \text{at rate } \left(\mu_{R_2} - \mu_{R_1}\right) I_{\{a_2>0\}} \\
(\boldsymbol{x}_1 + e_2,\, \boldsymbol{x}_2 + e_2) \\
\qquad \text{at rate } \lambda_{T_2} + \lambda_2 I_{\{\boldsymbol{x}_1 \in T^{(1)}, \boldsymbol{x}_2 \in T^{(2)}\}} \\
(\boldsymbol{x}_1 + e_2,\, \boldsymbol{x}_2) \\
\qquad \text{at rate } (\lambda_{T_1} - \lambda_{T_2}) + (\lambda_1 - \lambda_2) I_{\{\boldsymbol{x}_1 \in T^{(1)}\}} \\
(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2 + e_1) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in R^{(1)}, \boldsymbol{x}_2 \in R^{(2)}\}} \\
(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2) \\
\qquad \text{at rate } (\lambda_1 - \lambda_2) I_{\{\boldsymbol{x}_1 \in R^{(1)}\}} \\
(\boldsymbol{x}_1 + e_2,\, \boldsymbol{x}_2 + e_1) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in T^{(1)}, \boldsymbol{x}_2 \in R^{(2)}\}} \\
(\boldsymbol{x}_1 + e_1,\, \boldsymbol{x}_2 + e_2\big) \\
\qquad \text{at rate } \lambda_2 I_{\{\boldsymbol{x}_1 \in R^{(1)}, \boldsymbol{x}_2 \in T^{(2)}\}} \\
\big((\boldsymbol{x}_1 - Ne_2) I_{\{l_1=N\}},\, (\boldsymbol{x}_2 - Ne_2) I_{\{l_2=N\}} \big) \\
\qquad \text{at rate } \mu_{T_2}(1 - I_{\{l_1 \neq N, l_2 \neq N\}}) \\
\big((\boldsymbol{x}_1 - Ne_2) I_{\{l_1=N\}},\, \boldsymbol{x}_2 \big) \\
\qquad \text{at rate } (\mu_{T_1} - \mu_{T_2}) I_{\{l_1=N\}}
\end{cases} \tag{4.13}
$$

Then $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ follow the laws of $X_{D^{(1)}}$ and $X_{D^{(2)}}$ respectively. In the joint process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ whenever possible a departure from Q_R in $\boldsymbol{X}^{(1)}$ is always accompanied by a departure from Q_R in $\boldsymbol{X}^{(2)}$, a departure from Q_T in $\boldsymbol{X}^{(2)}$ is always accompanied by a departure from Q_T in $\boldsymbol{X}^{(1)}$, a committed

user's arrival to Q_T in$\boldsymbol{X}^{(2)}$ is always accompanied by committed user's arrival to Q_T in $\boldsymbol{X}^{(1)}$ and a general arrival to $\boldsymbol{X}^{(2)}$ is always accompanied by a general arrival to $\boldsymbol{X}^{(1)}$. The transitions for for the process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ are such that if $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})(0) = \big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ with $0 \leq l_2 \leq l_1 < N-1$ and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy one of the conditions *0* - *IV* then w.p.1 $L_1(t) \geq L_2(t)$ until $\boldsymbol{X}^{(1)}$ reaches the set of states H_N.

As previously, let $(\boldsymbol{x}_1', \boldsymbol{x}_2')$ denote the state of the process $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ immediately after a transition out of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Then we will show that if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies any of the conditions *0* - *IV* then so does $(\boldsymbol{x}_1', \boldsymbol{x}_2')$.

We consider first the transitions due to service completions in Q_R. If $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 - e_1, \boldsymbol{x}_2 - e_1)$ then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ satisfies whichever condition $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied. If $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1, \boldsymbol{x}_2 - e_1)$ then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ satisfies the condition *III* if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied the conditions *0* or *III*, it satisfies the condition *IV* if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied the conditions *I* or *IV* and if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied the condition *II* then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ either preserves the same condition (if $a_2 > 1$) or satisfies the condition *I* (if $a_2 = 1$). If $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 - e_1, \boldsymbol{x}_1)$, which occurs if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied conditions *III* or *IV*, then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ either preserves the condition that $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied if $a_1 > 1$ or it satisfies the condition *0*, if $a_1 = 1$ and $l_1 = l_2$, or *I*, if $a_1 = 1$ and $l_1 > l_2$.

Now, consider transitions due to arrivals from either the committed train users' stream or the general users' stream. If $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_2)$ or $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_1)$ and $l_1' \leq N-1$ then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ satisfies whichever condition $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied. If $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2)$ and $l_1' \leq N-1$ then $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ satisfies the condition *I* if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied the conditions *0* or *I*, $\big(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\big)$ satisfies the same condition if $\big(\boldsymbol{x}_1, \boldsymbol{x}_2\big)$ satisfied the condition *II* and

$\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *IV* if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied the conditions *III* or *IV*. If $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_2, \boldsymbol{x}_2 + e_1)$ and $l_1' \leq N - 1$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *II* if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied the conditions *0* , *I* or *II*. Furthermore, if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied the conditions *III* or *IV* then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *I*, if $a_1 = a_2 + 1$, or $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *IV*, if $a_1 > a_2 + 1$. If $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2 + e_2)$ (which occurs only if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied the condition *II*, since the two policies are slice monotone, with $T^{(2)} \subset T^{(1)}$ and $l_1' < N - 1$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ preserves the same condition if $a_2 > a_1 + 1$ and $l_1 > l_2 + 1$, $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *I* if $a_2 = a_1 + 1$ and $l_1 > l_2 + 1$ or $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies the condition *0* if $a_2 = a_1 + 1$ and $l_1 = l_2 + 1$. Finally if $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 + e_1, \boldsymbol{x}_2)$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies condition *III* if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied conditions 0 or *III*, $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfies condition *IV* if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied conditions *I* or *IV* and if $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied condition *II* then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ either preserves condition *II* (if $a_2 > a_1 + 1$) or satisfies condition *I* (if $a_2 = a_1 + 1$). If $l_1' = N$ then the process stops and $\varpi_{D^{(1)}}^{(1)}(\boldsymbol{x}_1; \Gamma^{(1)}) = 0 \leq \varpi_{D_k^{(2)}}^{(2)}(\boldsymbol{x}_2; \Gamma^{(2)})$ w.p.1. In each case, either $l_1' = N$ or $l_1' \geq l_2'$ with $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfying one of the conditions *0* - *IV*.

Finally, consider the transitions due to service completions in Q_T, which can only occur if $n_i > N$, $i = 1, 2$. If $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - Ne_2, \boldsymbol{x}_2 - Ne_2)$ then $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ continues to satisfy whichever condition $\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)$ satisfied and both processes are one batch closer to being served. If $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1 - Ne_2, \boldsymbol{x}_2)$ then $\boldsymbol{X}^{(1)}$ is one batch closer to being served than $\boldsymbol{X}^{(2)}$ (note that the case $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right) = (\boldsymbol{x}_1, \boldsymbol{x}_2 - Ne_2)$ cannot occur). In both cases, $k_1' \leq k_2'$.

Thus all possible $\left(\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}\right)$ satisfy one of the conditions *0*-*IV* and so we are done. The argument given earlier then completes the proof. ■

4.4.2 Monotonicity of $D^\star$ in λ, μ_R, λ_T and μ_T.

We are now ready to show that the user optimal policy $D^\star \in \mathcal{D}^\star$ changes monotonically with changes in parameters $\lambda, \mu_R, \lambda_T$ and μ_T.

Theorem 4.4.1 *Let $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ be two processes with parameters $\Gamma_1 = (\lambda_1, \mu_{R_1}, \lambda_{T_1}, \mu_{T_1})$ and $\Gamma_2 = (\lambda_2, \mu_{R_2}, \lambda_{T_2}, \mu_{T_2})$ respectively, and common batch size N. Furthermore, suppose $\lambda_1 \geq \lambda_2$, $\mu_{R_1} \leq \mu_{R_2}$, $\lambda_{T_1} \geq \lambda_{T_2}$ and $\mu_{T_1} \geq \mu_{T_2}$. Let $D^{\star(1)}$ and $D^{\star(2)}$ be the user optimal policies for $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ respectively. Then $a_k^{\star(1)}(n) \leq a_k^{\star(2)}(n)$ for all $kN \leq n \leq (k+1)N - 2$ and $k \geq 0$.*

Proof. As in the proof of Theorem 4.3.1, we use proof by induction on $\xi = N - l$ such that $\xi = 1, 2, \ldots, N$ and $k \geq 0$.

Again, consider first the case $k = 0$. For this case $n \in \{0, 1, \ldots, N-1\}$. We begin with $\xi = 1$, that is, $n = N - 1$. Then $z_{D^{\star(i)}}((a, N-1); \Gamma_i) = 1/\mu_{T_i}$ and $y_{D^{\star(i)}}((a, N-1); \Gamma_i) = (a+1)/\mu_{R_i}$, $i = 1, 2$ for all $a \geq 0$ and since $\mu_{R_1} \leq \mu_{R_2}$ and $\mu_{T_1} \geq \mu_{T_2}$ we have $a_0^{\star(1)}(N-1) \leq a_0^{\star(2)}(N-1)$, where $a_0^{\star(i)}(N-1) = \lceil \frac{\mu_{R_i}}{\mu_{T_i}} - 1 \rceil$, $i = 1, 2$.

Now suppose that $a_0^{\star(1)}(n) \leq a_0^{\star(2)}(n)$ holds for $N-m \leq n \leq N-1$. We will show that the required inequality holds for $\xi = m+1$, that is, $n = N-m-1$. Let $D \in \mathcal{D}_{0m}^{\star(1)}$ be a 0^{th} slice $m-$level user optimal policy for $\boldsymbol{X}^{(1)}$. Remember that $y_D(\boldsymbol{x}; \Gamma_1) = (a+1)/\mu_{R_1}$ is independent of both D and l. Similarly to Lemma 4.3.1 let $\beta_1 = \min\{a : \ z_D((a, N-m-1); \Gamma_1) \leq y((a, N-m-1); \Gamma_1)\}$. By the same arguments as in the proof of Lemma 4.3.1, $\beta_1 = a_0^{\star(1)}(N-m-1)$ and then by Lemma 4.4.1 $z_D((a, N-m-1); \Gamma_1) \leq z_{D^{\star(2)}}((a, N-m-1); \Gamma_2)$

for $a \geq 0$. Finally, since $z_D((a, N-m-1); \Gamma_1)$ and $z_{D_2^\star}((a, N-m-1); \Gamma_2)$ are both decreasing in a, and $y_D((a, N-m-1); \Gamma_1) = (a+1)/\mu_{R_1} \geq (a+1)/\mu_{R_2} = y_{D_2^\star}((a, N-m-1); \Gamma_2)$ we have $a_0^{\star(1)}(N-m-1) \leq a_0^{\star(2)}(N-m-1)$. Thus by the principle of induction $a_0^{\star(1)}(n) \leq a_0^{\star(2)}(n)$ holds for $k = 0$ and all $n = 0, 1, \ldots, N-1$.

Now consider the case $k > 0$. The argument for $\xi = 1$ follows exactly as for the case $k = 0$, but with $z_{D^{\star(i)}}((a, N-1); \Gamma_i) = (k+1)/\mu_{T_i}$ and $a_0^{\star(i)}(N-1) = \lceil \frac{\mu_{R_i}(k+1)}{\mu_{T_i}} - 1 \rceil$, $i = 1, 2$. The argument for $\xi > 1$ also follows that given above, but with the inductive hypothesis now being that $a_i^{\star(1)} \leq a_i^{\star(2)}$ holds for all i such that $0 \leq i < k$ and $a_k^{\star(1)}(n) \leq a_k^{\star(2)}(n)$ holds for $N-m \leq n \leq N-1$. Consider now a $D \in \mathcal{D}_{km}^{\star(1)}$ and the argument follows exactly that given above for $k = 0$, but with k substituted for 0 in the requisite statements. ■

4.5 Discussion and Conclusion

We have given proofs of existence, uniqueness and monotonicity properties for the state-dependent user optimal policy, $D^\star$, for the system we study in this thesis with $\cdot|M|1$ and $\cdot|M^{(N)}|1$ queues. For any given parameters, $N \geq 2$, $\lambda > 0$, $\lambda_T > 0$, $\mu_R > 0$ and $\mu_T > 0$, we can find a user optimal policy that is unique and slice monotone. Furthermore, $D^\star$ changes monotonically with changes in parameter values as shown in Theorem 4.4.1.

Figure 4.2 plots part of the user optimal policy for the system with a $\cdot|M|1$ and a $\cdot|G^{(N)}|\infty$ queues (i) and part of the user optimal policy for the system with a $\cdot|M|1$ queue and a $\cdot|M^{(N)}|1$ queue (ii) with the same parameter values. For the system with a single batch server $\cdot|M^{(N)}|1$ queue we have only shown

the 0^{th} and 1^{st} slices for any level 1 user optimal policy $D_1^\star \in \mathcal{D}_1^\star$. From Chapter 2 the user optimal policy $D^\star \in \mathcal{D}^\star$ for the system with an infinite server $\cdot|G^{(N)}|\infty$ queue always needs only the 0^{th} slice. For these user optimal policies we let $N = 3$, $\lambda = \lambda_T = 1$ and $\mu_R = \mu_T = 2$.

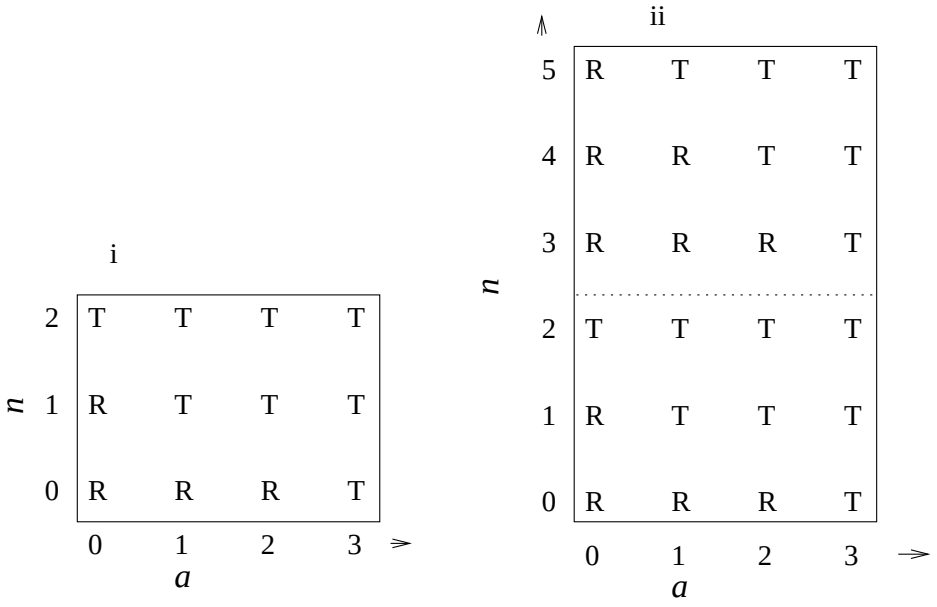


Figure 4.2: Plots i and ii are the unique user optimal policies, $D^\star$ and $D_1^\star$, for the system with $\cdot|G^{(N)}|\infty$ and $\cdot|M^{(N)}|1$ queues respectively, when $\lambda = \lambda_T = 1$, $\mu_R = \mu_T = 2$, $N = 3$. The switching curve $a^\star$ for plot (i) is $a^\star = (3, 1, 0)$ and those for plot (ii) are $a_0^\star = (3, 1, 0)$ and $a_1^\star = (3, 2, 1)$.

Chapter 5

Probabilistic Routing for $\cdot|M|1$ and $\cdot|M^{(N)}|1$ queues

5.1 Introduction and model description

In this chapter we consider the user equilibrium for $M|M|1$ and $M|M^{(N)}|1$ queues when arriving general users choose the queue to join probabilistically as in Chapter 3. Existence, uniqueness and stability properties of the user equilibrium were completely analyzed in [2] for a system with an $M|M|1$ queue and an infinite-server batch service $\left(M|G^{(N)}|\infty\right)$ queue instead of an $M|M^{(N)}|1$ queue, and we also explored them for a system consisting of $M|G|1$ and $M|G^{(N)}|\infty$ queues in Chapter 3. In those analyses the plentiful number of servers available in the batch service queue, Q_T, simplified the equations for the expected transit time for a general user joining the $M|G^{(N)}|\infty$ queue and the analysis is further simplified if the single server queue, Q_R, is an $M|M|1$ queue

(see [2]). However, it is more realistic to consider a finite number of servers for Q_T. Due to the complicated nature of the equations for the expected transit time for arriving general users who use this queue, W_T, we consider only the case when it has a single server. For the same reason we also consider an $M|M|1$ queue instead of an $M|G|1$ queue (see Chapter 3).

The two queues, Q_R and Q_T, are modelled as an $M|M|1$ queue with service rate μ_R and an $M|M^{(N)}|1$ queue with service rate μ_T and batch size N, respectively. As defined in Chapter 3 we assume that general users arrive to the system at rate λ and choose Q_R with probability p_R and Q_T with probability $1-p_R$ independently of everyone else, while confirmed train users arrive to the system at rate λ_T and always join Q_T. The service times at Q_R and Q_T are both exponentially distributed with $\mu_R > \lambda\, p_R$ and $N\mu_T > \lambda\,(1-p_R) + \lambda_T$ to ensure stability of this system. All inter-arrival times, routing decisions and service times are independent of one another. Let $(A(t),\, N(t))$ be the state of the process at time t, where $A(t) \in \mathbb{Z}^+$ is the number waiting in Q_R including any commuter in service and $N(t) \in \mathbb{Z}^+$ is the number awaiting for service in Q_T including those in the batch being served. A service cannot start at Q_T unless the batch of N commuters is completed. So if the train (server) is not busy and there are $N-1$ commuters waiting for service the arrival of the N^{th} commuter triggers the commencement of service. If the server is busy then the number of commuters in the queue is greater than or equal to N and the N commuters at the head of the queue are in service leaving $n-N$ in the queue waiting for service to commence. The process $\{(A(t),\, N(t)); t \geq 0\}$ is Markov

with state space $\mathcal{S} \in \mathbb{Z}^+ \times \mathbb{Z}^+$ and transitions as follows.

$$
\begin{array}{rcll}
 & \overset{\lambda p_R}{\nearrow} & (a+1,\, n) & \\
 & \overset{\Lambda}{\nearrow} & (a,\, n+1), & \text{where } \Lambda = \lambda\,(1-p_R) + \lambda_T \\
(a,\, n) & \xrightarrow{\mu_R} & (a-1,\, n) & \text{if } a \geq 1 \\
 & \overset{\mu_T}{\searrow} & (a,\, n-N) & \text{if } n \geq N
\end{array}
$$

We use results from Medhi [43] to obtain expressions for the expected transit time through a single server batch service queue. Medhi [43] considered a $\cdot|M^{(a,b)}|1$ queue and obtained for this system the distribution of the number in the queue, the distribution of the waiting time and the expected waiting time for an arriving user in steady state, where a and b corresponds to the minimum and maximum number of units for each batch, respectively. This rule is referred to as the *general rule* for batch service by, for example, Neuts [48] and Medhi [42, 43]. The rule is that the server can start service as soon as a minimum number of units 'a' is present in the queue and has maximum batch service capacity 'b'. Neuts [48] studied the transient state distribution for the same network model. Raj and Manoharan [51] considered a similar approach for a single and batch service $M|M|1$ queue with a control limit policy, 'c', where the server serves a single customer if the queue length is less than or equal to a threshold c and serves all the customers together in a batch if the queue length is greater than c. They then obtained the steady state probabilities, the waiting time distribution and the expected waiting time of an arriving unit. Lee, Lee and Chae [36] and Park, Kim and Chae [49] considered a single-server batch service queue with a fixed batch size as we are considering here. The service times in [36] and [49] are both generally distributed while the service

times of the system we study here are exponentially distributed. Lee, Lee and Chae [36] derived the decompositions of the queue length distributions for a single-server batch service queue with single and multiple vacations. Park, Kim and Chae [49] considered a single-server, two-phase queueing system. In their model, the server will not start the service in the first-phase unless there are k customers in the queue. After the first-phase service, the k customers in the batch receive individual services according to a first-come first-served (FCFS) principle. Any customer arriving during the service in the first-phase or the second-phase has to wait for the next batch service. They derived the steady state distribution for the system's queue length and determined the optimal batch size that minimizes the long-run average cost under a linear cost structure. Murugan and Kalyanaraman [46] considered an $M|M|1$ queue with server vacations. The single server not only provides first come first served service but also provides an additional optional service to the customers in batches of fixed size greater than or equal to 1. They obtained the probability generating function for the queue length and used Little's Law to obtain the expected waiting time of a customer in the queue. A similar approach was used by Medhi [43] to obtain the expected queue length and expected waiting time for a general user going via the $M|M^{(N)}|1$ queue.

The structure of this chapter is as follows. We provide some preliminary results in section 5.2. We calculate the user equilibria for this system model and provide numerical examples to support these results in section 5.3. We then conclude in section 5.4.

5.2 Preliminaries

From Chapter 3 if the coefficient of variation equals 1, that is, $c^2 = 1$, then, in equilibrium, the expected transit time of an arriving general user who goes to the destination via Q_R is

$$W_R(p_R) = \frac{1}{\mu_R - \lambda\, p_R} \tag{5.1}$$

From Medhi [43] (equation 4.3.31) if $a = b = N$ then that equation gives the expected waiting time, plus the expected service time $1/\mu_T$ to give W_T as follows.

$$W_T = \frac{1}{\mu_T} + \frac{1}{\Lambda}\left[\frac{N-1}{2} + \frac{r^{N+1}}{1-r}\right] \tag{5.2}$$

where $\Lambda = \lambda_T + \lambda(1 - p_R)$ and r is real and unique if and only if $\rho = \frac{\Lambda}{N\mu_T} < 1$, and satisfies (5.3)

$$N\,\rho \;=\; \frac{\Lambda}{\mu_T} = \frac{r(1-r^N)}{1-r}, \qquad \text{(see Medhi [43] equation 4.3.10)} \tag{5.3}$$

$$\text{for} \qquad \rho \in (0,\, 1).$$

$$\text{Also} \qquad \rho \le r \le \rho^{2/(N+1)} \tag{5.4}$$

In order for this system to be stable we need $p_R < \mu_R$ and $1 - p_R + \lambda_T < N\mu_T$.

Figure 5.1 plots W_T versus the total arrival rate, $\Lambda = \lambda_T + (1 - p_R)\,\lambda$, to Q_T when $\lambda = 0, p_R = 0, \lambda_T \in [0.1, 5.9], \mu_T = 3, N = 2$. The figure plots W_T for both an infinite-server batch service queue and a single-server batch service

queue. In this numerical example we see that W_T is decreasing in Λ for the infinite-server batch service queue, while it first decreases and then increases in Λ for the single-server batch service queue.

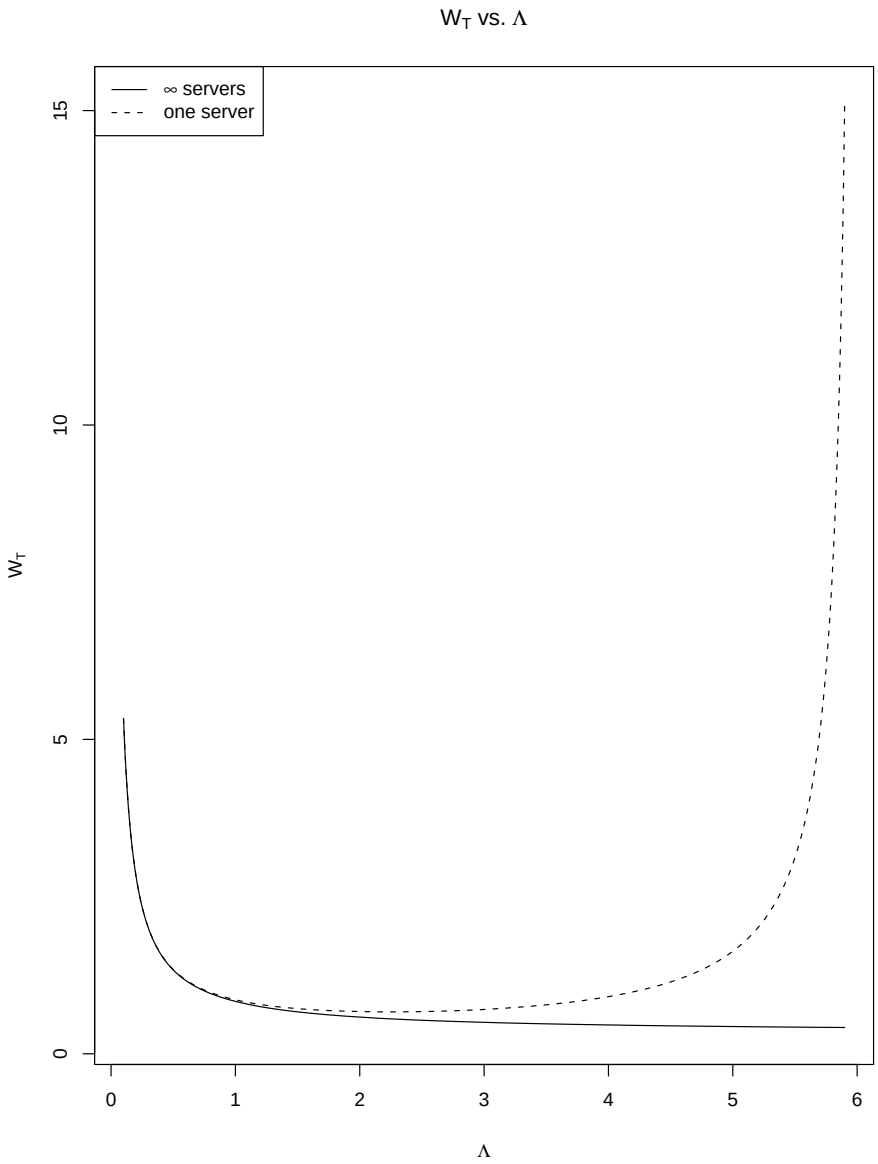


Figure 5.1: Plot of W_T vs Λ with $\lambda = 0$, $p_R = 0, \lambda_T = [0.1, 5.9]$, $\mu_T = 3$, $N = 2$.

5.3 User equilibrium

A value $p_R \in (0,\, 1)$ is a mixed user equilibrium, p_R^{EQ}, if $W_T(p_R) = W_R(p_R)$ as defined in Chapter 3. Also $p_R^{EQ} = 0$ is a pure equilibrium if $W_T(0) < W_R(0)$. Similarly $p_R^{EQ} = 1$ is a pure equilibrium if $W_T(1) > W_R(p_R)$. Furthermore, p_R^{EQ} is stable if there exists $\epsilon > 0$ such that for $p \in \big(p_R^{EQ},\, \min(p_R^{EQ}+\epsilon,\, 1)\,\big), W_R(p) < W_T(p)$ and for $p \in \big(p_R^{EQ},\, \max(0,\, p_R^{EQ} - \epsilon)\,\big),\ W_R(p) > W_T(p)$.

As in Chapter 3 we consider the difference between W_T and W_R since our definitions of the user equilibrium and stable user equilibrium are in terms of the comparison between $W_T(p_R)$ and $W_R(p_R)$. Also without loss of generality, we take $\lambda = 1$. Then the difference $W_T - W_R$ is as follows

$$
\begin{aligned}
[W_T - W_R](p_R) &= \frac{1}{\mu_T} + \frac{1}{\Lambda}\left[\frac{N-1}{2} + \frac{r^{N+1}}{(1-r)}\right] - \frac{1}{\mu_R - p_R} \\
&= \big(\mu_R - p_R\big)\left(1 + \lambda_T - p_R + \big(\frac{N-1}{2} + \frac{r^{N+1}}{1-r}\,\big)\mu_T\right) \\
&\qquad - \big(1 + \lambda_T - p_R\big)\,\mu_T \\
&= \frac{Q(p_R)}{\vartheta} = \frac{p_R^2 - 2\,\beta_1\,p_R + \beta_0}{\vartheta}, \qquad (5.5)
\end{aligned}
$$

where $\beta_1 = \frac{1}{2}\big(1 + \lambda_T + \mu_R + (\frac{N-3}{2} + \frac{r^{N+1}}{1-r})\,\mu_T\big)$, $\beta_0 = \big(1 + \lambda_T + (\frac{N-1}{2} + \frac{r^{N+1}}{1-r}\,)\mu_T\big)\,\mu_R - (1+\lambda_T)\,\mu_T$ and $\vartheta = (1 + \lambda_T - p)\,(\mu_R - p_R)\,\mu_T$.

Since r in the equation of $Q(\cdot)$ is a function of p_R, $Q(\cdot)$ here is not a quadratic in p_R and so we use (5.5) to provide numerical examples for this system. The following examples revisit some of the numerical examples in [2].

In the following numerical examples multiple user equilibria also exist for some parameter values.

The following numerical examples were produced using the R code in Appendix A.2. This code was checked using some simple case such as when $\lambda = \lambda_T = \mu_R = \mu_T = 1$ and $N = 2$.

Figure 5.2 plots W_T and W_R on the same axis against p_R in the first column and the values of the difference, $W_T - W_R$ against p_R in the second column when $\lambda = 1$, $\lambda_T = 0.1$, $\mu_T = 3$ and $N = 2$ for all plots with μ_R varying. For the first row of plots $\mu_R = 1.2$ and an unstable equilibrium occurs at $p_R = 0.98306$ and two stable user equilibria occur at $p_R = 0$ and $p_R = 1$. For the second row of plots $\mu_R = 1.3$ and $p_R = 0.20186$ and $p_R = 1$ are the stable user equilibria while $p_R = 0.79814$ is an unstable user equilibrium. Figure 5.3 plots W_T and W_R on the same axis for some parameter values.

Figure 5.4 plots the expected transit time, $W(p_R^\star) = p_R^\star W_R(p_R^\star) + (1 - p_R^\star) W_T(p_R^\star)$ as the service rate at Q_R, μ_R, varies from 0.1 to 3, with $\lambda = 1, \lambda_T = 0.1, N = 3$ and $\mu_T = 0.4$ for the first plot and $\mu_T = 1$ for the second plot. For both plots W is calculated at increments of 0.01. In the first example where $\mu_T = 0.4$ all general arrivals join Q_R for $\mu_R \geq 1.079$ and their expected transit time is then $W = 1/(\mu_R - 1)$ and for $\mu_R < 1.08$ the general arrivals are split between Q_R and Q_T. For the latter case W is decreasing for $\mu_R < 0.781$ and then increasing up to $\mu_R < 1.08$. In the second example where $\mu_T = 1$ all general arrivals join Q_T for $\mu_R < 0.48$ and $W = W_T = 2.1459$; for $\mu_R > 1.09$, all general arrivals join Q_R and $W = 1/(\mu_R - 1)$; for the intermediate values of μ_R the general arrivals are split between Q_R and Q_T and W increasing on this interval. Figure 5.5 plots W as the general users arrival rate, λ, varies

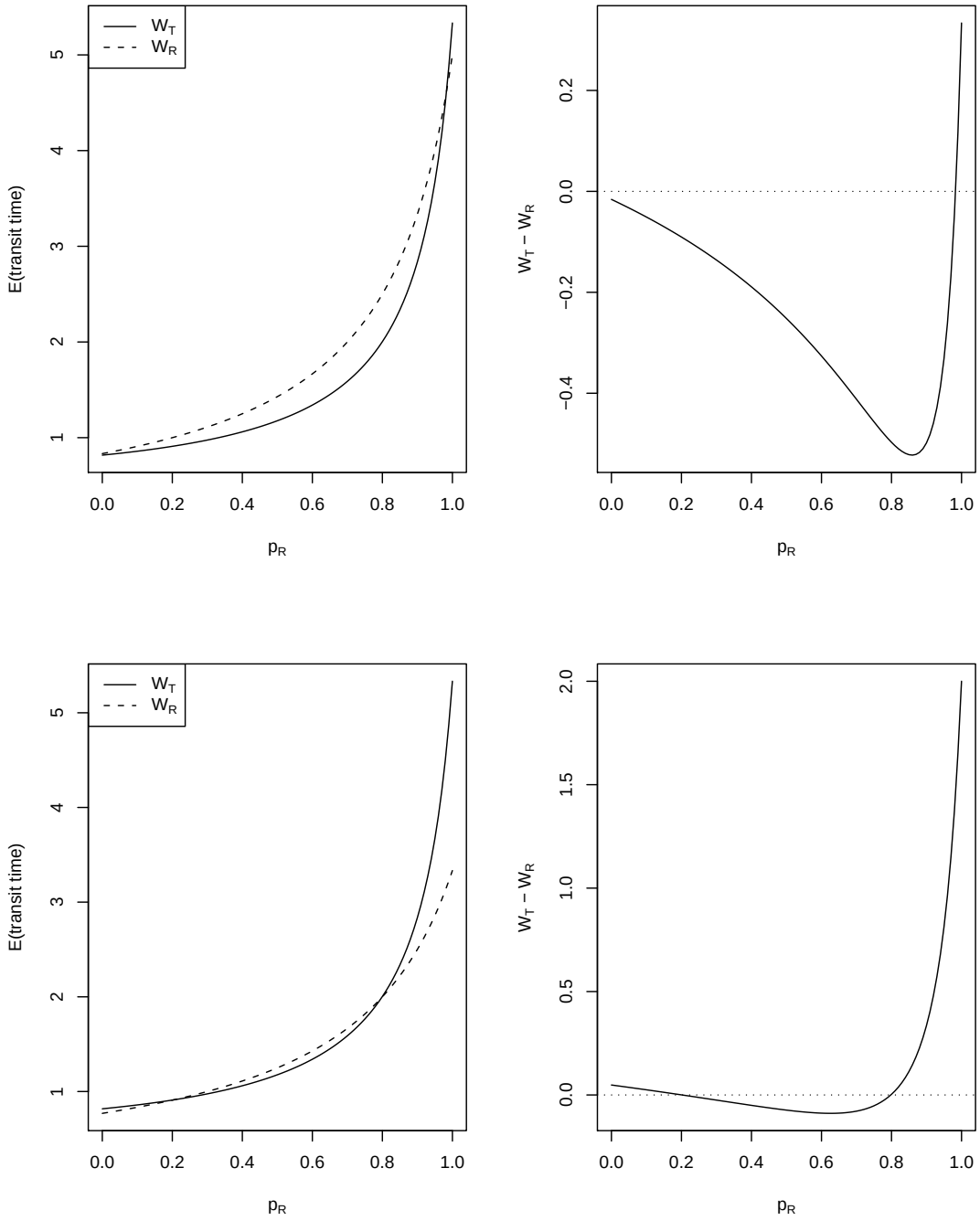


Figure 5.2: The first column contains plots of W_T and W_R on the same axis and the second column contains the plots of $W_T - W_R$ when $\lambda = 1, \lambda_T = 0.1$, $\mu_T = 3, N = 2$ and $\mu_R = 1.2$ for the first row of plots and $\mu_R = 1.3$ for the second row of plots. For the first row, $p_R = 0.98306$ is an unstable equilibrium and $p_R = 0$ and $p_R = 1$ are the stable equilibria. For the second row, $p_R = 0.20186$ and $p_R = 1$ are the stable equilibria and $p_R = 0.79814$ is an unstable user equilibrium.

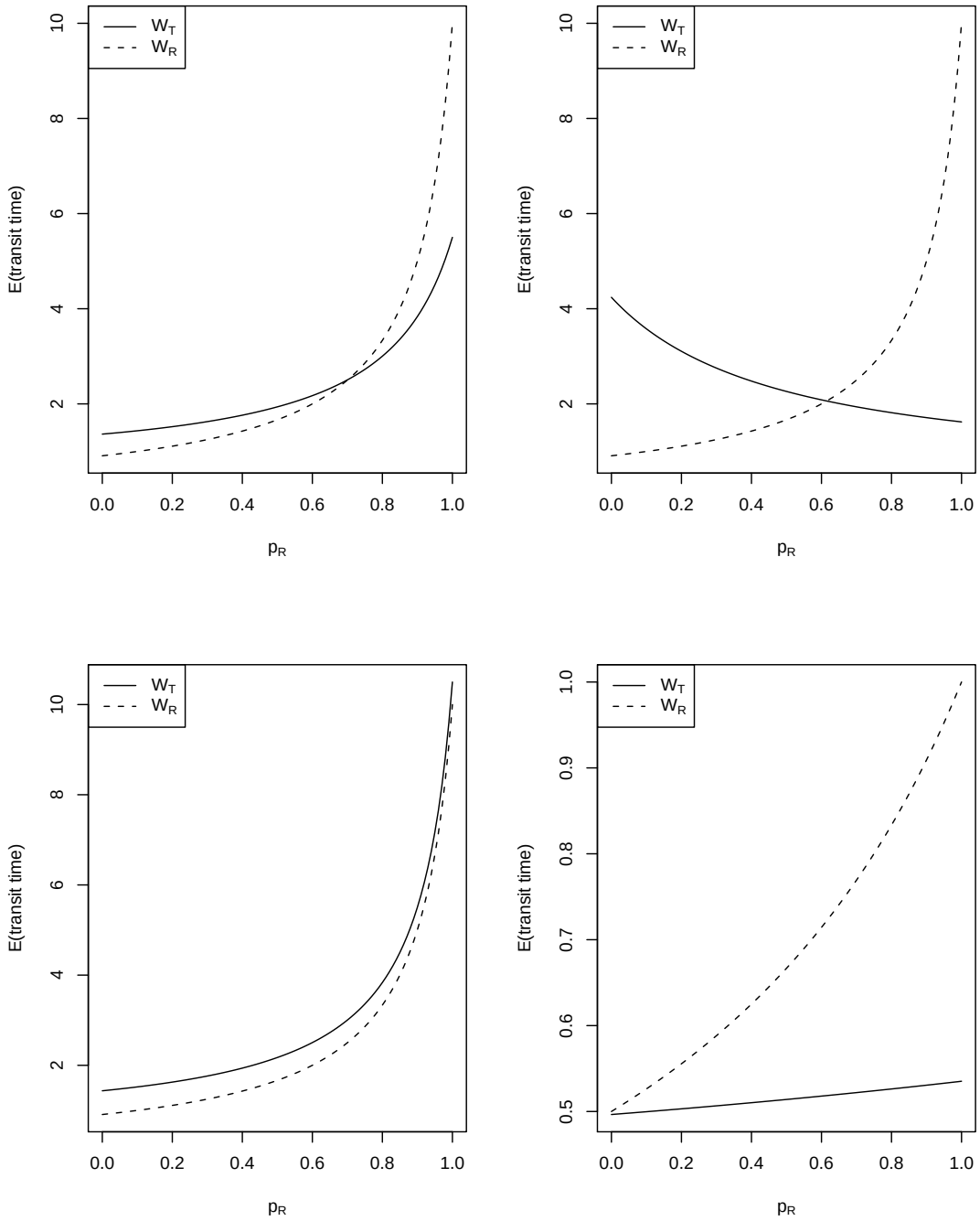


Figure 5.3: Plots of W_T and W_R on the same axis for different parameter values. For the top row of plots and the bottom left corner plot, $\lambda = 1$, $\mu_R = 1.1$, $\mu_T = 2$, $N = 3$ and $\lambda_T = 0.2$ for the top left corner plot, $\lambda_T = 0.1$ for the top right corner plot and $\lambda_T = 4.5$ for the bottom left corner plot. The stable user equilibria for each respective plot occurs at $p_R = 0.7018$, $p_R = 0.61483$ and $p_R = 1$. For the last plot, $\lambda = 1$, $\mu_R = 2$, $\mu_T = 5$, $\lambda_T = 6.2$ and $N = 5$ and the stable equilibrium occurs at $p_R = 0$.

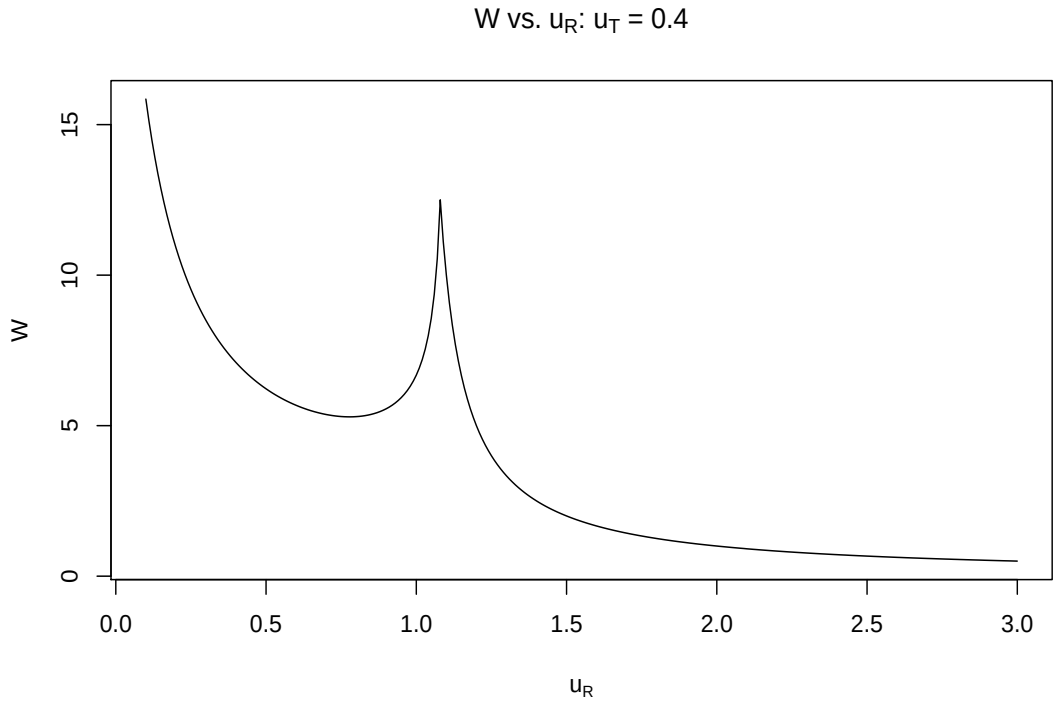
from 0.1 to 2.5 with $\mu_R = 1, \lambda_T = 0.1, \mu_T = 1$ and $N = 3$. It also plots W as the committed train users' arrival rate, λ_T, varies between 0.1 and 4 with $\mu_R = 1, \lambda = 1, \mu_T = 2$ and $N = 3$. We see the Downs-Thomson effect appearing in these numerical examples.

5.4 Discussion and conclusion

In this chapter we considered numerical examples for the network consisting of $\cdot|M|1$ and $\cdot|M^{(N)}|1$ queues with general users choosing their route probabilistically. In order to ensure stability of the system we need $p_R\,\lambda < \mu_R$ and $\lambda_T + (1-p_R)\,\lambda < N\,\mu_T$, where $p_R\,\lambda$ and $\lambda_T + (1-p_R)\,\lambda$ are the arrival rates into Q_R and Q_T respectively. If we consider the network with a $\cdot|G^{(N)}|\infty$ queue instead of a $\cdot|M^{(N)}|1$ queue we only need $p_R\lambda < \mu_R$ to ensure the stability of the system, since Q_T has an infinite number of servers and service of a complete batch commences as soon as that batch is full.

From the numerical examples multiple user equilibria can exist for this system, in contrast with the state-dependent routing case, where the user equilibrium is always unique. In addition, the Downs-Thomson effect can be clearly seen in the numerical examples.

We see from the numerical examples that the expected delay W does not always decrease as the confirmed train users' rate λ_T is increased as we expected. This was not the case for the system with a $\cdot|G^{(N)}|\infty$ queue (see [2]). As for the network with $\cdot|G^{(N)}|\infty$ queue, numerical examples with batch size $N > 2$ do not have multiple user equilibria and so we conjecture that this is the case here also.



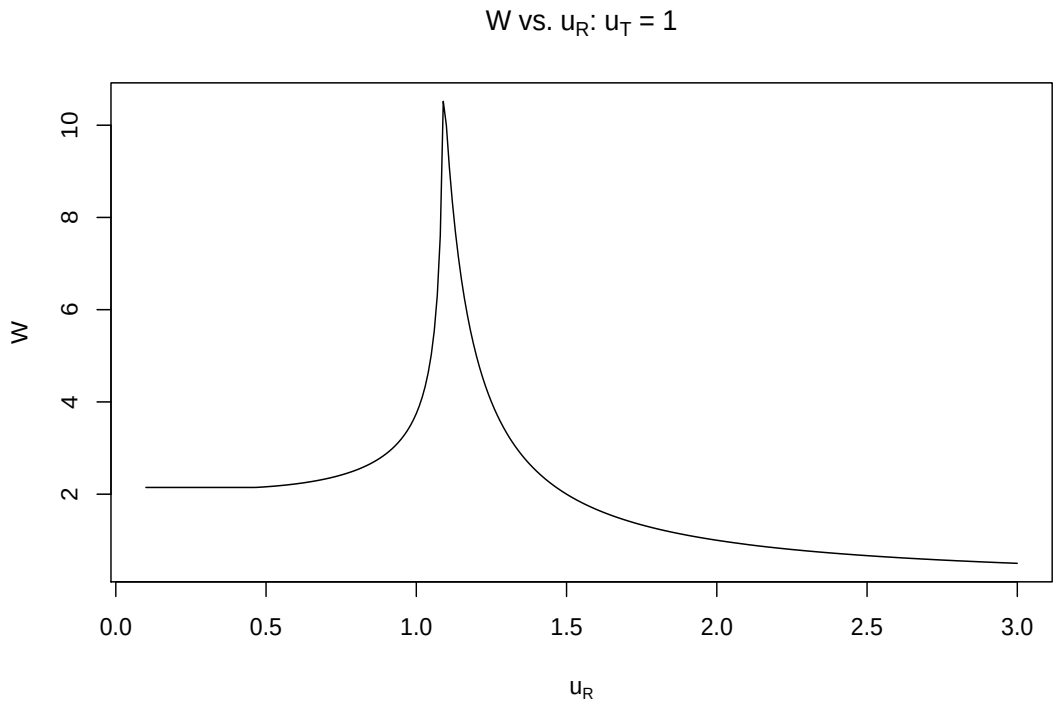


Figure 5.4: The expected transit time, W, versus μ_R when $\lambda = 1, \lambda_T = 0.1, N = 3$ and $\mu_T = 0.4$ and $\mu_T = 1$ for the first plot and second plot respectively. For both plots W is calculated at increments of 0.01.

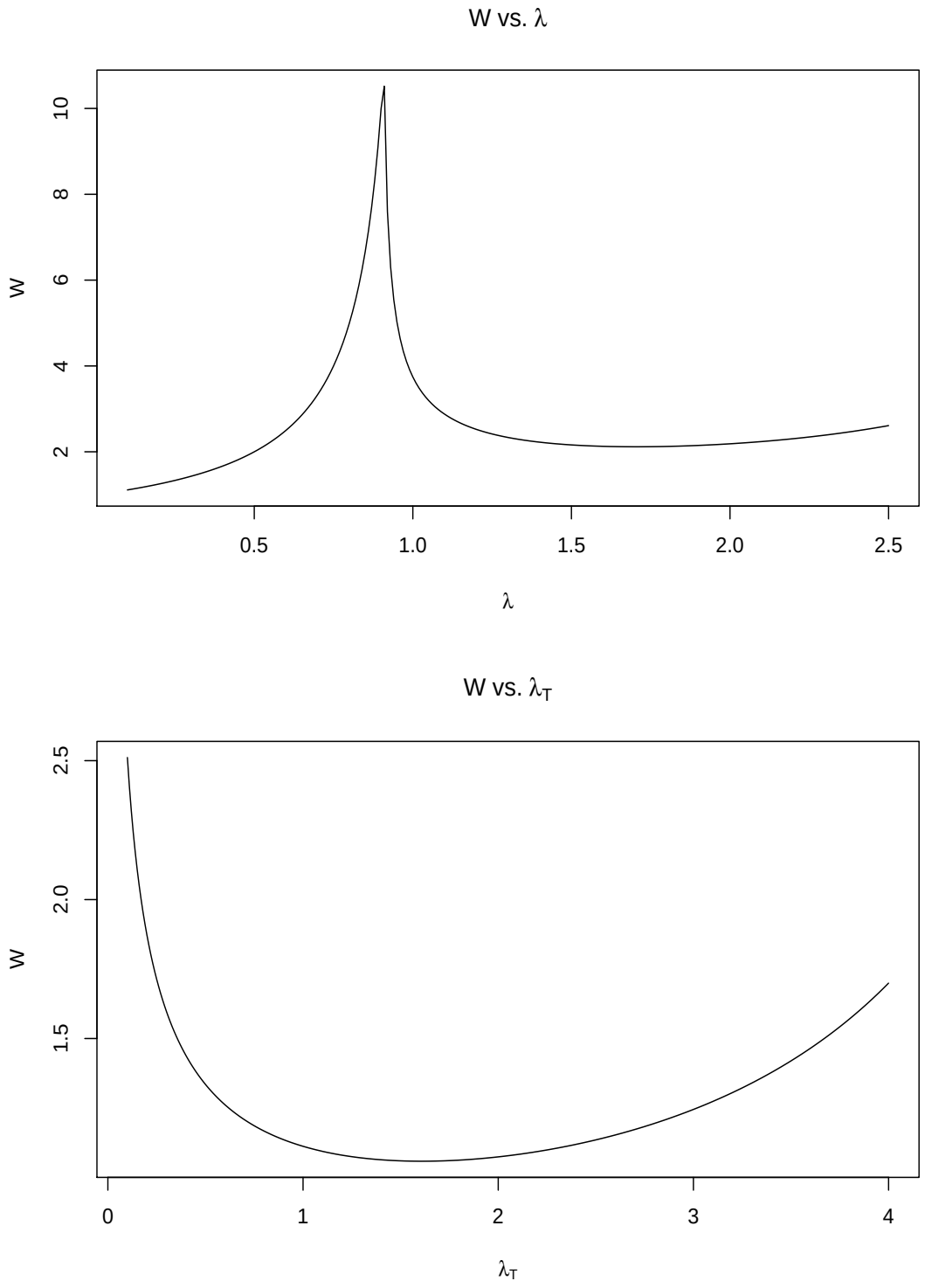


Figure 5.5: First plot: The expected transit time, W, versus λ when $\mu_R = 1, \lambda_T = 0.1, \mu_T = 1, N = 3$ and $0.1 \leq \lambda \leq 2.5$. Second plot: W versus λ_T when $\mu_R = 1, \lambda = 1, \mu_T = 2, N = 3$ and $0.1 \leq \lambda_T \leq 4$. For both plots W is calculated at increments of 0.01.

Chapter 6

Discussion and Conclusion

In this chapter we summarize and discuss our findings and some possible extensions to the system studied in this thesis.

We have studied existence, uniqueness and monotonicity properties of the user optimal policy (user equilibrium) for a system with a single-server FCFS queue (Q_R) and a batch service queue (Q_T) under state-dependent routing in Chapters 2 and 4. We also studied existence, uniqueness and stability properties of the user equilibrium for the above system under probabilistic routing in Chapters 3 and 5. In Chapters 2, 4 and 5, we considered Q_R as a $\cdot|M|1$ queue while in Chapter 3 we considered it as a $\cdot|G|1$ queue, that is, for Chapter 3 the service time of Q_R is generally distributed. We modelled Q_T as a $\cdot|M^{(N)}|\infty$ queue, that is Q_T has infinitely many servers, in Chapters 2 and 3, and it was modelled as a $\cdot|M^{(N)}|1$ queue in Chapters 4 and 5. In Chapter 2 in the state-dependent case we have shown that a unique user optimal policy exists and that it possesses various monotonicity properties using coupling and stochastic comparison arguments for the system with a $\cdot|M|1$ queue and a

$\cdot|M^{(N)}|\infty$ queue. In Chapter 4 we have also shown that a unique user optimal policy exists and that it possesses some monotonicity properties using coupling arguments as we did in Chapter 2.

The monotonicity properties for the systems with a $\cdot|G^{(N)}|\infty$ queue and that with a $\cdot|M^{(N)}|1$ queue are different due to the possible number in the queue seen by a marked general commuter waiting for service in Q_T. Since the $\cdot|G^{(N)}|\infty$ queue has a plentiful supply of servers, service of a completed batch commences immediately when the batch is full and so the marked commuter can only see $0, 1, 2, \ldots, N-1$ commuters waiting for service in Q_T. For a $\cdot|M^{(N)}|1$ queue, since arriving general users not only have to wait for the completion of their own batch but may also have to wait for the completion of the services of the batches ahead of their own batch, that is, they have to wait until the server is free, arrivals to Q_T may see any number of commuters ahead of them including those currently in service at Q_T. We have shown that the unique user optimal policy is slice monotone for the system with a $\cdot|M^{(N)}|1$ queue in Chapter 4. Furthermore, we have shown that the user optimal policy changes monotonically with changes in the parameters, λ, λ_T, μ_R and μ_T for both systems. From the numerical examples in Chapter 2 the expected delay for all general users, W, is not monotone in μ_R and λ but in the examples it is monotone in λ_T and μ_T for the system with a $\cdot|M|1$ queue and a $\cdot|G^{(N)}|\infty$ queue — this needs further investigation for this system.

Routing the general arrival via Q_T, that is $\boldsymbol{x} \in T$ for $\boldsymbol{x} \in \Omega$, in those states where the expected delay via both routes is equal (i.e. $y(\boldsymbol{x}) = z(\boldsymbol{x})$) guarantees the uniqueness of the user optimal policy (user equilibrium) for these systems. The uniqueness of the user optimal policy for this state-dependent network is

interesting, since for the equivalent network with probabilistic routing, multiple user equilibria exist as seen in [2] and Chapter 5. The existence, uniqueness and monotonicity of the user optimal policy we discussed above still hold if we set $\boldsymbol{x} \in R$ instead of $\boldsymbol{x} \in T$ in those cases where $y(\boldsymbol{x}) = z_{D}(\boldsymbol{x})$, that is, to route via Q_{R} instead of Q_{T}. However, the user optimal policy will now differ from that obtained if the decision is to route via Q_{T}. This model is deterministic and unlike others studied in the literature where randomization is required (see for example, [4]). More generally, it would be possible to set a parameter p such that in those states where the delay via both routes is equal, general users choose Q_{R} with probability p and Q_{T} with probability $1 - p$. The case $p = 0$ is the one studied in Chapter 2 and 4. For this more general case, the user optimal policy exists, is unique and monotone. For $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues a slight generalization of Theorem 2.3.2 shows that the user optimal policy will be monotone in p. Thus any choice of randomization parameter p produces a unique decision policy, unlike the system studied in [27]. It is interesting that even when the batch-service queue has only a single server, randomization is not needed to guarantee uniqueness.

Under probabilistic routing multiple equilibria may exist for the system described above with a $\cdot|M|1$ queue and a $\cdot|G^{(N)}|\infty$ queue if $N = 2$ (see [2]). In Chapter 3 we consider a $\cdot|G|1$ queue, that is, generally distributed service time for Q_{R}, with Q_{T} as a $\cdot|G^{(N)}|\infty$ queue. Although the Pollaczek-Khinchin formula gives the expected transit time in Q_{R}, an exact analysis is more difficult. However, numerical examples there show that multiple user equilibria exist even for $N > 2$. Similarly, multiple user equilibria exist for some parameter values if we consider the same system with a $\cdot|M|1$ queue and a $\cdot|M^{(N)}|1$ queue instead of a $\cdot|G^{(N)}|\infty$ queue (see Chapter 5). Furthermore,

from the numerical examples in Chapter 5 the Downs-Thomson paradox is observed for the latter system for the parameters μ_R and λ.

The numerical examples from Chapter 2 for a $\cdot|M|1$ queue and a $\cdot|G^{(N)}|\infty$ queue illustrated some interesting features of these user optimal policies. We observed that the state-dependent routing mitigates the effects of the Downs-Thomson paradox observed under probabilistic routing. This observation is of practical interest, since it suggests that traffic networks where users have information about the current state of the network may sometimes perform better than networks where only information about average delay is available. Hassin [22, 23] and Hassin and Haviv [27] have observed that in some systems additional information may lead to *worse* performance rather than better. From our examples delays under state-dependent routing, although mitigating the worst of the Downs-Thomson effect, are sometimes slightly greater than that under probabilistic routing. This needs to be investigated further for this system, and for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|1$ queues.

It is interesting to compare this model in greater detail with the batch-service queues example discussed by Hassin and Haviv [26, 27]. Their model also consists of two queues. One of them is a batch-service queue (shuttle bus) with batch size $N = 7$ but the other queue, rather than being a single-server FCFS service queue, could be thought of as a single-server batch service queue, where all commuters waiting for service are served together no matter how many they are. The latter queue could be a reasonable model for a standard bus service, where buses arrive as a Poisson process and then commuters compare the time to wait until either a bus arrives or a shuttle bus is full. Under probabilistic routing, the authors defined p to be the probability

of choosing the shuttle bus and with $1-p$ the probability of choosing the standard bus. They observed that the two pure user equilibria (that is, $p=0$ and $p=1$) are ESS and the mixed user equilibrium, $p \in (0,1)$, is not ESS. Furthermore, the two pure user equilibria are also socially optimal. Under probabilistic routing we have shown that the pure stable equilibria and the unique stable mixed equilibrium are always ESS but the stable mixed equilibrium from multiple equilibria is not ESS. Under the state-dependent case, the authors observed that the user optimal policy is either for all arrivals to use the shuttle bus or none and for some arrival parameter values the user optimal policy and socially optimum policy do not coincide. In their model there is no stream of dedicated arrivals to the shuttle bus queue. If a stream of dedicated shuttle bus arrivals is added, as we did here, then, more generally, there will be a threshold (say C) at the shuttle bus queue, above which general users choose to join the shuttle bus queue rather than wait for the bus. In the terminology used here, if we now compare only the waiting times until service commences in the two queues then this would correspond to setting $y(\boldsymbol{x}) = 1/\mu_R$, where $1/\mu_R$ is the expected time until the next standard bus arrives. Thus the user optimal policy would still be monotone but with $R^\star = \{\boldsymbol{x} \in \Omega : n \leq C\}$, where C is the threshold above which general users are routed to the shuttle bus. In our model here in the state-dependent routing case, since $y(\boldsymbol{x})$ is increasing in a, even if the stream of dedicated train users is removed, that is, even if $\lambda_T = 0$, general users will choose to join Q_T if the number in Q_R grows sufficiently large. Thus, even with $\lambda_T = 0$ the decision policy where all users join Q_R is not a user optimal policy for our model.

The user optimal policy and the socially optimal policy were compared for $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues in the probabilistic routing case in [2] and Chapter

3. In summary, the value of p_R at which the socially optimal policy occurs is less than or equal to the value of p_R at which the unique stable user equilibrium occurs or the smallest stable user equilibrium of the multiple equilibria occurs. We have not compared the user optimal policy with the socially optimal policy for the system models we consider here in the state-dependent routing case, though, Su and Ziedins compare these policies for $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues in [63]. The numerical results in [63] suggest that the user equilibrium for this model may often be close to socially optimal.

In addition to the open questions discussed above some obvious extensions to the state-dependent model are as follows. The first is to modify the discipline for Q_T by allowing a batch to start service before it is completed. For example, let p_i be the probability that for a batch with i users already in it, the next arrival triggers a service commencement with $0 \leq p_1 \leq p_2 \leq, \ldots, \leq p_{N-1}$. The second is to add a third stream of Poisson arrivals at rate λ_R, say, of dedicated road users to the system. Then there will be a threshold level for Q_R, above which general users will no longer choose to join it but will join Q_T instead. If the network has a $\cdot|G^{(N)}|\infty$ queue then, provided that $\lambda_T < \mu_R$, the system will be stable, but now the set of recurrent states is infinite. Slightly modified versions of the proofs for the state-dependent routing case in Chapter 2 still carry through to show that the user optimal policy exists, is unique and monotone in structure. The results obtained in [2] for the probabilistic routing case also held for the case where a stream of dedicated road users stream is added. This additional arrival stream does not affect Q_T at all while the expected delay via Q_R, $W_R(p_R)$, becomes $1/(\mu_R - \lambda\, p_R - \lambda_R)$ instead of $\mu_R - \lambda\, p_R$ (see [2] for more details of this case when the network has a $\cdot|G^{(N)}|\infty$ queue).

Appendix A

Matlab and R functions

A.1 Matlab code

1. **Function for calculating the user optimal policy (uop) for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues.**

```
% function for computing the user-optimal policy
function[y,z,astar,Ra,Rn] = uop(lambda,lambdat,mur,mut,N)
 % define theta1 and theta2 since they use many times here
theta1 = lambda+lambdat;
theta2 = theta1+mur;
% aa = prob. of jump fr. a = 0 when lambda arrival occurs
aa = lambda/theta1;
% bb = prob. of jump fr. a = 0 when lambdat arrival occurs
bb = lambdat/theta1;
%c =  prob. of jump fr. a>0 when service muR occurs
c = mur/theta2;
```

```
% d = prob. of jump when lambda arrival occurs
d = lambda/theta2;
% e = prob. of jump when lambdat arrival occurs
e = lambdat/theta2;
% f = prob. of time spend in the state (0,n) b4 next jump
f = 1/theta1;
 % g = prob. of time spend in the state (a,n) b4 next jump
g = 1/theta2;
% h = prob. of jump from (a,n) to (a,n+1) if (a,n) in T
h = d+e;
% calculate the max a
asta = ceil(mur*(1/mut+(N-1)/lambdat)-1)+1;
z1 = zeros(asta,N);
astar1 = ones(1,N);
% calculate the Dt-marker on N-1 row
 % add 1 to avoid reference to index 0
astar1(N) = ceil(mur/mut - 1)+1;
z1( :,N) = 1/mut; % z(a,N-1) = 1/mut
y = (1:asta)/mur; % y(a,n) = (a+1)/mur
% calculate the Ra and Rn for user-optimal policy
for n = (N-1):-1:1
   an1 = astar1(n+1);
    if an1 ==1
        z1(1,n) = f + z1(1,n+1);
        for a = 2:asta
            z1(a,n) = g + h*z1(a,n+1) + c*z1(a-1,n);
```

```
    end;
    temp = 1;
    while y(temp) < z1(temp,n)
        temp = temp + 1;
    end
    astar1(n) = temp;
else
    A = eye(an1) - diag(ones(1,an1-1)*c, -1) -
         diag(ones(1,an1-1)*d,1);
    A(1,2) = -aa;
    B = eye(an1)*e;
    B(1,1) = bb;
    B(an1,an1) = h;
    C = ones(an1,1)*g;
    C(1) = f;
    zn1 = z1(1:an1,n+1);
    zn = A\(B*zn1+C); % A\B = A*inv(B)
    z1(1:an1,n) = zn;
    for a = (an1+1):asta
        z1(a,n) = g + h*z1(a,n+1) + c*z1(a-1,n);
    end;
    temp = 1;
    while y(temp) < z1(temp,n)
        temp = temp + 1;
    end
    astar1(n) = temp;
```

```
    end
end
Ra = [ ];
Rn = [ ];
% Ra and Rn
for n = 1: N
    for a = 1:(astar1(n)-1)
        Ra = [Ra (a-1)];
        Rn = [Rn (n-1)];
    end
end
z = flipud(z1');
astar =  flipud(astar1')-1;
```

2. **Function for calculating the expected delay W (w) for the system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queues.**

```
% function for W with any policy
function[End, y, zzz, ppp, W] = w(lambda,lambdat,mur,mut,N,DRa,
            DRn,forceEnd,Endin)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% For a network with parameters lambda, lambdat,mur,mut,N
% and decision policy specified by DRa and DRn,
% calculate the persistent states via End (if forceEnd == 0)
% End = a*(0)
% the steady state distribution p
% the long-run average origin-destination time W
% y(a,n) = the expected transit time via Q_R for an arrival from
```

```
% the general users stream who sees the system in state (a,n)
% z(a,n) = the expected transit time via Q_T for an arrival from
% the general users stream who sees the system in state (a,n)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
sDRa= unique(DRa);
if forceEnd == 0
    End = 0;
    while (End < length(sDRa)) & (sDRa(End+1)==End)
        End = End+1;
    end
    % persistent states SD = {0,1,...,End} x {0,1,...,N-1}
else
    End = Endin;
end
    %otherwise {0,1,...,End}x{0,1,...,N-1}
    End;
    d1 = lambda + lambdat;
    %t1 = lambda+lambdat;
    block1 = -diag(ones(1,N)*d1) + diag(ones(1,N-1)*d1,1);
    block1(N,1) = d1;
    d2 = d1+mur;
    %theta2 = theta1+mur;
    block2 = -diag(ones(1,N)*d2) + diag(ones(1,N-1)*d1,1);
    block2(N,1) = d1;
    % Q matrix for DL = []
    Q = kron(diag(ones(1,End+1)),block2);
```

```
Q(1:N,1:N) = block1;
Q = Q + diag(ones(1,End*N)*mur, -N);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Q matrix modified for current policy (via DRa & DRn)
% rows converts policy states (a,n) to the
%correct row index in Q
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
rows = N*DRa + DRn+1;
for i = 1:length(DRa)
    if DRa(i) < End
        if DRn(i) < N-1
            Q(rows(i),rows(i)+1) = lambdat;
            Q(rows(i),rows(i)+N) = lambda;
        else
            Q(rows(i),rows(i)-N+1) = lambdat;
            Q(rows(i),rows(i)+N) = lambda;
        end
    end
end
Q( :,N*(End+1)) = 1;
rhs = zeros(1,N*(End+1));
rhs(N*(End+1))=1;
p = rhs/Q; % equiv to rhs*inv(Q)
% create M matrix & rhs ( assuming all states in DT)
block1 = diag(ones(1,N-1)) - diag(ones(1,N-2),1);
rhs1 = 1/(lambda+lambdat);
```

```
d2 = (lambda + lambdat)/(lambda+lambdat+mur);
block2 = diag(ones(1,N-1)) - diag(ones(1,N-2)*d2,1);
rhs2 = 1/(lambda+lambdat+mur);
rhs3 = rhs2 + d2/mut;
M = kron(diag(ones(1,End+1)),block2);
M(1:(N-1),1:(N-1)) = block1;
d3 = mur/(lambda+lambdat+mur);
M = M - diag(ones(1,End*(N-1))*d3,-N+1);
rhs = ones((N-1)*(End+1),1)*rhs2;
if N > 2
    rhs(1:(N-2)) = rhs1;
end
rhs(N-1) = rhs1 + 1/mut;
for i = 1:End
    rhs((N-1)*(i+1)) = rhs3;
end
M;
rhs;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% now alter the M matrix & rhs for states in R
% we only reqire the z(a,n)'s for (a,n) in SD
% also state (a,n) in R affects the "z(a,n-1) equations"
% for 1<=n<=N-1
% also states (a,0) in R do not alter any z equation
% we weed out of the R list ( in DRa & DRn) states (a,n) that
% are not in SD or having n = 0.
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
DRaz = [ ]; % a co-ords of all (a,n) in R with n <=N-2
DRnz = [ ]; % corresponding n co-ords
for i = 1:length(DRa)
    if (DRa(i) < End) & (DRn(i)> 0)
        DRaz = [DRaz DRa(i)];
        DRnz = [DRnz DRn(i)];
    end
end
DRaz;
DRnz;
rows = (N-1)*DRaz + DRnz;
for i = 1:length(DRaz)
  if DRaz(i) == 0
     % check whether bracket is needed here
     if (DRnz(i) < (N-1))
       M(rows(i),rows(i)+1) = -lambdat/(lambda+lambdat);
       M(rows(i),rows(i)+N-1) = -lambda/(lambda+lambdat);
     else
       rhs(rows(i)) = rhs(rows(i)) - lambda/((lambda +
              lambdat)*mut);
       M(rows(i),rows(i)+N-1)= -lambda/(lambda+lambdat);
        end
    else
     if (DRnz(i) < N-1)
       M(rows(i),rows(i)+1) = -lambdat/(lambda+lambdat+mur);
```

```
        M(rows(i),rows(i)+N-1) = -lambda/(lambda+lambdat+mur);
      else
        rhs(rows(i)) = rhs(rows(i)) -
              lambda/((lambda+lambdat+mur)*mut);
        M(rows(i),rows(i)+N-1)=-lambda/(lambda+lambdat+mur);
         end
    end
end
M;
rhs;
z = M\rhs;
zz = zeros(1,N*(End+1));
onemut = 1/mut;
for i = 1:(End+1)
  zz((N*(i-1)+1): (N*i-1))= z(((N-1)*(i-1)+1) : ((N-1)*i));
  zz(N*i) = onemut;
end
zz;
zzz = flipud(reshape(zz',N,End+1));
ppp = flipud(reshape(p',N,End+1));
a = 0:End;
y=(a+1)/mur;
rows = N*DRa+DRn+1;
for i = 1:length(DRa)
    if DRa(i)<End
        zz(rows(i)) = (DRa(i)+1)/mur;
```

```
        end
     end
    % End = End +1;
     W=sum(p.*zz);
```

3. **Function for calculating user equilibrium under probabilistic case for a system with $\cdot|M|1$ and $\cdot|G^{(N)}|\infty$ queue.**

```
% function for calculating the user equilibrium values
%    for prob. case
%specific with parameter values when c^2 = 1 that is,
%    M|M|1 queue
function[p] = ueq(lambda,lambdat,mur, mut,N)
  gamma = (lambda + lambdat + 1/2*(N-1)*mut)*mur
        - (lambda + lambdat)*mut;
  pmin = (lambda + mur + lambdat + 1/2 * (N-3)*mut)/2;
  mu0 = mut*((lambda+lambdat)/(lambda + lambdat
       + 1/2*(N-1)*mut));
  mu1 = lambda + (lambdat * mut)/(lambdat + 1/2 * (N-1)*mut);
            if mur <= mu0
                p = 0;
            elseif mur >= mu1
                p = 1;
            elseif (mur > mu0 && mur < mu1)
                p = (pmin - sqrt(pmin^2 - gamma))/(lambda);
            end
%%%%%%%%%%%%% end of ueq func. %%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

4. **Matlab file for plotting W against μ_R under both state-dependent case and probabilistic case.**

   Note that similar matlab files were use to produce the plots of W versus λ and W versus λ_T.

```
% plot W for state-dep & prob routing on the same axes when
% mur changes and other parameter values
% (lambda, lambdat, mut, N) are fixed
lambda = 1; lambdat = 0.1; mut = 1; N= 3;
start = 0.01;
step = 0.001;
finish = 3;
% ww = vector of expected transit times, W,
%   for state-dependent case
% pProb = vector of equilibrium for prob. case
%wr and wt are W_R and W_T for prob. case
% wwProb = vector of W for prob. case
% pval = a equilibrium for given parameter values
%     in prob. case
 ww = []; pProb = []; wr = []; wt = [];
 wwProb = []; pval = [];
for mur = start:step:finish % mur = [0.01, 3]
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  %%%%%%%%%%% for stateDep routing %%%%%%%%%%%
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  % uop function spit out the R =  {(a,n)} for
  % an user optimal policy D, values of y(a,n) and z_D(a,n)
```

```
  % and astar for each n
  [y z astar Ra Rn] = uop(lambda,lambdat, mur, mut,N);
  forceEnd = 0; Endin = 5;
  % w function gives the values of astar(0), y(a,n),
  %  z(a,n), pi(a,n) and W for
  [astar0,y,z,pii,W] = w(lambda,lambdat,mur, mut, N, Ra, Rn,
       forceEnd, Endin);
  ww = [ww, W];
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  %%%%%%%%% probabilistic routing %%%%%%%%%%%%%
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  pProb = [pProb,ueq(lambda,lambdat,mur,mut,N)];
  pval = ueq(lambda,lambdat,mur,mut,N);
  if pval == 0
    wwProb = [wwProb, 1/mut + (N-1)/(2*(lambdat + lambda))];
  elseif pval == 1
    wwProb = [wwProb,1/(mur - lambda)];
  else
    wr = 1/(mur - lambda*pval);
    wt = 1/mut + (N-1)/(2*(lambdat + (1 - pval)*lambda));
    wwProb = [wwProb,pval*wr + (1- pval)*wt];
  end
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% set up the plot                                          %
% the legends for this plot is written using               %
```

```
% the legend function on the figure window            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
mur = start:step:finish;
hold on % use when draw > 1 plot on the same axes
plot_one = plot(mur, wwProb,':k'); % :k  = black dotted line
xlabel('\mu_{R}');
ylabel('W');
title('W vs \mu_{R}');
plot_two = plot(mur, ww, '-k'); %-k = black hard line
hold off % end the hold function
```

5. **Matlab file for plotting W against λ under both state-dependent case and probabilistic case**

```
% produce the plot of W for state-dependent case and prob. case
% on the same axis when lambda varies but the other parameters
% are fixed.
lambdat = 0.1; mur = 2; mut = 2; N = 3;
start = .01;
step = 0.01;
finish = 6;
% ww = vector of W for state-dep. case
% wwProb = vec. of W for prob. case
% pProb = vec. of equilibria for prob. case
ww = []; pProb = []; wwProb = [];
for lambda = start:step:finish %lambda = [0.01, 6]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% state-dependent case %%%%%%%%%%%%%%%%%%%
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 [y z astar Ra Rb] = uop(lambda,lambdat, mur, mut,N);
 forceEnd = 0; Endin = 10;
 [astar0,y,z,pii,W] = w(lambda,lambdat, mur, mut, N, Ra, Rb,
     forceEnd, Endin);
  ww = [ww W];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%% probabilistic routing %%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  pProb = [pProb,ueq(lambda,lambdat,mur,mut,N)];
    pval = ueq(lambda,lambdat,mur,mut,N);
    if pval == 0
        wwProb = [wwProb, 1/mut + (N-1)/
            (2*(lambdat + lambda))];
    elseif pval == 1
        wwProb = [wwProb,1/(mur - lambda)];
    else
    wr = 1/(mur - lambda*pval);
    wt = 1/mut + (N-1)/(2*(lambdat + (1 - pval)*lambda));
    wwProb = [wwProb,pval*wr + (1- pval)*wt];
    end
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% set up the plot of W for prob & stateDep            %
% the same axes,                                      %
% the legends for this plot is written using          %
```

```
% the legend function on the figure window            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lambda = start:step:finish;
max_x = finish;
max_y = max(wwProb)+ 1;
hold on;
axis([0 max_x 0 max_y]);
plot(lambda, wwProb, ':k')
plot(lambda, ww, '-k')
xlabel('\lambda');
ylabel('W');
title('W vs \lambda');
hold off;
```

6. **Matlab file for plotting W against λ_T under both state-dependent case and probabilistic case.**

```
% produce the plot of W for state-dependent case
% and prob. case
% on the same axis when lambda_T varies but the
% other parameters
% are fixed.
lambda = 1; mur = 1.2; mut = 1; N = 3;
start = 0.01;
step = 0.001;
finish = 5;
% ww = vector of W for state-dep. case
% wwProb = vec. of W for prob. case
```

```
% pProb = vec. of equilibria for prob. case
ww = []; pProb = []; wwProb = [];
for lambdat = start:step:finish % lambda_T = [0.01, 5]
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 %%%%%%%%%% state-dependent case %%%%%%%%%%%%%%%%%%%
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 [y z astar Ra Rb] = uop(lambda,lambdat, mur, mut,N);
 forceEnd = 0; Endin = 6;
 [astar0,y,z,pii,W] = w(lambda,lambdat,mur,mut, N,Ra,Rb,
       forceEnd, Endin);
  ww = [ww W];
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  %%%%%%%%% probabilistic routing %%%%%%%%%%%%%%%
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  pProb = [pProb,ueq(lambda,lambdat,mur,mut,N)];
    pval = ueq(lambda,lambdat,mur,mut,N);
    if pval == 0
        wwProb = [wwProb, 1/mut + (N-1)/
                (2*(lambdat + lambda))];
    elseif pval == 1
        wwProb = [wwProb,1/(mur - lambda)];
    else
    wr = 1/(mur - lambda*pval);
    wt = 1/mut + (N-1)/(2*(lambdat + (1 - pval)*lambda));
    wwProb = [wwProb,pval*wr + (1- pval)*wt];
    end
```

```
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% set up the plot of W for prob & stateDep            %
% the same axes,                                      %
% the legends for this plot is written using          %
% the legend function on the figure window            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lambdat = start:step:finish;
max_x = finish;
max_y = max(wwProb)+ 0.5;
hold on;
axis([0 max_x 0 max_y]);
plot(lambdat, wwProb, ':k')
plot(lambdat, ww, '-k')
xlabel('\lambda_T');
ylabel('W');
title('W vs \lambda_T');
hold off;
```

A.2 R code for Chapters 3 and 5

1. R code for an $M|G|1$ queue and $M|M|1$ queue

```
############# W_R for M|G|1 queue #########################
wrG.func <- function(lambda = 1, mur = 1, c = 0, pr = 0){  #
  1/mur + (lambda*pr*(1+c^2))/(2*mur*(mur - lambda*pr))    #
```

```
}                                                                    #
#####################################################################
#################### W_R for M|M|1 model ############################
# wr.func calculate the expected transit time via Q_R, W_R,#
# with the parameters arrival rate, lambda, service rate   #
# at Q_R, mur, and prob. of choosing the Q_R, pr.          #
wr.func <- function( mur = 1.3, lambda = 1, pr = 0)        #
  1/(mur - lambda * pr)                                    #
#####################################################################
```

2. R code for an $M|G^{(N)}|\infty$ queue

```
#####################################################################
#         For Q_T with infinte many servers                #
# wt.infinite.func calculate the expected transit time via #
# Q_T, W_T, with the parameters arrival rate, lambda,      #
# service rate at Q_T,mut, committed Q_T arrival rate,     #
# lambdat and prob. of choosing the Q_T, (1 - pr)          #
wt.infinite.func <- function(lambda = 1, lambdat = 0.1,
                             mut = 3, pr = 0.1,N = 2)        #
  (1/(lambdat + lambda*(1- pr))) * ((N-1)/2 ) + 1/mut      #
#####################################################################
```

3. R code for searching value of p_R that minimizes the expected transit time W and the user equilibria p_R^{EQ}.

```
#####################################################################
# w.phatsoc.func search for the value of p that minimize W #
```

```
#  and the user equilibria.                                    #
w.phatsoc.func <- function(lambda = 1, mur = 1, c = 1,        #
                        lambdat = 0.1, mut = 3,N = 2,peq.u    #
                           = 0.5, peq.l = 0.5, l.end = 0,     #
                           u.end = 1, pr = 0){                #
          wr <- wrG.func(lambda = lambda, mur = mur, c = c,   #
                       pr = pr)                               #
          wt <- wt.infinity.func(lambda = lambda, lambdat =   #
                              lambdat, mut = mut, N = N,      #
                              pr = pr)                        #
          wr0 <- wrG.func(lambda = lambda, mur = mur, c = c,  #
                        pr = 0)                               #
          wt0 <- wt.infinity.func(lambda = lambda, lambdat =  #
                           lambdat, mut = mut,N = N, pr =0)   #
           wr1 <- wrG.func(lambda = lambda, mur = mur, c = c,#
                         pr = 1)                              #
          wt1 <- wt.infinity.func(lambda = lambda, lambdat =  #
                            lambdat, mut = mut,N = N,pr=1)    #
          #beta.i, i = 1,. ..6 are coefficients of p in WT-WR#
          beta1 <- mur + 0.5* (c^2 -1)*mut                    #
          beta2 <- mur^2 + (1 + lambdat + 0.5*(N-3)*mut)*mur+#
                            0.5*(1+ lambdat)*(c^2 - 1)*mut    #
          beta3 <- (1 + lambdat + 0.5*(N-1)*mut)*mur^2 -(1+   #
                                          lambdat)*mut*mur#
          beta4 <- mur^2 * mut * (1+lambdat)                  #
          beta5 <- (1+lambdat  + mur)*mur*mut                 #
```

```
beta6 <- mur*mut                                          #
##              w.eq = WT - WR                            #
w.eq <- wt - wr                                           #
p.end <- c()                                              #
w.eq.pos <- w.eq[w.eq > 0]                                #
w.eq.neg <- w.eq[w.eq < 0]                                #
p.star1 <- c()                                            #
p.star2 <- c()                                            #
if (length(w.eq.pos) > 0 & length(w.eq.neg) == 0){ #
 p.star1 <- 1}                                            #
else {                                                    #
  if (length(w.eq.pos) == 0 & length(w.eq.neg) > 0)#
      p.star1 <- 0                                        #
  else {                                                  #
    if(length(w.eq.pos) > 0 & length(w.eq.neg)> 0){#
        pEQ.func <- function(p)                           #
          (beta1*p^2 - beta2*p + beta3)/(beta4 -   #
                                beta5*p + beta6*p^2)#
# search for 2nd user mixed user equilibrium       #
#  w.eq[1] = WT(0)  - WR(0)                                 #
        up.test <- c()                                    #
       lower.test <- c()                                 #
        for (i in 2:(length(w.eq)-1)){                    #
          if (w.eq[i] == min(w.eq))                       #
            up.test <- pr[i]+ 0.02                        #
          if (w.eq[i] == max(w.eq))                       #
```

```
            lower.test <- pr[i]- 0.02                #
          }
          if (min(w.eq)< w.eq[1] & min(w.eq)<        #
                          w.eq[length(pr)]){         #

           p.star1 <- uniroot(pEQ.func, c(0,up.test),#
                              tol = 10 ^(-30))$root #
           p.star2 <- "WT - WR has a min. val"       #
         }
          else {                                     #
          #search for 1st user mixed user equilibrium#
           if (max(w.eq) > w.eq[1] & max(w.eq) >     #
                  w.eq[length(pr)]){                 #
           p.star1<-uniroot(pEQ.func,c(lower.test,1),#
                             tol = 10 ^(-30))$root  #
           p.star2 <- "WT - WR has a max. val"       #
         } } }                                       #
  } #end of 2nd else                                 #
# end of 1st else statements                         #
#alpha.i, i = 1,2,..,4 are coefficients of p for W #
alpha.1 <- mur - (1-c^2)*mut/2                       #
alpha.2 <- (mur^2 + (lambdat + 2+(N-3)*mut/2)*mur- #
                        (1+ lambdat)*(1-c^2)*mut/2)#
alpha.3<-((lambdat + 2 + (N-1)*mut/2)*mur^2 + (1 + #
          lambdat +(N-3)*mut/2 - lambdat * mut)*mur)#
alpha.4 <- (1 + lambdat + (N-1)*mut/2)*mur^2         #
```

```
#w.soc = W = p*WR + (1-p)*WT                                    #
# if p = 0, W = WT(0) and if p = 1, W = WR(1)                   #
w.soc <- pr * wr + (1 - pr)*wt                                  #
p <- c()                                                        #
   if (wt0 <= min(w.soc) & wt0 <= max(w.soc)){                  #
     p <- 0                                                     #
   }
   else {                                                       #
        if(wr1 <= min(w.soc) & wr1 <= max(w.soc)){              #
     p <- 1                                                     #
   }
     else {                                                     #
        if(wt0 > min(w.soc) & min(w.soc) < wr1){                #
# p.func will be use to search for value of p that             #
# make W' = 0                                                   #
      p.func <- function(p)                                     #
        (-alpha.1 * p^4 + 2*alpha.1*(1 + lambdat + #
        mur) * p^3 - (3*(1 + lambdat)*mur*alpha.1 +#
        (1 + lambdat + mur)*alpha.2 - alpha.3)*p^2+#
        2*((1 + lambdat)*mur*alpha.2 - alpha.4)*p +#
        (1+ lambdat + mur)*alpha.4 - (1 + lambdat  #
                                    )*mur*alpha.3#
## search for the root of W within [0,1] using                  #
# function uniroot                                              #
     p <-  uniroot(p.func , c(l.end, u.end),tol =  #
                   10 ^(-20))$root                              #
```

```
                } # end of last if                              #
              }# end of the last if statement                   #
            } ## end of 2nd if and else statement               #
          # the following if statements use to get the social#
          #        optimum p.hat                                #
           if (length(p.star2) > 0)
            cat("p.star = ",p.star1 , " , ", p.star2, ",      #
                  p.hat = ", p,'\n')                           #
          else                                                 #
            cat("p.star = ",p.star1 ,  " p.hat = ", p,'\n')   #
          list(pstar =  p.star1, phat = p)                     #
}##### end of soc.diff.func                                   #
#################################################################
```

4. R code for an $M|M^{(N)}|1$ queue.

```
#################################################################
#          For Q_T with single server                           #
# wt.one.func calculate WT for system with single server       #
# batch service queue. If wtout = 1 wt.one.func spits out      #
# WT and if wtout = 0 and spits out the value of r.            #
wt.one.func <- function(lambda = 1, lambdat = 0.1, mut = 3,#
                        pr = 0.1, N = 3, wtout = 1){         #
  # vectors wt.vec and r.vec collect the values of WT and     #
  # r for thecorresponding parameters.                        #
    r.vec <- c(); wt.vec <- c();                             #
  #  arrival2qt (= arrival to QT)                            #
    arrival2qt <- lambda *(1-pr) + lambdat                   #
```

```
    for (i in 1:length(pr)){                                    #
  #   r.f = func we use 2 search for r value that satisfy       #
  #   the arrival2qt/mut = r(1-r^N)/(1-r)                       #
    r.f <- function(r)                                          #
     arrival2qt[i] - (arrival2qt[i] + mut)*r + mut*r^(N+1)      #
   rho <- arrival2qt[i]/(N*mut)                                 #
   if (rho >= 1) stop("rho greater than 1 \n ")                 #
    r <-  uniroot(r.f , c(rho , rho ^(2/(N+1))),tol =           #
           10^(-20))$root                                       #
   wt <-(1/arrival2qt[i])*((N-1)/2+r ^(N+1)/(1-r ))+1/mut       #
   wt.vec <- c(wt.vec,wt)                                       #
   r.vec <- c(r.vec,r)                                          #
 } # end of for loop                                           #
 if (wtout == 0)                                                #
   r.vec                                                        #
 else{                                                          #
  if (wtout == 1)                                               #
 wt.vec }                                                       #
} ## end of function wt.one.func                               #
################################################################
```

5. R code for searching the user equilibria for W when Q_T is an $M|M^{(N)}|1$ queue.

```
################################################################
# peq.one.func search for the user equilibrium for given       #
# parameter values. If all values of w.diff = WT - WR >= 0     #
# stable eq. = 1, if all values of w.diff <= 0                 #
```

```
# stable eq. = 0  o.w. stable eq. in (0,1)                    #
peq.one.func <-function(lambda = 1,lambdat = 0.1,mur = 1.2,#
                mut = 3, N = 2, pr = 0 ){                     #
  peq <- c(); wr.pos <- c()                                   #
  wr.val <- wr.func(lambda = lambda, mur = mur, pr = pr)      #
  wr.pos <- wr.val[wr.val >=0]                                #
  pr.pos <- pr[wr.val >= 0]                                   #
  wt.val <- wt.one.func(lambda = lambda, lambdat = lambdat,#
                mut = mut,N= N,pr = pr.pos)                   #
  w.diff <- wt.val - wr.pos                                   #
  if (all(w.diff >= 0))                                       #
      peq <- c(peq,1)                                         #
  else{                                                       #
      if (all (w.diff <= 0))                                  #
         peq <- c(peq, 0 )                                    #
      else {                                                  #
        for (i in 1:(length(pr.pos)-1)){                      #
          if (w.diff[i] > 0 & w.diff[i+1] < 0){               #
            temp.p <- c(pr.pos[i],pr.pos[i+1])                #
            temp.diff <-  c(abs(w.diff[i]),abs(w.diff[i+1]))
     peq <- c(peq,temp.p[temp.diff == min(temp.diff)])
           }
          else{                                               #
          if (w.diff[i] < 0 & w.diff[i+1] > 0){               #
            temp.p <- c(pr.pos[i],pr.pos[i+1])
            temp.diff <-  c(abs(w.diff[i]),abs(w.diff[i+1]))#
```

```
      peq <- c(peq,temp.p[temp.diff == min(temp.diff)])
            } }                                                    #
        }
      }                                                            #
     }
   peq                                                             #
}# end of peq.one.func                                            #
#################################################################
```

Bibliography

[1] Abraham, J. E. and Hunt, J. D. (2001) Transit system management, equilibrium mode split and the Downs-Thompson paradox. Technical report. Department of Civil Engineering, University of Calgary.

[2] Afimeimounga, H., Solomon, W. and Ziedins, I. (2005) The Downs-Thomson paradox: Existence, uniqueness and stability of user equilibria. *Queueing Systems* **49**, 321-334.

[3] Afimeimounga, H., Solomon, W. and Ziedins, I. (2010) User equilibria for a parallel queueing system with state dependent routing. *Queueing Systems* **66**, 169-193.

[4] Altman, E. and Shimkin, N. (1998) Individual equilibrium and learning in processor sharing. *Operations Research* **46**, 776-784.

[5] Armony, M., Shimkin, N. and Whitt, W. (2009) The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**, 66-81.

[6] Arnott, R. and Small, K. (1997) The economics of traffic congestion. *American Scientist* **82**, 446-455.

[7] Arnott, R. and Yan, A. (2000) The two-mode problem: Second-best pricing and capacity. *Review of Urban and Regional Development Studies* **12**, 170-199.

[8] Bean, N. G., Kelly, F. P. and Taylor, P. G. (1997) Braess's paradox in a loss network. *Journal of Applied Probability* **34**, 155-159.

[9] Bell, C. E. and Stidham, S. (1983) Individual versus social optimization in the allocation of customers to alternative servers. *Management Science* **29**, 831-839.

[10] Ben-Shahar, I., Orda, A. and Shimkin, N. (2000) Dynamic service sharing heterogeneous preferences. *Queueing Systems* **35**, 83-103.

[11] Borst, S. C., Jonckheere, M. and Leskelä, L. S. (2008) Stability of parallel queueing systems with coupled service rates. *Discrete Event Dynamic Systems* **18**, 447-472.

[12] Brady, M. E. (1993) Dynamic stability, traffic equilibrium and the law of peak-hour expressway congestion. *Transportation Research* **27B**, 229-236.

[13] Braess, D., Nagurney, A. and Wakolbinger, T. (2005) On a paradox of traffic planning. *Transportation Science* **39**, 446-450.

[14] Brooms A. C. (2005) On the Nash equilibria for the FCFS queueing system with load-increasing service rate. *Advances in Applied Probability* **37**, 461-481.

[15] Calvert, B. (1997) The Downs-Thomson effect in a Markov process. *Probability in the Engineering and Informational Sciences* **11**, 327-340.

[16] Calvert, B., Solomon, W. and Ziedins, I. (1997) Braess's paradox in a queueing network with state-dependent routing. *Journal of Applied Probability* **34**, 134-154.

[17] Cohen, J. E. and Kelly, F. P. (1990) A paradox of congestion in a queueing network. *Journal of Applied Probability* **27**, 730-734.

[18] Correa, J. R., Schulz, A. S. and Stier-Moses, N. E. (2004) Selfish routing in capacitated networks. *Mathematics of Operations Research* **29**, 961-976.

[19] Downs, A. (1962). The law of peak-hour expressway congestion. *Traffic Quarterly* **16**, 393-409.

[20] Guo, P. and Zipkin, P. (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* **53**, 962-970.

[21] Hajek, B. (1984) Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control* **29**, 491-499.

[22] Hassin, R. (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**, 1185-1195.

[23] Hassin, R. (2007) Information and uncertainty in a queuing system. *Probability in the Engineering and Informational Sciences* **21**, 361-380.

[24] Hassin, R. (2009) Equilibrium customers' choice between FCFS and random servers. *Queueing Systems* **62**, 243-254.

[25] Hassin, R. and Haviv, M. (1995) Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* **45**, 966-973.

[26] Hassin, R. and Haviv, M. (2002) Nash equilibrium and subgame perfection in observable queues. *Annals of Operations Research* **113**, 15-26.

[27] Hassin, R. and Haviv, M. (2003) *To Queue or not to Queue.* Kluwer.

[28] Haviv, M. and Kerner, Y. (2007) On balking from an empty queue. *Queueing Systems* **55**, 239-249.

[29] Ho, B. (2003) Existence, uniqueness and monotonicity of the state-dependent user optimal policy for a simple Markov transport network. MSc Thesis, The University of Auckland.

[30] Holden, D. J. (1989) Wardrop's third principle: Urban traffic congestion and traffic policy. *Journal of Transport Economics and Policy* **23**, 239-262.

[31] Horowitz, J. L. (1984) The stability of stochastic equilibrium in a two-link equilibrium in transportation network. *Transportation Research* **18B**, 245-252.

[32] Koole, G. (1998) Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems* **30**, 323-339.

[33] Kraus, M. (2003) A new look at the two-mode problem. *Journal of Urban Economics* **53**, 511-530.

[34] Kumar, P. R. and Walrand, J. (1985) Individually optimal routing in parallel systems. *Journal of Applied Probability* **22**, 989-995.

[35] Last, G. and Szekli, R. (1998) Stochastic comparison of repairable systems by coupling. *Journal of Applied Probability* **35**, 348-370.

[36] Lee, H. W., Lee, S. S. and Chae, K. C. (1996) A fixed size batch service queue with vacations. *Journal of Applied Mathematics and Stochastic Analysis* **9**, 205-219.

[37] Lee, S., Lee, S. and Lee, Y. (2006) Innovative public transport oriented policies in Seoul. *Transportation* **33**, 189-204.

[38] Lee, S., Ryu, S. and Lee, S. (2010) Bus rapid transit operation analysis using the Downs-Thomson paradox. *Journal of Advanced Transportation* **44**, 205-215.

[39] Lin, W. and Lo, H. K. (2009) Investigating Braess's paradox with time-dependent queues. *Transportation Science* **43**, 117-126.

[40] Lindvall, T. (1992) *Lectures on the Coupling Method*. Wiley.

[41] Maynard-Smith, J. (1982) *Evolution and Game Theory.* Cambridge University Press.

[42] Medhi, J. (1975) Waiting time distribution in a Poisson queue with a general bulk service rule. *Management Science* **21**, 777-782.

[43] Mehdi, J. (2003) *Stochastic Models in Queueing Theory.* Elsevier Science.

[44] Mogridge, M. J. H. (1990) Planning for optimum urban efficiency: The relationship between congestion on the roads and public transport. *Transportation Planning Systems* **1**, 11-19.

[45] Mogridge, M. J. H. (1997) The self-defeating nature of urban road capacity policy. *Transportation Policy* **4**, 5-23.

[46] Murugan, S. P. B. and Kalyanaraman, R. (2009) A vacation queue with additional optional service in batches. *Applied Mathematical Sciences* **3**, 1203-1208.

[47] Naor, P. (1969) The regulation of queue size by levying tolls. *Econometrica* **37**, 15-24.

[48] Neuts, M. F. (1967) A general class of bulk queues with Poisson input. *The Annals of Mathematical Statistics* **38**, 759-770.

[49] Park, H., Kim, T. and Chae, K. C. (2010) Analysis of a two-phase queueing system with a fixed-size batch policy. *European Journal of Operational Research* **206**, 118-122.

[50] Parlaktürk, A. K. and Kumar, S. (2004) Self-interested routing in queueing networks. *Management Science* **50**, 949-966.

[51] Raj, C. B. and Manoharan, M. (1998) On the waiting time distribution of a single server and batch service $M|M|1$ queue. *Information and Management Sciences* **9**, 59-65.

[52] Ross, S. M. (1996) *Stochastic Processes.* Wiley.

[53] Roughgarden, T. (2002) How bad is selfish routing? *Journal of the ACM* **49**, 236-259.

[54] Roughgarden, T. (2006) On the severity of Braess's paradox: Designing networks for selfish users is hard. *Journal of Computer and System Sciences* **72**, 922-953.

[55] Shaked, M. and Shanthikumar, J. G. (1994) *Stochastic Orders and Their Applications.* Academic.

[56] Sharp, C. H. (1967) The choice between cars and buses on urban roads. *Journal of Transportation Economics and Policy* **1**, 104-111.

[57] Sheu, R. S. and Ziedins, I. (2010) Asymptotically optimal control of parallel tandem queues with loss. *Queueing Systems* **65**, 211-227.

[58] Smith, M. J. (1979) The existence, uniqueness and stability of traffic equilibria. *Transportation Research* **13B**, 295-304.

[59] Smith, M. J. (1984) The stability of a dynamic model of traffic assignment-an application of a method of Lyapunov. *Transportation Science* **18**, 245-252.

[60] Sparaggis, P. D. and Cassandras, C. G. (1992) Monotonicity properties of cost functions in queueing networks. *Discrete Event Dynamic Systems: Theory and Applications* **1**, 315-327.

[61] Spicer, S. and Ziedins, I. (2006) User optimal state-dependent routeing in parallel tandem queues with loss. *Journal of Applied Probability* **43**, 274-281.

[62] Stidham, S. and Weber, R. (1993) A survey of Markov decision models for control of networks of queues. *Queueing Systems* **13**, 291-314.

[63] Su, K. and Ziedins, I. (2010) Optimal controls for queueing networks. In preparation.

[64] Thomson, J. M. (1977) *Great Cities and Their Traffic.* Gollancz.

[65] Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration.* Springer.

[66] Veeraraghavan, S. and Debo, L. (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing and Service Operations Management* **11**, 543-562.

[67] Walting, D. (1996) Asymmetric problems and stochastic process models of traffic assignments. *Transportation Research* **30B**, 339-357.

[68] Wardrop, J. G. (1952) Some theoretical aspects of road traffic research. *Proceedings, Institution of Civil Engineers, II* **1**, 325-378.

[69] Whitt, W. (1986) Deciding which queue to join: Some counterexamples. *Queueing Systems* **34**, 55-63.

[70] Whitt, W. (1999) Improving service by informing customers about anticipated delays. *Management Science* **45**, 192-207.

[71] Winston, W. (1977) Optimality of the shortest line discipline. *Journal of Applied Probability* **14**, 181-189.

[72] Ye, H. Q. (2007) A paradox for admission control of multiclass queuing network with differentiated service. *Journal of Applied Probability* **44**, 321-331.

[73] Zhuang, W. and Li, M. Z. F. (2010) A new method of proving structural properties for certain class of stochastic dynamic control problems. *Operations Research Letters* **38**, 462-467.

[74] Ziedins, I. (2007) A paradox in a queueing network with state-dependent routing and loss. *Journal of Applied Mathematics and Decision Science* **2007**, ID: 68280.