Journal of the Operational Research Society (2015), 1–15

© 2015 Operational Research Society Ltd. All rights reserved. 0160-5682/15

www.palgrave-journals.com/jors/

Open

Containment of socially optimal policies in multiple-facility Markovian queueing systems

Rob Shone*, Vincent A Knight, Paul R Harper, Janet E Williams and John Minty

Cardiff University, Cardiff, UK

We consider a Markovian queueing system with N heterogeneous service facilities, each of which has multiple servers available, linear holding costs, a fixed value of service and a first-come-first-serve queue discipline. Customers arriving in the system can be either rejected or sent to one of the N facilities. Two different types of control policies are considered, which we refer to as 'selfishly optimal' and 'socially optimal'. We prove the equivalence of two different Markov Decision Process formulations, and then show that classical $M/M/1$ queue results from the early literature on behavioural queueing theory can be generalized to multiple dimensions in an elegant way. In particular, the state space of the continuous-time Markov process induced by a socially optimal policy is contained within that of the selfishly optimal policy. We also show that this result holds when customers are divided into an arbitrary number of heterogeneous classes, provided that the service rates remain non-discriminatory.

Journal of the Operational Research Society advance online publication, 25 November 2015; doi:10.1057/jors.2015.98

Keywords: queues with balking; Markov Decision Processes; equilibrium strategies; optimal strategies; dynamic programming

The online version of this article is available Open Access

1. Introduction

One of the most persistent themes in the literature on behavioural queueing theory is the sub-optimality of greedy or 'selfish' customer behaviour in the context of overall social welfare. In order to induce the most favourable scenario for society as a whole, customers are typically required to deviate in some way from the actions that they would choose if they were motivated only by their own interests. This principle has been observed in many of the classical queueing system models, including $M/M/1$, $GI/M/1$, $GI/M/s$ and others (see, eg, Naor, 1969; Yechiali, 1971; Knudsen, 1972; Yechiali, 1972; Littlechild, 1974; Edelson and Hildebrand, 1975; Lippman and Stidham, 1977; Stidham, 1978). More recently, this theme has been explored in applications including queues with setup and closedown times (Sun *et al*, 2010), queues with server breakdowns and delayed repairs (Wang and Zhang, 2011), vacation queues with partial information (Guo and Li, 2013), queues with compartmented waiting space (Economou and Kanta, 2008) and routing in public services (Knight *et al*, 2012; Knight and Harper, 2013). More generally, the implications of selfish and social decision making have been studied in various applications of economics and computer science; Roughgarden's (2005) monograph provides an overview of this work and poses some open problems.

The first author to compare 'self-optimization' with 'overall optimization' in a queueing setting was Naor (1969), whose classical model consists of an $M/M/1$ system with linear waiting costs and a fixed service value. The general queueing system that we consider in this paper may be regarded as an extension of Naor's model to a higher-dimensional space. We consider a system with $N \geqslant 2$ heterogeneous *service facilities* in parallel, each of which has its own queue and operates with a cost and reward structure similar to that of Naor's single-server model (see Figure 1). In addition, we generalize the system by assuming that each facility i may serve up to c_i customers simultaneously, so that we are essentially considering a network of non-identical $M/M/c_i$ queues.

The inspiration for our work is derived primarily from public service settings in which customers may receive service at any one of a number of different locations. For example, in a healthcare setting, patients requiring a particular operation procedure might choose between various different healthcare providers (or a choice might be made on their behalf by a central authority). In this context, the ith provider is able to treat up to c_i patients at once, and any further arrivals are required to join a waiting list, or seek treatment elsewhere. A further application of this work involves the queueing process at immigration control at ports and/or airports. These queues are often centrally controlled by an officer aiming to ensure that congestion is reduced. Finally, computer data traffic provides yet another application of this work. When transferring packets of data

*Correspondence: Rob Shone, School of Mathematics, Cardiff University, Senghennydd Road, Cardiff CF24 4AG, UK.

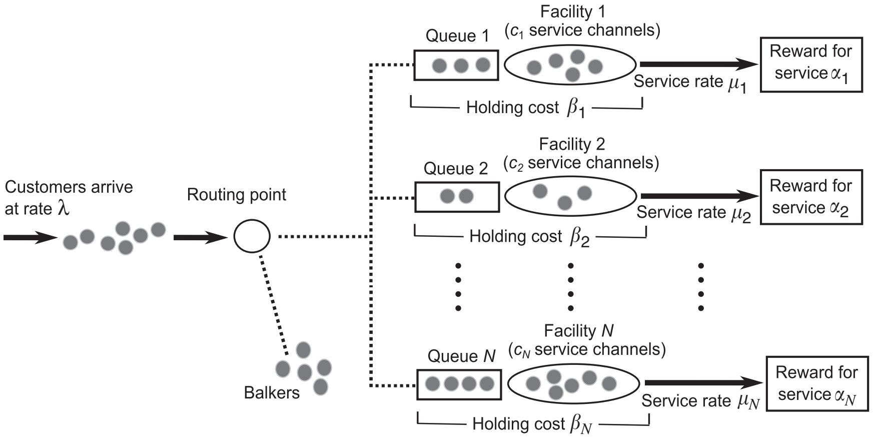


Figure 1 A diagrammatic representation of the queueing system.

over a network, there arise instances at which choices of available servers can have a major impact on the efficacy of the entire network.

The queueing system that we consider in this work evolves stochastically according to transitions which we assume are governed by Markovian distributions. We address the problem of finding an optimal routing and admission control policy and model this as a *Markov Decision Process* (*MDP*) (see, eg, Puterman, 1994) for a complete description and rigorous theoretical treatment of MDPs). Stidham and Weber (1993) provide an overview of MDP models for the control of queueing networks. It is well-known that optimal policies for the allocation of customers to parallel heterogeneous queues are not easy to characterize; for this reason, heuristic approaches have been developed which have achieved promising results in applications (see Argon *et al*, 2009; Glazebrook *et al*, 2009). In attempting to identify or approximate an optimal policy, one aims to find a dynamic decision-making scheme which optimizes the overall performance of the system with respect to a given criterion; we refer to such a scheme as a *socially optimal* solution to the optimization problem associated with the MDP. In this paper our objective is to draw inferences about the nature of a socially optimal solution from the structure of the corresponding *selfishly optimal* solution. A selfishly optimal solution may be regarded as a simple heuristic rule which optimizes a customer's immediate outcome without giving due consideration to long-term consequences. The remaining sections in this paper are organized as follows:

- In Section 2 we provide an MDP formulation of our queueing system and define all of the input parameters. We also offer an alternative formulation and show that it is equivalent.
- In Section 3 we define 'selfishly optimal' and 'socially optimal' policies in more detail. We then show that our model satisfies certain conditions which imply the existence of a *stationary* socially optimal policy, and prove an important relationship between the structures of the selfishly and socially optimal policies.
- In Section 4 we draw comparisons between the results of Section 3 and known results for systems of *unobservable* queues.
- In Section 5 we show that the results of Section 3 hold when customers are divided into an arbitrary number of heterogeneous classes. These classes are heterogeneous with respect to demand rates, holding costs and service values, but not service rates.
- Finally, in Section 6, we discuss the results of this paper and possible avenues for future research.

2. Model formulation

We consider a queueing system with N service facilities. Customers arrive from a single demand node according to a stationary Poisson process with demand rate $\lambda > 0$. Let facility i (for $i = 1, 2, \dots, N$) have c_i identical service channels, a linear holding cost $\beta_i > 0$ per customer per unit time, and a fixed value of service (or fixed reward) $\alpha_i > 0$. Service times at any server of facility i are assumed to be exponentially distributed with mean μ_i^{-1}. We assume $\alpha_i \geqslant \beta_i/\mu_i$ for each facility i in order to avoid degenerate cases where the reward for service fails to compensate for the expected costs accrued during a service time. When a customer arrives, they can proceed to one of the N facilities or, alternatively, exit from the system without receiving service (referred to as *balking*). Thus, there are $N+1$ possible decisions that can be made upon a customer's arrival. The decision chosen is assumed to be *irrevocable*; we do not allow reneging or jockeying between queues. The queue discipline at each facility is first-come-first-served (FCFS). A diagrammatic representation of the system is given in Figure 1.

We define $S := \{\mathbf{x} = (x_1, x_2, \dots x_N) : x_1, x_2, \dots, x_N \in \mathbb{N}_0\}$ to be the *state space* of our system, where x_i (the ith component of the vector $\mathbf{x}$) is the number of customers present (including those in service and those waiting in the queue) at facility i.

It is assumed that the system state is always known and can be used to inform decision making.

No binding assumption is made in this paper as to whether decisions are made by individual customers themselves, or whether actions are chosen on their behalf by a central controller. It is natural to suppose that *selfish* decision making occurs in the former case, whereas *socially optimal* behaviour requires some form of central control, and the discussion in this paper will tend to be consistent with this viewpoint; however, the results in this paper remain valid under alternative perspectives (eg, socially optimal behaviour might arise from selfless co-operation between customers).

We do not assume any upper bound on the value of λ in terms of the other parameters. However, the types of policies that we consider in this work always induce system stability. For convenience, we will use the notation $\mathbf{x}^{i+}$ to denote the state which is identical to $\mathbf{x}$ except that one extra customer is present at facility i; similarly, when $x_i \geqslant 1$, we use $\mathbf{x}^{i-}$ to denote the state with one fewer customer present at facility i. That is:

$$\mathbf{x}^{i+} := \mathbf{x} + \mathbf{e}_i$$
$$\mathbf{x}^{i-} := \mathbf{x} - \mathbf{e}_i$$

where $\mathbf{e}_i$ is the ith vector in the standard orthonormal basis of $\mathbb{R}^N$.

Let us discretize the system by defining:

$$\Delta = \left(\lambda + \sum_{i=1}^{N} c_i \mu_i \right)^{-1}$$

and considering an MDP which evolves in discrete time steps of size Δ. Using the well-known technique of *uniformization*, usually attributed to Lippman (1975) (see also Serfozo, 1979), we can analyse the system within a discrete-time framework, in which arrivals and service completions occur only at the ‘jump times’ of the discretized process. At any time step, the probability that a customer arrives is $\lambda\Delta$, and the probability that a service completes at facility i is either $c_i\mu_i\Delta$ or $x_i\mu_i\Delta$, depending on whether or not all of the channels at facility i are in use. At each time step, an action $a \in \{0, 1, 2, \ldots, N\}$ is chosen which represents the destination of any customer who arrives at that particular step; if $a = 0$ then the customer balks from the system, and if $a = i$ (for $i \in \{1, 2, \ldots, N\}$) then the customer joins facility i. This leads to the following definition for the *transition probabilities* $p_{\mathbf{xy}}(a)$ for transferring from state $\mathbf{x}$ to $\mathbf{y}$ in a single discrete time step, given that action a is chosen:

$$p_{\mathbf{xy}}(a) = \begin{cases} \lambda\Delta, & \mathbf{y} = \mathbf{x}^{i+} \text{ and } a = i \neq 0, \\ \min(x_i, c_i)\mu_i\Delta, & \mathbf{y} = \mathbf{x}^{i-}, \\ 1 - \left(I(a \neq 0)\lambda\Delta + \sum\limits_{i=1}^{N} \min(x_i, c_i)\mu_i\Delta \right), & \mathbf{y} = \mathbf{x}, \\ 0, & \text{otherwise} \end{cases}$$

Here we have used I to denote the indicator function. Since the units of time can always be re-scaled, we may assume $\Delta = 1$ without loss of generality, and we therefore suppress Δ in the remainder of this work. Figure 2 illustrates these transition probabilities diagrammatically.

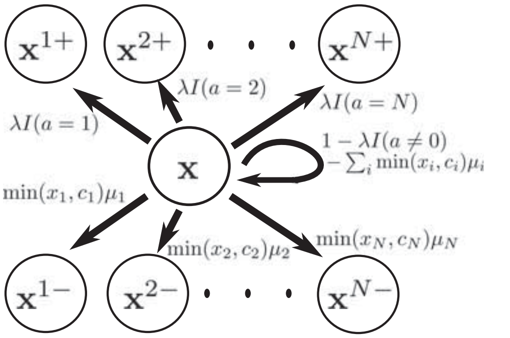


Figure 2 Transition probabilities (marked next to arrows) from an arbitrary state $\mathbf{x} \in S$.

If the system is in some state $\mathbf{x} \in S$ at a particular time step, the sum of the holding costs incurred is $\Sigma_{i=1}^{N} \beta_i x_i$; meanwhile, services are completing at an overall rate $\Sigma_{i=1}^{N} \min(x_i, c_i)\mu_i$ and the value of service at facility i is α_i. This leads to the following definition for the *single-step expected net reward* $r(\mathbf{x})$ associated with being in state $\mathbf{x}$ at a particular time step:

$$r(\mathbf{x}) := \sum_{i=1}^{N} (\min(x_i, c_i)\alpha_i\mu_i - \beta_i x_i) \tag{1}$$

The system may be controlled by means of a *policy* which determines, for each $n \in \mathbb{N}_0$, the action a_n to be chosen after n time steps. In this paper we focus on *stationary non-randomized* policies, under which the action a_n is chosen deterministically according to the accompanying system state $\mathbf{x}_n$, and is not dependent on other factors (such as the history of past states and actions, or the time index n). In Section 3 it will be shown that it is always possible to find a policy of the aforementioned type which achieves optimality in our system. Let $\mathbf{x}_n$ and a_n be, respectively, the state of the system and accompanying action chosen after n time steps, and let us use θ to denote the stationary policy being followed. The *long-run average net reward* $g_\theta(\mathbf{x}, r)$ per time step, given an initial state $\mathbf{x}_0 = \mathbf{x}$ and reward function r, is given by:

$$g_\theta(\mathbf{x}, r) = \lim_{t \to \infty} t^{-1} E_\theta \left[\sum_{n=0}^{t-1} r(\mathbf{x}_n) \mid \mathbf{x}_0 = \mathbf{x} \right] \tag{2}$$

where the dependence of $\mathbf{x}_n$ on the previous state $\mathbf{x}_{n-1}$ and action a_{n-1} is implicit. Before proceeding, we will show that an alternative definition of the reward function r yields the same long-run average reward (assuming that the same policy is followed). If a customer joins facility i under system state $\mathbf{x}$, then their *individual expected net reward*, taking into account the expected waiting time, holding cost β_i and value of service α_i, is given by $\alpha_i - \beta_i/\mu_i$ if they begin service immediately, and

$\alpha_i - \beta_i(x_i+1)/(c_i\mu_i)$ otherwise. Given that the probability of a customer arriving at any time step is λ, this suggests the possibility of a new reward function $\hat{r}$, which (unlike the function r defined in (1)) depends on the chosen action a in addition to the state $\mathbf{x}$:

$$\hat{r}(\mathbf{x}, a) = \begin{cases} \lambda\left(\alpha_i - \frac{\beta_i}{\mu_i}\right), & a = i \neq 0, x_i < c_i, \\ \lambda\left(\alpha_i - \frac{\beta_i(x_i+1)}{c_i\mu_i}\right), & a = i \neq 0, x_i \geqslant c_i, \\ 0, & a = 0 \end{cases} \tag{3}$$

The two reward functions in (1) and (3) look very different at first sight, but both formulations are entirely logical. The original definition in (1) is based on the real-time holding costs and rewards accrued during the system's evolution, while the alternative formulation in (3) is based on an unbiased estimate of each individual customer's contribution to the aggregate net reward, made at the time of their entry to the system. We will henceforth refer to the function r in (1) as the *real-time* reward function, and the function $\hat{r}$ in (3) as the *anticipatory* reward function. Our first result proves algebraically that these two reward formulations are equivalent.

Lemma 1 *For any stationary policy θ we have*:

$$g_\theta(\mathbf{x}, r) = g_\theta(\mathbf{x}, \hat{r}) \tag{4}$$

where r and $\hat{r}$ are defined as in (1) and (3) respectively. That is, the long-run average net reward under θ is the same under either reward formulation.

Proof We assume the existence of a stationary distribution $\{\pi_\theta(\mathbf{x})\}_{x\in S}$, where $\pi_\theta(\mathbf{x})$ is the steady-state probability of being in state $\mathbf{x} \in S$ under the stationary policy θ and $\sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x}) = 1$. If no such distribution exists, then the system is unstable under θ and both quantities in (4) are infinite. Under steady-state conditions, we can write:

$$g_\theta(\mathbf{x}, r) = \sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})r(\mathbf{x}),$$

$$g_\theta(\mathbf{x}, \hat{r}) = \sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})\hat{r}(\mathbf{x}, \theta(\mathbf{x}))$$

noting, as before, that $\hat{r}$ (unlike r) has a dependence on the action $\theta(\mathbf{x})$ associated with $\mathbf{x}$. For each $\mathbf{x} \in S$, the steady-state probability $\pi_\theta(\mathbf{x})$ is the same under either reward formulation since we are considering a fixed stationary policy. Our objective is to show:

$$\sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})r(\mathbf{x}) = \sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})\hat{r}(\mathbf{x}, \theta(\mathbf{x}))$$

We begin by partitioning the state space S into disjoint subsets. For each facility $i \in \{1, 2, \ldots, N\}$, let S_i denote the (possibly empty) set of states at which the action chosen under the policy θ is to join i. Then $S_i = S_{i-} \cup S_{i+}$, where:

$$S_{i-} := \{\mathbf{x} \in S : \theta(\mathbf{x}) = i \,\text{and}\, x_i < c_i\}$$

$$S_{i+} := \{\mathbf{x} \in S : \theta(\mathbf{x}) = i \,\text{and}\, x_i \geqslant c_i\}$$

We also let S_0 denote the set of states at which the action chosen under θ is to balk. Now let $g_\theta(\mathbf{x}, r)$ and $g_\theta(\mathbf{x}, \hat{r})$ be divided into 'positive' and 'negative' constituents in the following way:

$$g_\theta^+(\mathbf{x}, r) := \sum_{\mathbf{x}\in S}\sum_{i=1}^{N}\pi_\theta(\mathbf{x})\min(x_i, c_i)\alpha_i\mu_i,$$

$$g_\theta^-(\mathbf{x}, r) := -\sum_{\mathbf{x}\in S}\sum_{i=1}^{N}\pi_\theta(\mathbf{x})\beta_i x_i,$$

$$g_\theta^+(\mathbf{x}, \hat{r}) := \lambda\sum_{i=1}^{N}\sum_{\mathbf{x}\in S_i}\pi_\theta(\mathbf{x})\alpha_i,$$

$$g_\theta^-(\mathbf{x}, \hat{r}) := -\lambda\sum_{i=1}^{N}\left(\sum_{\mathbf{x}\in S_{i-}}\pi_\theta(\mathbf{x})\frac{\beta_i}{\mu_i} + \sum_{\mathbf{x}\in S_{i+}}\pi_\theta(\mathbf{x})\frac{\beta_i(x_i+1)}{c_i\mu_i}\right)$$

By referring to (1) and (3), it can be checked that $g_\theta(\mathbf{x}, r) = g_\theta^+(\mathbf{x}, r) + g_\theta^-(\mathbf{x}, r)$ and $g_\theta(\mathbf{x}, \hat{r}) = g_\theta^+(\mathbf{x}, \hat{r}) + g_\theta^-(\mathbf{x}, \hat{r})$. It will be sufficient to show that $g_\theta^+(\mathbf{x}, r) = g_\theta^+(\mathbf{x}, \hat{r})$ and $g_\theta^-(\mathbf{x}, r) = g_\theta^-(\mathbf{x}, \hat{r})$. Let $S_{i,k} \subseteq S_i$ (for $k = 0, 1, 2, \ldots$) be the set of states at which the action chosen under θ is to join facility i, given that there are k customers present there. That is:

$$S_{i,k} := \{\mathbf{x} \in S : \theta(\mathbf{x}) = i \,\text{and}\, x_i = k\}$$

Using the detailed balance equations for ergodic Markov chains under steady-state conditions (see, eg, Cinlar, 1975) we may assert that for every facility i and $k \geqslant 0$, the total flow from all states $\mathbf{x} \in S$ with $x_i = k$ up to states with $x_i = k+1$ must equal the total flow from states with $x_i = k+1$ down to $x_i = k$. Hence:

$$\lambda\sum_{\mathbf{x}\in S_{i,k}}\pi_\theta(\mathbf{x}) = \sum_{\substack{\mathbf{x}\in S \\ x_i = k+1}}\pi_\theta(\mathbf{x})\min(x_i, c_i)\mu_i \tag{5}$$

Summing over all $k \in \mathbb{N}_0$, we obtain:

$$\lambda\sum_{\mathbf{x}\in S_i}\pi_\theta(\mathbf{x}) = \sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})\min(x_i, c_i)\mu_i \tag{6}$$

which holds for $i \in \{1, 2, \ldots, N\}$. The physical interpretation of (6) is that, under steady-state conditions, the rate at which customers join facility i is equal to the rate at which service completions occur at i. Multiplying both sides of (6) by α_i and summing over $i \in \{1, 2, \ldots, N\}$, we have:

$$\lambda\sum_{i=1}^{N}\sum_{\mathbf{x}\in S_i}\pi_\theta(\mathbf{x})\alpha_i = \sum_{i=1}^{N}\sum_{\mathbf{x}\in S}\pi_\theta(\mathbf{x})\min(x_i, c_i)\alpha_i\mu_i$$

which states that $g_\theta^+(\mathbf{x},\hat{r}) = g_\theta^+(\mathbf{x},r)$ as required. It remains for us to show that $g_\theta^-(\mathbf{x},\hat{r}) = g_\theta^-(\mathbf{x},r)$. We proceed as follows: in (5) (which holds for all $k \in \mathbb{N}_0$ and $i \in \{1,2,..,N\}$), put $k = c_i$ to obtain:

$$\lambda \sum_{\mathbf{x}\in S_{i,c_i}} \pi_\theta(\mathbf{x}) = \sum_{\substack{\mathbf{x}\in S \\ x_i = c_i+1}} \pi_\theta(\mathbf{x}) c_i \mu_i \tag{7}$$

Suppose we multiply both sides of (7) by c_i+1. Since the sum on the left-hand side is over $\mathbf{x} \in S_{i,c_i}$ and the sum on the right-hand side is over states with $x_i = c_i + 1$, this is equivalent to multiplying each summand on the left-hand side by $x_i + 1$ and each summand on the right-hand side by x_i. In addition, multiplying both sides by $\beta_i/(c_i\mu_i)$ yields:

$$\lambda \sum_{\mathbf{x}\in S_{i,c_i}} \pi_\theta(\mathbf{x}) \frac{\beta_i(x_i+1)}{c_i\mu_i} = \sum_{\substack{\mathbf{x}\in S \\ x_i = c_i+1}} \pi_\theta(\mathbf{x}) \beta_i x_i \tag{8}$$

We can write similar expressions with $k = c_i + 1$, $c_i + 2$ and so on. Recall that $\bigcup_{k=c_i}^\infty S_{i,k} = S_{i+}$ by definition. Hence, by summing over all $k \geqslant c_i$ in (8) we obtain:

$$\lambda \sum_{\mathbf{x}\in S_{i+}} \pi_\theta(\mathbf{x}) \frac{\beta_i(x_i+1)}{c_i\mu_i} = \sum_{\substack{\mathbf{x}\in S \\ x_i \geqslant c_i+1}} \pi_\theta(\mathbf{x}) \beta_i x_i \tag{9}$$

Note also that multiplying both sides of (5) by β_i/μ_i and summing over all $k < c_i$ (and recalling that $\bigcup_{k=0}^{c_i-1} S_{i,k} = S_{i-}$) gives:

$$\lambda \sum_{\mathbf{x}\in S_{i-}} \pi_\theta(\mathbf{x}) \frac{\beta_i}{\mu_i} = \sum_{\substack{\mathbf{x}\in S \\ x_i \leqslant c_i}} \pi_\theta(\mathbf{x}) \beta_i x_i \tag{10}$$

Hence, from (9) and (10) we have:

$$\lambda\left(\sum_{\mathbf{x}\in S_{i-}} \pi_\theta(\mathbf{x}) \frac{\beta_i}{\mu_i} + \sum_{\mathbf{x}\in S_{i+}} \pi_\theta(\mathbf{x}) \frac{\beta_i(x_i+1)}{c_i\mu_i}\right) = \sum_{\mathbf{x}\in S} \pi_\theta(\mathbf{x}) \beta_i x_i$$

Summing over $i \in \{1,2,\ldots,N\}$ gives $g_\theta^-(\mathbf{x},\hat{r}) = g_\theta^-(\mathbf{x},r)$ as required. We have already shown that $g_\theta^+(\mathbf{x},\hat{r}) = g_\theta^+(\mathbf{x},r)$, so this completes the proof that $g_\theta(\mathbf{x},\hat{r}) = g_\theta(\mathbf{x},r)$. □

It follows from Lemma 1 that any policy which is optimal among stationary policies under one reward formulation (either r or $\hat{r}$) is likewise optimal under the other formulation, with the same long-run average reward. The interchangeability of these two reward formulations will assist us in proving later results.

3. Containment of socially optimal policies

Let us define what we will refer to as 'selfishly optimal' and 'socially optimal' policies. The terminology used in this paper is slightly incongruous to that which is typically found in the literature on MDPs, and the main reason for this is that we wish to draw analogies with the work of Naor (1969). The policies which we describe as 'socially optimal' are those which satisfy the well-known *Bellman optimality equations* of dynamic programming (introduced by Bellman, 1957), and would be referred to by many authors simply as 'optimal' policies; on the other hand, the 'selfishly optimal' policies that we will describe could alternatively be referred to as 'greedy' or 'myopic' policies.

We begin with *selfishly* optimal policies. Suppose that each customer arriving in the system is allowed to make his or her own decision (as opposed to being directed by a central decision-maker). It is assumed throughout this work that the queueing system is *fully observable* and therefore the customer is able to observe the exact state of the system, including the length of each queue and the occupancy of each facility (the case of *unobservable queues* is a separate problem; see, eg, Bell and Stidham, 1983; Haviv and Roughgarden, 2007; Shone *et al*, 2013). Under this scenario, a customer may calculate their expected net reward (taking into account the expected cost of waiting and the value of service) at each facility based on the number of customers present there using a formula similar to (3); if they act selfishly, they will simply choose the option which maximizes this expected net reward. If the congestion level of the system is such that all of these expected net rewards are negative, we assume that the (selfish) customer's decision is to balk. This definition of selfish behaviour generalizes Naor's simple decision rule for deciding whether to join or balk in an *M*/*M*/1 system. We note that since the FCFS queue discipline is assumed at each facility, a selfish customer's behaviour depends only on the existing state, and is not influenced by the knowledge that other customers act selfishly.

Taking advantage of the 'anticipatory' reward formulation in (3), we can define a *selfishly optimal* policy $\tilde{\theta}$ by:

$$\tilde{\theta}(\mathbf{x}) \in \underset{a\in\{0,1,2,\ldots,N\}}{\arg\max}\ \hat{r}(\mathbf{x},a) \qquad (\mathbf{x}\in S)$$

In the case of ties, we assume that the customer joins the facility with the smallest index i; however, balking is never chosen over joining facility i when $\hat{r}(\mathbf{x},i) = 0$. This is in keeping with Naor's convention.

A *socially optimal* policy, denoted θ^*, is any policy which maximizes the long-run average net reward defined in (2). The optimality equations for our system, derived from the classical Bellman optimality equations for average reward problems (see, eg, Puterman, 1994) and assuming the real-time reward formulation in (1), may be expressed as:

$$g^* + h(\mathbf{x}) = r(\mathbf{x}) + \lambda \max_a\{h(\mathbf{x}^{a+})\} + \sum_{i=1}^N \min(x_i,c_i)\mu_i h(\mathbf{x}^{i-}) + \left(1 - \lambda - \sum_{i=1}^N \min(x_i,c_i)\mu_i\right) h(\mathbf{x}) \qquad (\mathbf{x}\in S) \tag{11}$$

where $h(\mathbf{x})$ is a *relative value function* and g^* is the optimal long-run average net reward. (We adopt the notational convention that $\mathbf{x}^{0+} = \mathbf{x}$ to deal with the case where balking is optimal in (11).) Under the anticipatory reward formulation in (3) these optimality equations are similar except that $r(\mathbf{x})$ is replaced by $\hat{r}(\mathbf{x}, a)$, which must obviously be included within the maximization operator. Indeed, by adopting $\hat{r}(\mathbf{x}, a)$ as our reward formulation we may observe the fundamental difference between the selfishly and socially optimal policies: the *selfish* policy simply maximizes the *immediate* reward $\hat{r}(\mathbf{x}, a)$, without taking into account the extra term $h(\mathbf{x}^{a+})$; this is why it may be called a *myopic* policy. The physical interpretation is that under the *selfish* policy, customers consider only the outcome to themselves, without taking into account the implications for future customers, who may suffer undesirable consequences as a result of their behaviour.

In this work we assume an *infinite time horizon*, but we use the *method of successive approximations* (see Ross, 1983) to treat the infinite horizon problem as the limiting case of a finite horizon problem. We therefore state the *finite horizon* optimality equations corresponding to the infinite horizon equations in (11):

$$v_{n+1}^*(\mathbf{x}) = r(\mathbf{x}) + \lambda \max_a \{v_n^*(\mathbf{x}^{a+})\} + \sum_{i=1}^{N} \min(x_i, c_i)\mu_i v_n^*(\mathbf{x}^{i-})$$
$$+ \left(1 - \lambda - \sum_{i=1}^{N} \min(x_i, c_i)\mu_i\right) v_n^*(\mathbf{x}) \qquad (\mathbf{x} \in S, n \geqslant 0) \qquad (12)$$

where $v^*_n(\mathbf{x})$ is the maximal expected total reward from a problem with n time steps, given an initial state $\mathbf{x} \in S$ (we define $v^*_0(\mathbf{x}) = 0$ for all $\mathbf{x} \in S$).

Remark It has already been shown (Lemma 1) that, in an infinite-horizon problem, a stationary policy earns the same long-run average reward under either of the reward formulations r and $\hat{r}$. However, this equivalence is lost when we consider *finite-horizon* problems. Indeed, given a finite horizon n, a policy which is optimal under reward function r may perform extremely poorly under $\hat{r}$. This is especially likely to be the case if n is small.

Given that selfish customers refuse to choose facility i if $\hat{r}(\mathbf{x}, i) < 0$, it follows that for $i = 1, 2, \ldots, N$ there exists an upper threshold b_i which represents the greatest possible number of customers at i under steady-state conditions. The value b_i can be derived from (3) as:

$$b_i := \left\lfloor \frac{c_i \alpha_i \mu_i}{\beta_i} \right\rfloor$$

where $\lfloor \cdot \rfloor$ denotes the integer part. Two important ways in which the selfishly optimal policy $\tilde{\theta}$ differs from a socially optimal policy are as follows:

1. The decisions made under $\tilde{\theta}$ are entirely independent of the demand rate λ.
2. The threshold b_i (representing the steady-state maximum occupancy at i) is independent of the parameters for the other facilities $j \neq i$.

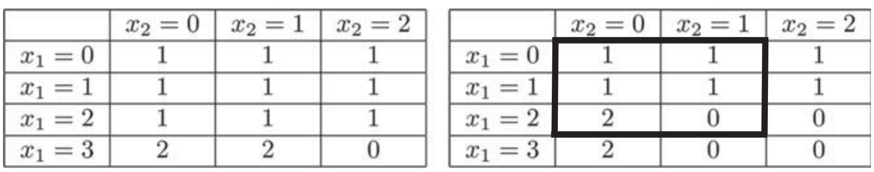
Because of the thresholds b_i, a selfishly optimal policy $\tilde{\theta}$ induces an ergodic Markov chain defined on a *finite* set of states $\tilde{S} \subset S$. Formally, we have:

$$\tilde{S} := \{(x_1, x_2, ..., x_N) : x_i \leqslant b_i \text{ for all } i\} \qquad (13)$$

We will refer to $\tilde{S}$ as the *selfishly optimal state space*. Note that, due to the convention that the facility with the smallest index i is chosen in the case of a tie between the expected net rewards at two or more facilities, the selfishly optimal policy $\tilde{\theta}$ is unique in any given problem. Changing the ordering of the facilities (and thereby the tie-breaking rules) affects the policy $\tilde{\theta}$, but does *not* alter the boundaries of $\tilde{S}$.

Let S_{θ^*} denote the set of positive recurrent states belonging to the Markov chain induced by a *socially optimal* policy θ^* satisfying the optimality equations in (11). The main result to be proved in this section is that S_{θ^*} is not only finite, but must also be contained in $\tilde{S}$.

Example 1 *Consider a system with demand rate $\lambda = 12$ and only two facilities. The first facility has two channels available ($c_1 = 2$) and a service rate $\mu_1 = 5$, holding cost $\beta_1 = 3$ and fixed reward $\alpha_1 = 1$. The parameters for the second facility are $c_2 = 2$, $\mu_2 = 1$, $\beta_2 = 3$ and $\alpha_2 = 3$, so it offers a higher reward but a slower service rate. We can uniformize the system by taking $\Delta = 1/24$, so that $(\lambda + \sum_i c_i \mu_i)\Delta = 1$. The selfishly optimal state space $\tilde{S}$ for this system consists of 12 states. Figure 3 shows the*



Selfish Policy

	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$
$x_1 = 0$	1	1	1
$x_1 = 1$	1	1	1
$x_1 = 2$	1	1	1
$x_1 = 3$	2	2	0

Social Policy

	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$
$x_1 = 0$	1	1	1
$x_1 = 1$	1	1	1
$x_1 = 2$	2	0	0
$x_1 = 3$	2	0	0

Figure 3 Selfishly and socially optimal policies for Example 1.
Note: For each state $\mathbf{x} = (x_1, x_2) \in \tilde{S}$, the corresponding decisions under the respective policies are shown.

decisions taken at these states under the selfishly optimal policy $\tilde{\theta}$, and also the corresponding decisions taken under a socially optimal policy θ^*.

By comparing the tables in Figure 3 we may observe the differences between the policies $\tilde{\theta}$ and θ^*. At the states (2, 0), (2, 1), (2, 2) and (3, 1), the socially optimal policy θ^* deviates from the selfish policy $\tilde{\theta}$ (incidentally, the sub-optimality of the selfish policy is about 22%). More striking, however, is the fact that under the socially optimal policy, some of the states in $\tilde{S}$ are actually unattainable under steady-state conditions. Indeed, the recurrent state space S_{θ^*} consists of only six states (enclosed by the bold rectangle in the figure). Thus, for this system, $S_{\theta^*} \subseteq \tilde{S}$ and in this section we aim to prove that this result holds in general.

It is known that for a general MDP defined on an infinite set of states, an average reward optimal policy need not exist, and that even if such a policy exists, it may be non-stationary. In 1983, Ross provides counter-examples to demonstrate both of these facts. Thus, it is desirable to establish the *existence* of an optimal stationary policy before aiming to examine its properties. Our approach in this section is based on the results of Sennott (1989), who has established sufficient conditions for the existence of an average reward optimal stationary policy for an MDP defined on an infinite state space (this problem has also been addressed by other authors; see, eg, Zijm, 1985; Cavazos-Cadena, 1989). We will proceed to show that Sennott's conditions are satisfied for our system, and then deduce that for any socially optimal policy θ^*, S_{θ^*} must be contained in $\tilde{S}$. Sennott's approach is based on the theory of *discounted reward* problems, in which a reward earned n steps into the future is discounted by a factor γ^n, where $0<\gamma<1$. A policy θ is said to be *γ-discount optimal* if it maximizes the *total expected discounted reward* (abbreviated henceforth as TEDR) over an infinite time horizon, defined (for reward function $\hat{r}$) as:

$$v_{\theta,\gamma}(\mathbf{x},\hat{r}) = E_\theta\left[\sum_{n=0}^{\infty}\gamma^n \hat{r}(\mathbf{x}_n,a_n) \mid \mathbf{x}_0=\mathbf{x}\right] \qquad (\mathbf{x}\in S) \quad (14)$$

Let θ_γ^* denote an optimal policy under discount rate γ, and let $v_\gamma^*(\mathbf{x},\hat{r})$ be the corresponding TEDR, so that $v_\gamma^*(\mathbf{x},\hat{r}) = \sup_\theta v_{\theta,\gamma}(\mathbf{x},\hat{r})$. It is known that $v_\gamma^*(\mathbf{x},\hat{r})$ satisfies the *discount optimality equations* (see Puterman, 1994):

$$v_\gamma^*(\mathbf{x},\hat{r}) = \max_a\left\{\hat{r}(\mathbf{x},a) + \gamma\sum_{\mathbf{y}\in S} p_{\mathbf{xy}}(a)v_\gamma^*(\mathbf{y},\hat{r})\right\} \qquad (\mathbf{x}\in S)$$

We proceed to show that in our system, the discount optimal value function $v^*{}_\gamma$ satisfies conditions which are sufficient for the existence of an *average reward optimal* stationary policy. In the proofs of the upcoming results, we adopt the anticipatory reward function $\hat{r}$ defined in (3).

Lemma 2 *For every state $\mathbf{x}\in S$ and discount rate $0<\gamma<1$:*

$$v_\gamma^*(\mathbf{x},\hat{r}) \geqslant 0$$

Proof Let θ_0 be the trivial policy of balking under every state. Each reward $\hat{r}(\mathbf{x}_n,\theta_0(\mathbf{x}_n))$ is zero and hence $v_{\theta_0,\gamma}(\mathbf{x},\hat{r}) = 0$. Since $v_\gamma^*(\mathbf{x},\hat{r}) = \sup_\theta v_{\theta,\gamma}(\mathbf{x},\hat{r})$ by definition, the result follows. □

The next result establishes an important monotonicity property of the function $v_\gamma^*(\mathbf{x},\hat{r})$ which, incidentally, does not hold for its counterpart $v_\gamma^*(\mathbf{x},r)$ under the real-time reward formulation (1).

Lemma 3 *For every state $\mathbf{x}\in S$, discount rate $0<\gamma<1$ and facility $i\in\{1,2,\ldots,N\}$, we have:*

$$v_\gamma^*\left(\mathbf{x}^{i+},\hat{r}\right) \leqslant v_\gamma^*(\mathbf{x},\hat{r})$$

Proof We rely on the finite horizon optimality equations (for discounted problems) and prove the result using induction on the number of stages. The finite horizon optimality equations are:

$$\begin{aligned} v_{\gamma,n+1}^*(\mathbf{x},\hat{r}) = {} & \max_a\left\{\hat{r}(\mathbf{x},a) + \gamma\lambda v_{\gamma,n}^*(\mathbf{x}^{a+},\hat{r})\right\} \\ & + \gamma\left[\sum_{i=1}^{N}\min(x_i,c_i)\mu_i v_{\gamma,n}^*\left(\mathbf{x}^{i-},\hat{r}\right)\right. \\ & \left. + \left(1-\lambda-\sum_{i=1}^{N}\min(x_i,c_i)\mu_i\right)v_{\gamma,n}^*(\mathbf{x},\hat{r})\right] \\ & (\mathbf{x}\in S,\ n\geqslant 0) \qquad (15) \end{aligned}$$

It is sufficient to show that for each state $\mathbf{x}\in S$, discount rate $0<\gamma<1$, facility $i\in\{1,2,\ldots,N\}$ and integer $n\geqslant 0$:

$$v_{\gamma,n}^*\left(\mathbf{x}^{i+},\hat{r}\right) \leqslant v_{\gamma,n}^*(\mathbf{x},\hat{r}) \qquad (16)$$

We define $v_{\gamma,0}^*(\mathbf{x},\hat{r}) = 0$ for all $\mathbf{x}\in S$. In order to show that (16) holds when $n=1$, we need to show, for $i=1,2,\ldots,N$:

$$\max_a \hat{r}\left(\mathbf{x}^{i+},a\right) \leqslant \max_b \hat{r}(\mathbf{x},b) \qquad (\mathbf{x}\in S)$$

Indeed, let $a^* \in \arg\max_a \hat{r}(\mathbf{x}^{i+},a)$. It follows from the definition of $\hat{r}(\mathbf{x},a)$ in (3) that $\hat{r}(\mathbf{x}^{i+},a) \leqslant \hat{r}(\mathbf{x},a)$ for any fixed action a and facility i. Hence:

$$\max_a \hat{r}\left(\mathbf{x}^{i+},a\right) = \hat{r}\left(\mathbf{x}^{i+},a^*\right) \leqslant \hat{r}\left(\mathbf{x},a^*\right) \leqslant \max_b \hat{r}(\mathbf{x},b)$$

Now let us assume that (16) also holds for $n=k$, where $k\geqslant 1$ is arbitrary, and aim to show $v^*_{\gamma,k+1}(\mathbf{x}^{i+},\hat{r})\leqslant v^*_{\gamma,k+1}(\mathbf{x},\hat{r})$. We have:

$$\begin{aligned}
& v^*_{\gamma,k+1}(\mathbf{x}^{i+},\hat{r})-v^*_{\gamma,k+1}(\mathbf{x},\hat{r})\\
&=\max_a\left\{\hat{r}\left(\mathbf{x}^{i+},a\right)+\gamma\lambda v^*_{\gamma,k}\left(\left(\mathbf{x}^{i+}\right)^{a+},\hat{r}\right)\right\}\\
&\quad-\max_b\left\{\hat{r}(\mathbf{x},b)+\gamma\lambda v^*_{\gamma,k}\left(\mathbf{x}^{b+},\hat{r}\right)\right\}\\
&\quad+\gamma\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\left(v^*_{\gamma,k}\left(\left(\mathbf{x}^{i+}\right)^{j-},\hat{r}\right)-v^*_{\gamma,k}\left(\mathbf{x}^{j-},\hat{r}\right)\right)\\
&\quad+\gamma\left(1-\lambda-\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\right)\left(v^*_{\gamma,k}\left(\mathbf{x}^{i+},\hat{r}\right)-v^*_{\gamma,k}(\mathbf{x},\hat{r})\right)\\
&\quad-\gamma I(x_i<c_i)\mu_i\left(v^*_{\gamma,k}\left(\mathbf{x}^{i+},\hat{r}\right)-v^*_{\gamma,k}(\mathbf{x},\hat{r})\right)
\end{aligned}\tag{17}$$

Note that the indicator term in (17) arises because, under state $\mathbf{x}^{i+}$, there may (or may not) be one extra service in progress at facility i, depending on whether or not $x_i<c_i$. Recall that we assume $\lambda+\Sigma_{i=1}^{N}c_i\mu_i=1$, hence $(1-\lambda-\Sigma_{j=1}^{N}\min(x_j,c_j)\mu_j-I(x_i<c_i)\mu_i)$ must always be non-negative. We also have $v^*_{\gamma,k}(\mathbf{x}^{i+},\hat{r})\leqslant v^*_{\gamma,k}(\mathbf{x},\hat{r})$ and $v^*_{\gamma,k}((\mathbf{x}^{i+})^{j-},\hat{r})\leqslant v^*_{\gamma,k}(\mathbf{x}^{j-},\hat{r})$ (for $j=1,2,\dots,N$) using our inductive assumption of monotonicity at stage k. Hence, in order to verify that (17) is non-positive, it suffices to show:

$$\begin{aligned}
&\max_a\left\{\hat{r}\left(\mathbf{x}^{i+},a\right)+\gamma\lambda v^*_{\gamma,k}\left(\left(\mathbf{x}^{i+}\right)^{a+},\hat{r}\right)\right\}\\
&\leqslant\max_b\left\{\hat{r}(\mathbf{x},b)+\gamma\lambda v^*_{\gamma,k}\left(\mathbf{x}^{b+},\hat{r}\right)\right\}
\end{aligned}\tag{18}$$

Here, let a^* be a maximizing action on the left-hand side, that is

$$a^*\in\arg\max_a\left\{\hat{r}\left(\mathbf{x}^{i+},a\right)+\gamma\lambda v^*_{\gamma,k}\left(\left(\mathbf{x}^{i+}\right)^{a+},\hat{r}\right)\right\}$$

By the monotonicity of $\hat{r}$ and our inductive assumption, we have:

$$\hat{r}\left(\mathbf{x}^{i+},a^*\right)\leqslant\hat{r}\left(\mathbf{x},a^*\right),$$

$$v^*_{\gamma,k}\left(\left(\mathbf{x}^{a^*+}\right)^{i+},\hat{r}\right)\leqslant v^*_{\gamma,k}\left(\mathbf{x}^{a^*+},\hat{r}\right)$$

Hence the left-hand side of (18) is bounded above by $\hat{r}(\mathbf{x},a^*)+\gamma\lambda v^*_{\gamma,k}(\mathbf{x}^{a^*+},\hat{r})$, which in turn is bounded above by $\max_b\left\{\hat{r}(\mathbf{x},b)+\gamma\lambda v^*_{\gamma,k}(\mathbf{x}^{b+},\hat{r})\right\}$. This shows that $v^*_{\gamma,k+1}(\mathbf{x}^{i+},\hat{r})\leqslant v^*_{\gamma,k+1}(\mathbf{x},\hat{r})$, which completes the inductive proof that (16) holds for all $n\in\mathbb{N}$. Using the method of 'successive approximations', Ross (1983) proves that $\lim_{n\to\infty}v^*_{\gamma,n}(\mathbf{x},\hat{r})=v^*_\gamma(\mathbf{x},\hat{r})$ for all $\mathbf{x}\in S$, and so we conclude that $v^*_\gamma(\mathbf{x}^{i+},\hat{r})\leqslant v^*_\gamma(\mathbf{x},\hat{r})$ as required. □

We require another lemma to establish a state-dependent lower bound for the relative value function h.

Lemma 4 *For every* $\mathbf{x}\in S$, *there exists a value* $M(\mathbf{x})>0$ *such that, for every discount rate* $0<\gamma<1$:

$$v^*_\gamma(\mathbf{x},\hat{r})-v^*_\gamma(\mathbf{0},\hat{r})\geqslant -M(\mathbf{x})$$

where $\mathbf{0}$ *denotes the 'empty system' state,* $(0,0,\dots,0)$.

Proof Let $\alpha_{\max}=\max_{i\in\{1,2,\dots,N\}}\alpha_i$ denote the maximum value of service across all facilities. For each discount rate $0<\gamma<1$ and policy θ, let us define a new function $w_{\theta,\gamma}$ by:

$$w_{\theta,\gamma}(\mathbf{x},\hat{r}):=E_\theta\left[\sum_{n=0}^{\infty}\gamma^n\left(\hat{r}(\mathbf{x}_n,a_n)-\lambda\alpha_{\max}\right)\ \middle|\ \mathbf{x}_0=\mathbf{x}\right]$$

$$(\mathbf{x}\in S)$$

By comparison with the definition of $v_{\theta,\gamma}$ in (14), we have:

$$w_{\theta,\gamma}(\mathbf{x},\hat{r})=v_{\theta,\gamma}(\mathbf{x},\hat{r})-\frac{\lambda\alpha_{\max}}{1-\gamma}$$

and since the subtraction of a constant from each single-step reward does not affect our optimality criterion, we also have:

$$w^*_\gamma(\mathbf{x},\hat{r})=v^*_\gamma(\mathbf{x},\hat{r})-\frac{\lambda\alpha_{\max}}{1-\gamma}\tag{19}$$

where $w^*_\gamma(\mathbf{x},\hat{r})=\sup_\theta w_{\theta,\gamma}(\mathbf{x},\hat{r})$. By the definition of $\hat{r}$ in (3) it can be checked that $\hat{r}(\mathbf{x},a)\leqslant\lambda\alpha_{\max}$ for all state-action pairs $(\mathbf{x},a)$. Therefore $w^*_\gamma(\mathbf{x},\hat{r})$ is a sum of non-positive terms and must be non-positive itself. Furthermore, $w*_\gamma$ is the TEDR function for a new MDP which is identical to our original MDP except that we replace each $\hat{r}(\mathbf{x}_n,a_n)$ (for $n=0,1,2,\dots$) by $\hat{r}(\mathbf{x}_n,a_n)-\lambda\alpha_{\max}$. Thus, $w*_\gamma$ satisfies:

$$w^*_\gamma(\mathbf{x},\hat{r})=\max_a\left\{\hat{r}(\mathbf{x},a)-\lambda\alpha_{\max}+\gamma\sum_{\mathbf{y}\in S}p_{\mathbf{xy}}(a)w^*_\gamma(\mathbf{y},\hat{r})\right\}$$

$$(\mathbf{x}\in S)\tag{20}$$

Consider $\mathbf{x}=\mathbf{0}^{i+}$, for an arbitrary $i\in\{1,2,\dots,N\}$. Using (20) we have, for all actions a:

$$w^*_\gamma(\mathbf{0}^{i+},\hat{r})\geqslant\hat{r}(\mathbf{0}^{i+},a)-\lambda\alpha_{\max}+\gamma\sum_{\mathbf{y}\in S}p_{\mathbf{0}^{i+},\mathbf{y}}(a)w^*_\gamma(\mathbf{y},\hat{r})$$

In particular, if the action $a=0$ is to balk then $\hat{r}(\mathbf{0}^{i+},a)=0$ and the only possible transitions are to states $\mathbf{0}$ or $\mathbf{0}^{i+}$. Hence:

$$w_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right) \geqslant -\lambda\alpha_{\max}+\gamma\mu_i w_\gamma^*(\mathbf{0},\hat{r})+\gamma(1-\mu_i)w_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right)$$

Then, since $\gamma\leqslant 1$ and by the non-positivity of $w_\gamma^*(\mathbf{0},\hat{r})$ and $w_\gamma^*(\mathbf{0}^{i+},\hat{r})$:

$$w_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right) \geqslant -\lambda\alpha_{\max}+\mu_i w_\gamma^*(\mathbf{0},\hat{r})+(1-\mu_i)w_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right) \tag{21}$$

From (19) and (21) we derive:

$$v_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right)-v_\gamma^*(\mathbf{0},\hat{r}) = w_\gamma^*\left(\mathbf{0}^{i+},\hat{r}\right)-w_\gamma^*(\mathbf{0},\hat{r}) \geqslant -\frac{\lambda\alpha_{\max}}{\mu_i} \tag{22}$$

so we have a lower bound for $v_\gamma^*(\mathbf{0}^{i+},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r})$ which is independent of γ as required. We need to show that for each $\mathbf{x}\in S$, a lower bound can be found for $v_\gamma^*(\mathbf{x},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r})$. Let us form a hypothesis as follows: for each state $\mathbf{x}\in S$, there exists a value $\psi(\boldsymbol{x})$ such that, for all γ:

$$v_\gamma^*(\mathbf{x},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r}) \geqslant -\lambda\alpha_{\max}\psi(\mathbf{x}) \tag{23}$$

We have $\psi(\mathbf{0})=0$ and, from (22), $\psi(\mathbf{0}^{i+})=\mu_i^{-1}$ for $i=1,$-$2,\ldots,N$. Let us aim to show that (23) holds for an arbitrary $\mathbf{x}\neq\mathbf{0}$, under the assumption that for all $j\in\{1,2,$-$\ldots,N\}$ with $x_j\geqslant 1$, (23) holds for the state $\mathbf{x}^{j-}$. Using similar steps to those used for $\mathbf{0}^{i+}$ earlier, we have:

$$w_\gamma^*(\mathbf{x},\hat{r}) \geqslant -\lambda\alpha_{\max}+\gamma\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j w_\gamma^*\left(\mathbf{x}^{j-},\hat{r}\right)$$
$$+\gamma\left(1-\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\right)w_\gamma^*(\mathbf{x},\hat{r})$$

and hence:

$$\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\left(w_\gamma^*(\mathbf{x},\hat{r})-w_\gamma^*\left(\mathbf{x}^{j-},\hat{r}\right)\right) \geqslant -\lambda\alpha_{\max}$$

Then, using our inductive assumption that, for each $j\in\{1,2,\ldots,N\}$, $w_\gamma^*(\mathbf{x}^{j-},\hat{r})-w_\gamma^*(\mathbf{0},\hat{r})$ is bounded below by $-\lambda\alpha_{\max}\psi(\mathbf{x}^{j-})$:

$$w_\gamma^*(\mathbf{x},\hat{r})-w_\gamma^*(\mathbf{0},\hat{r}) \geqslant -\lambda\alpha_{\max}\left(\frac{1+\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\psi(\mathbf{x}^{j-})}{\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j}\right) \tag{24}$$

Using (19), we conclude that the right-hand side of (24) is also a lower bound for $v_\gamma^*(\mathbf{x}^{j-},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r})$. Therefore we can define:

$$\psi(\mathbf{x}) := \frac{1+\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j\psi(\mathbf{x}^{j-})}{\sum_{j=1}^{N}\min\left(x_j,c_j\right)\mu_j}$$

with the result that $v_\gamma^*(\mathbf{x},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r})$ is bounded below by an expression which depends only on the system input parameters λ, $\alpha_{\max}$ and the service rates $\mu_1,\mu_2,\ldots,\mu_N$ as required. Using an inductive procedure, we can derive a lower bound of this form for every $\mathbf{x}\in S$. □

Lemma 5 *For all states $\mathbf{x}\in S$ and actions $a\in\{0,1,2,\ldots,N\}$:*

$$\sum_{\mathbf{y}\in S}p_{\mathbf{xy}}(a)M(\mathbf{y})<\infty$$

Where $-M(\mathbf{y})$ is the lower bound for $v_\gamma^*(\mathbf{y},\hat{r})-v_\gamma^*(\mathbf{0},\hat{r})$ derived in Lemma 4.

Proof This is immediate from Lemma 4 since, for any $\mathbf{x}\in S$, the number of 'neighbouring' states $\mathbf{y}$ that can be reached via a single transition from $\mathbf{x}$ is finite (regardless of the action chosen), and each $M(\mathbf{y})$ is finite. □

The results presented in this section thus far confirm that our system satisfies Sennott's (1989) conditions for the existence of an average reward optimal stationary policy. We now state this as a theorem.

Theorem 1 *Consider a sequence of discount rates (γ_n) converging to 1, with $(\theta_{\gamma_n}^*)$ the associated sequence of discount-optimal stationary policies. There exists a subsequence (η_n) of (γ_n) such that the limit*

$$\theta^* := \lim_{n\to\infty}\theta_{\eta_n}^*$$

exists, and the stationary policy θ^* is average reward optimal. Furthermore, the policy θ^* yields an average reward $g^*=\lim_{\gamma\uparrow 1}(1-\gamma)v_\gamma^*(\mathbf{x})$ which, together with a function $h(\mathbf{x})$, satisfies the optimality equations:

$$g^*+h(\mathbf{x}) = \max_a\left\{\hat{r}(\mathbf{x},a)+\sum_{\mathbf{y}\in S}p_{\mathbf{xy}}(a)h(\mathbf{y})\right\} \quad (\mathbf{x}\in S) \tag{25}$$

Proof We refer to Sennott (1989), who presents four assumptions which (together) are sufficient for the existence of an average reward optimal stationary policy in an MDP with an infinite state space. From Lemma 2 we have $v_\gamma^*(\mathbf{x},\hat{r})\geqslant 0$ for every $\mathbf{x}\in S$ and $\gamma\in(0,1)$, so a stronger version of Assumption 1 in Sennott (1989) holds. From Lemma 3 we have $v_\gamma^*(\mathbf{x}^{i+},\hat{r})\leqslant v_\gamma^*(\mathbf{x},\hat{r})$ for all $\mathbf{x}$, $i\in\{1,2,\ldots,N\}$ and γ, which implies $v_\gamma^*(\mathbf{x},\hat{r}) - v_\gamma^*(\mathbf{0},\hat{r}) \leqslant 0$ using an inductive argument. Therefore Assumption 2 in Sennott (1989) also holds. Assumptions 3 and 3* in Sennott (1989) follow directly from Lemmas 4 and 5. □

Our next result establishes the containment property of socially optimal policies which we alluded to earlier in this section.

Theorem 2 *There exists a stationary policy θ^* which satisfies the average reward optimality equations and which induces an ergodic Markov chain on some finite set S_{θ^*} of states contained in $\tilde{S}$.*

Informally, we say 'the socially optimal state space is contained within the selfishly optimal state space'.

Proof From the definition of $\tilde{S}$ in (13) we note that it is sufficient to show that for some stationary optimal policy θ^*, the action $\theta^*(\mathbf{x})$ prescribed under state $\mathbf{x} \in S$ is never to join facility i when $x_i = b_i$ $(i = 1, 2, \ldots, N)$. The policy θ^* described in Theorem 1 is obtained as a limit of the discount-optimal stationary policies $(\theta^*_{\eta_n})$. It follows that for every state $\mathbf{x} \in S$ there exists an integer $U(\mathbf{x})$ such that $\theta^*(\mathbf{x}) = \theta^*_{\eta_n}(\mathbf{x})$ for all $n \geqslant U(\mathbf{x})$, and therefore it suffices to show that for any discount rate $0 < \gamma < 1$, the discount-optimal policy θ^*_γ forbids joining facility i under states $\mathbf{x}$ with $x_i = b_i$. For a contradiction, suppose $x_i = b_i$ and $\theta^*_\gamma(\mathbf{x}) = 1$ for some state $\mathbf{x}$, facility i and discount rate γ. Then the discount optimality equations in (15) imply:

$$\hat{r}(\mathbf{x}, i) + \gamma\lambda v_\gamma^*(\mathbf{x}^{i+}, \hat{r}) \geqslant \gamma\lambda v_\gamma^*(\mathbf{x}, \hat{r}) \qquad (26)$$

that is, joining i is preferable to balking at state $\mathbf{x}$. Given that $x_i = b_i$, we have $\hat{r}(\mathbf{x}, i) < 0$ and therefore (26) implies $v_\gamma^*(\mathbf{x}^{i+}, \hat{r}) > v_\gamma^*(\mathbf{x}, \hat{r})$, but this contradicts the result of Lemma 3. □

Having shown that some socially optimal policy exists which induces a Markov chain with a positive recurrent class of states contained in $\tilde{S}$, we proceed to show that, in fact, *any* socially optimal policy has this property.

Lemma 6 *Any stationary policy θ^* which maximizes the long-run average reward defined in (2) induces an ergodic Markov chain on some set of states contained in $\tilde{S}$.*

Proof Suppose, for a contradiction, that we have a stationary policy θ which maximizes (2) and $\theta(\bar{\mathbf{x}}) = i$ for some state $\bar{\mathbf{x}} \in S$ with $\bar{x}_i = b_i$ and $\pi_\theta(\bar{\mathbf{x}}) > 0$. We proceed using a sample path argument. We start two processes at an arbitrary state $\mathbf{x}_0 \in S$ and apply policy θ to the first process, which follows path $\mathbf{x}(t)$. Let $(\mathbf{x}(t), t)$ denote the state-time of the system. Since θ is stationary, we may abbreviate $\theta(\mathbf{x}(t), t)$ to $\theta(\mathbf{x}(t))$. We also apply a *non-stationary* policy ϕ to the second process, which follows path $\mathbf{y}(t)$. The policy ϕ operates as follows: it chooses the same actions as θ at all times, *unless* the first process is in state $\bar{\mathbf{x}}$, in which case ϕ chooses to balk instead of joining facility i. In notation:

$$\phi(\mathbf{y}(t), t) = \begin{cases} \theta(\mathbf{x}(t)) & \text{if } \mathbf{x}(t) \neq \bar{\mathbf{x}}, \\ 0 & \text{otherwise} \end{cases}$$

Initially, $\mathbf{x}(0) = \mathbf{y}(0) = \mathbf{x}_0$. Let t_1 denote the first time, during the system's evolution, that the first process is in state $\bar{\mathbf{x}}$. At this point the process earns a *negative* reward $\hat{r}(\bar{\mathbf{x}}, i)$ by choosing action i; meanwhile, the second process earns a reward of zero by choosing to balk. An arrival may or may not occur at t_1; if it does, the first process acquires an extra customer, and if not, both processes remain in state $\bar{\mathbf{x}}$ (but nevertheless, due to the reward formulation in (3), the second process earns a greater reward at time t_1). Let u_1 denote the time of the next visit (after time t_1) of the first process to the regenerative state $\mathbf{0}$. In the interval $(t_1, u_1]$, the first process may acquire a certain number of extra customers at facility i (possibly more than one) in comparison to the second process due to further arrivals occurring under state $\bar{\mathbf{x}}$. Throughout the interval $(t_1, u_1]$, $\mathbf{x}(t)$ dominates $\mathbf{y}(t)$ in the sense that every facility has at least as many customers present under $\mathbf{x}(t)$ as under $\mathbf{y}(t)$. Consequently, at time u_1 or earlier, the processes are coupled again. At each of the time epochs $t_1 + 1, t_1 + 2, \ldots, u_1$ we note that the reward earned by the first process cannot possibly exceed the reward earned by the second process; this is because the presence of extra customers at facility i results in either a smaller reward (if facility i is chosen) or an equal reward (if a different facility, or balking, is chosen). Therefore the total reward earned by the first process up until time u_1 is smaller than that earned by the second process.

Using similar arguments, we can say that if t_2 denotes the time of the next visit (after u_1) of the first process to state $\bar{\mathbf{x}}$, the second process must earn a greater total reward than the first process in the interval $(t_2, u_2]$, where u_2 is the time of the next visit (after t_2) of the first process to state $\mathbf{0}$. Given that $\pi_\theta(\bar{\mathbf{x}}) > 0$, the state $\bar{\mathbf{x}}$ is visited infinitely often. Hence, by repetition of this argument, it is easy to see that θ is strictly inferior to the non-stationary policy ϕ in terms of expected long-run average reward. We know (by Theorem 1) that an optimal stationary policy exists, so there must be another stationary policy which is superior to θ. □

Theorem 1 may be regarded as a generalisation of a famous result which is due to Naor. In 1969, Naor shows (in the context of a single *M*/*M*/1 queue) that the selfishly optimal and socially optimal strategies are both threshold strategies, with thresholds n_s and n_o, respectively, and that $n_o \leqslant n_s$. This is the *M*/*M*/1 version of the containment property which we have proved for multiple, heterogeneous facilities (each with multiple service channels allowed). We also note that Theorem 1 assures us of being able to find a socially optimal policy by searching within the class of stationary policies which remain 'contained' in the finite set $\tilde{S}$. This means that we can apply the established techniques of dynamic programming (eg, value iteration, policy improvement) by restricting the state space so that it only includes states in $\tilde{S}$; any policy that would take us outside $\tilde{S}$ can be ignored, since we know that such a policy would be sub-optimal. For example, when implementing value iteration, we loop over all states in $\tilde{S}$ on each iteration and simply restrict the set of actions so that joining facility i is not allowed at any state $\mathbf{x}$ with $x_i = b_i$. This 'capping' technique enables us to avoid the use of alternative techniques which have been proposed in the literature for searching for optimal policies on infinite state spaces (see, eg, the method of 'approximating sequences' proposed by Sennott (1991), or Ha's (1997) method of approximating the limiting behaviour of the value function).

4. Comparison with unobservable systems

The results proved in Section 3 bear certain analogies to results which may be proved for systems of *unobservable* queues, in which routing decisions are made independently of the state of the system. In this section we briefly discuss the case of unobservable queues, in order to draw comparisons with the observable case. Comparisons between selfishly and socially optimal policies in unobservable queueing systems have already received considerable attention in the literature (see, eg, Littlechild, 1974; Edelson and Hildebrand, 1975; Bell and Stidham, 1983; Haviv and Roughgarden, 2007; Knight and Harper, 2013).

Consider a multiple-facility queueing system with a formulation identical to that given in Section 2, but with the added stipulation that the action a_n chosen at time step n must be selected independently of the system state $\mathbf{x}_n$. In effect, we assume that the system state is hidden from the decision-maker. Furthermore, the decision-maker lacks the ability to 'guess' the state of the system based on the waiting times of customers who have already passed through the system, and must simply assign customers to facilities according to a vector of routing probabilities $(p_1, p_2, \ldots, p_N)$ which remains constant over time. We assume that $\Sigma_{i=1}^{N} p_i \leqslant 1$, where p_i is the probability of routing a customer to facility i. Hence, $p_0 := 1 - \Sigma_{i=1}^{N} p_i$ is the probability that a customer will be rejected.

Naturally, the arrival process at facility $i \in \{1, 2, \ldots, N\}$ under a randomized admission policy is a Poisson process with demand rate $\lambda_i := \lambda p_i$, where (as before) λ is the demand rate for the system as a whole. Let $g_i(\lambda_i)$ denote the expected average net reward per unit time at facility i, given that it operates with a Poisson arrival rate λ_i. Then:

$$g_i(\lambda_i) = \lambda_i \alpha_i - \beta_i L_i(\lambda_i) \tag{27}$$

where $L_i(\lambda_i)$ is the expected number of customers present at i under demand rate λ_i. In this context, a socially optimal policy is a vector $(\lambda^*_1, \lambda^*_2, \ldots, \lambda^*_N)$ which maximizes the sum $\Sigma_{i=1}^{N} g_i(\lambda^*_i)$. On the other hand, a *selfishly* optimal policy is a vector $(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_N)$ which causes the system to remain in equilibrium, in the sense that no self-interested customer has an incentive to deviate from the randomized policy in question (see Bell and Stidham, 1983, p 834). More specifically, individual customers make decisions according to a probability distribution $\{\tilde{p}_i\}$ (where $\lambda \tilde{p}_i = \tilde{\lambda}_i$ for each $i \in \{1, 2, \ldots, N\}$) and, in order for equilibrium to be maintained, it is necessary for all of the actions chosen with non-zero probability to yield the same expected net reward.

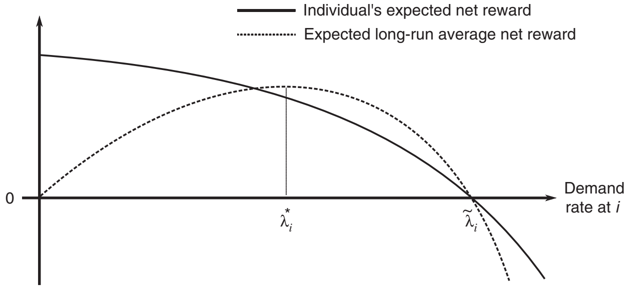


Figure 4 The general shapes of $w_i(\lambda_i)$ and $g_i(\lambda_i)$ as functions of λ_i.

First of all, it is worth making the point that no theoretical upper bound exists for the number of customers who may be present at any individual facility i under a Poisson demand rate λ_i which is independent of the system state (unless, of course, $\lambda_i = 0$). Indeed, standard results for *M*/*M*/*c* queues (see Gross and Harris, 1998, p 69) imply that the steady-state probability of n customers being present at a facility with a positive demand rate is positive for each $n \geqslant 0$. As such, the positive recurrent state spaces under the selfishly and socially optimal policies are both unbounded in the unobservable case, and there is no prospect of being able to prove a 'containment' result similar to that of Theorem 2. However, it is straightforward to prove an alternative result involving the total effective admission rates under the two policies which is consistent with the general theme of socially optimal policies generating 'less busy' systems than their selfish counterparts.

Figure 4 illustrates the general shapes of the expected net reward for an individual customer (henceforth denoted $w_i(\lambda_i)$) and the expected long-run average reward $g_i(\lambda_i)$ as functions of the Poisson queue-joining rate λ_i at an individual facility i. Naturally, $w_i(\lambda_i)$ is a strictly decreasing function of λ_i and, assuming that the demand rate for the system is sufficiently large, the joining rate at facility i under an equilibrium (selfish) policy is the unique value $\tilde{\lambda}_i$ which equates $w_i(\lambda_i)$ to zero. Indeed, if this were not the case, then a selfish customer would deviate from the equilibrium policy by choosing to join the queue with probability 1 (if $w_i(\lambda_i)$ was positive) or balk with probability 1 (if $w_i(\lambda_i)$ was negative). On the other hand, it is known from the queueing theory literature (see Grassmann, 1983; Lee and Cohen, 1983) that the expected queue length $L_i(\lambda_i)$ is a strictly convex function of λ_i, and hence the function $g_i(\lambda_i)$ in (27) is strictly concave in λ_i. Under a socially optimal policy, the joining rate at facility i is the unique value λ^*_i which maximizes $g_i(\lambda_i)$ (assuming, once again, that the demand rate for the system as a whole is large enough to permit this flow of traffic at facility i).

It is worth noting that the theory of non-atomic routing games (see Roughgarden, 2005) assures us that the equilibrium and socially optimal policies both exist and are unique. This allows a simple argument to be formed in order to show that the sum of the joining rates at the individual facilities under a socially optimal policy (let us denote this by η^*) cannot possibly exceed the corresponding sum under an equilibrium policy (denoted $\tilde{\eta}$). Indeed, it is clear that if the system demand rate λ is sufficiently large, then the selfishly and socially optimal joining rates at any individual facility i will attain their 'ideal' values $\tilde{\lambda}_i$ and $\lambda^*{}_i$ (as depicted in Figure 4), respectively, and so in this case it follows trivially that $\eta^* \leqslant \tilde{\eta}$. On the other hand, suppose that λ is *not* large enough to permit $w_i(\lambda_i) = 0$ for all facilities i. In this case, $w_i(\lambda_i)$ must be strictly positive for some facility i, and therefore the probability of a customer balking under an equilibrium strategy is zero (since balking is unfavourable in comparison to joining facility i). Hence, one has $\tilde{\eta} = \lambda$ in this case, and since η^* is also bounded above by λ the result $\eta^* \leqslant \tilde{\eta}$ follows.

The conclusion of this section is that, while the 'containment' property of observable systems proved in Section 3 does not have an exact analogue in the unobservable case, the general principle that selfish customers create 'busier' systems still persists (albeit in a slightly different guise).

5. Heterogeneous customers

An advantage of the anticipatory reward formulation in (3) is that it enables the results from Section 3 to be extended to a scenario involving heterogeneous customers without a re-description of the state space S being required. Suppose we have $M \geqslant 2$ *customer classes*, and customers of the ith class arrive in the system via their own independent Poisson process with demand rate λ_i ($i = 1, 2, .., M$). In this case we assume, without loss of generality, that $\sum_i \lambda_i + \sum_j c_j \mu_j = 1$. For convenience we will define $\lambda := \sum_i \lambda_i$ as the total demand rate. We allow the holding costs and fixed rewards in our model (but *not* the service rates) to depend on these customer classes; that is, the fixed reward for serving a customer of class i at facility j is now α_{ij}, and the holding cost (per unit time) is β_{ij}. Various physical interpretations of this model are possible; for example, suppose we have a healthcare system in which patients arrive from various different geographical locations. Then the parameters α_{ij} and β_{ij} may be configured according to the distance of service provider j from region i (among other factors), so that patients' commuting costs are taken into account.

Suppose we wanted to use a 'real-time' reward formulation, similar to (1), for the reward $r(\cdot)$ in our extended model. Then the system state would need to include information about the classes of customers in service at each facility, and also the total number of customers of each class waiting in each queue. However, using an 'anticipatory' reward formulation, we can allow the state space representation to be the same as before; that is, $S = \{\mathbf{x} = (x_1, x_2, ..., x_N) : x_1, x_2, ..., x_N \in \mathbb{N}_0\}$, where x_j is simply the number of customers present (irrespective of class) at facility j, for $j = 1, 2, \ldots, N$. On the other hand, the set of actions A available at each state $\mathbf{x} \in S$ has a more complicated representation. We now define A as follows:

$$A = \{\mathbf{a} = (a_1, a_2, ..., a_M) : a_1, a_2, ..., a_M \in \{0, 1, ..., N\}\}$$

That is, the action $\mathbf{a}$ is a vector which prescribes, for each customer class $i \in \{1, 2, \ldots, M\}$, the destination a_i of any customer of class i who arrives at the present epoch of time, with the system having been *uniformized* so that it evolves in discrete time steps of size $\Delta = (\sum_i \lambda_i + \sum_j c_j \mu_j)^{-1}$. The reward $\hat{r}(\mathbf{x}, \mathbf{a})$ for taking action $\mathbf{a} \in A$ at state $\mathbf{x}$ is then:

$$\hat{r}(\mathbf{x}, \mathbf{a}) = \sum_{i=1}^{M} \hat{r}_i(\mathbf{x}, a_i)$$

where a_i is the i^{th} component of $\mathbf{a}$, and:

$$\hat{r}_i(\mathbf{x}, a_i) := \begin{cases} \lambda_i\left(\alpha_{ij} - \frac{\beta_{ij}}{\mu_j}\right), & a_i = j \neq 0,\ x_j < c_j \\ \lambda_i\left(\alpha_{ij} - \frac{\beta_{ij}(x_j+1)}{c_j \mu_j}\right), & a_i = j \neq 0,\ x_j \geqslant c_j \\ 0, & a_i = 0 \end{cases}$$

for $i = 1, 2 \ldots, M$. We note that expanding the action set in this manner is not the only possible way of formulating our new model (with heterogeneous customers) as an MDP, but it is the natural extension of the formulation adopted in the previous section. An alternative approach would be to augment the state space so that information about the class of the most recent customer to arrive is included in the state description; actions would then need to be chosen only at arrival epochs, and these actions would simply be integers in the set $\{0, 1, \ldots, N\}$ as opposed to vectors (see Puterman, 1994, p 568) for an example involving admission control in an *M*/*M*/1 queue). By keeping the state space S unchanged, however, we are able to show that the results of Section 3 can be generalized very easily.

Under our new formulation, the discount optimality equations (using the anticipatory reward functions $\hat{r}_i$) are as follows:

$$\begin{aligned} v_\gamma^*(\mathbf{x}, \hat{r}) = & \sum_{i=1}^{M} \max_{a_i \in A}\left\{\hat{r}_i(\mathbf{x}, a_i) + \gamma \lambda_i v_\gamma^*(\mathbf{x}^{a_i+}, \hat{r})\right\} \\ & + \gamma\left[\sum_{j=1}^{N} \min(x_j, c_j)\mu_j v_\gamma^*(\mathbf{x}^{j-}, \hat{r})\right. \\ & \left. + \left(1 - \lambda - \sum_{j=1}^{N} \min(x_j, c_j)\mu_j\right) v_\gamma^*(\mathbf{x}, \hat{r})\right] \end{aligned} \tag{28}$$

Note that the maximization in (28) can be carried out in a componentwise fashion, so that instead of having to find the maximizer among *all* vectors $\mathbf{a}$ in A (of which the total number is $(N+1)^M$), we can simply find, for each customer class $i \in M$, the 'marginal' action a_i which maximizes $\hat{r}_i(\mathbf{x}, a_i) + \gamma\lambda_i v_\gamma^*(\mathbf{x}^{a_i+}, \hat{r})$. This can be exploited in the implementation of dynamic programming algorithms (eg, value iteration), so that computation times increase only in proportion to the number of customer classes M.

As before, we define the *selfishly optimal policy* to be the policy under which the action chosen for each customer arriving in the system is the action which maximizes $\hat{r}_i(\mathbf{x}, a_i)$ (obviously this action now depends on the customer class). A selfish customer of class i accepts service at facility j if and only if, prior to joining, the number x_j of customers at facility j, satisfies $x_j \leqslant b_{ij}$, where:

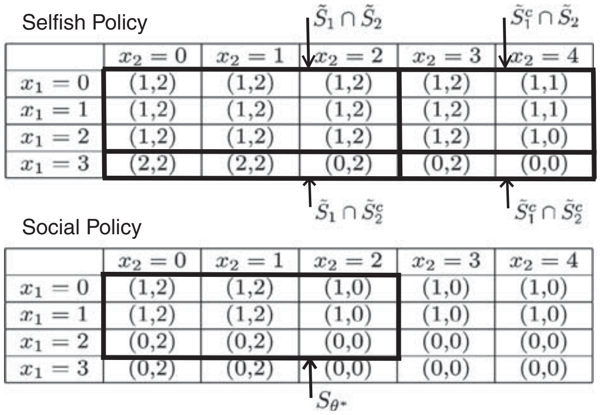
$$b_{ij} := \left\lfloor \frac{c_j \alpha_{ij} \mu_j}{\beta_{ij}} \right\rfloor$$

Consequently, under steady-state conditions, the number of customers present at facility j is bounded above by $\max_i b_{ij}$. It follows that we now have the following definition for the selfishly optimal state space $\tilde{S}$:

$$\tilde{S} := \left\{ (x_1, x_2, \ldots, x_N) : x_j \leqslant \max_i b_{ij},\ j = 1, 2, \ldots, N \right\}$$

Example 2 *We modify Example 1 from earlier so that there are now two classes of customer, with demand rates $\lambda_1 = 12$ and $\lambda_2 = 10$ respectively. The first class has the same cost and reward parameters as in Example 1; that is, $\beta_{11} = 3$, $\alpha_{11} = 1$ (for the first facility) and $\beta_{12} = 3$, $\alpha_{12} = 3$ (for the second facility). The second class of customer has steeper holding costs and a much greater value of service at the second facility: $\beta_{21} = \beta_{22} = 5$, $\alpha_{21} = 1$, $\alpha_{22} = 12$. Both facilities have two service channels and the service rates $\mu_1 = 5$ and $\mu_2 = 1$ remain independent of customer class. We take $\Delta = (\sum_i \lambda_i + \sum_j c_j \mu_j)^{-1} = 1/34$ to uniformize the system.*

We have previously seen that customers of the first class acting selfishly will cause the system state to remain within a set of 12 states under steady-state conditions, with $x_1 \leqslant 3$ and $x_2 \leqslant 2$ at all times. The incorporation of a second class of customer has no effect on the selfish decisions made by the first class of customer, so (as shown in Figure 5) these decisions remain the same as shown in Figure 3 previously. The first table in Figure 5 shows that selfish customers of the second class are unwilling to join the first facility when $x_1 \geqslant 2$; however, under certain states they will choose to join the second facility when $x_2 = 3$ (but never when $x_2 > 3$). As a result, the selfish state space $\tilde{S}$ is expanded from 12 states to 20.


Selfish Policy
$\tilde{S}_1 \cap \tilde{S}_2$
$\tilde{S}_1^c \cap \tilde{S}_2$
$\tilde{S}_1 \cap \tilde{S}_2^c$
$\tilde{S}_1^c \cap \tilde{S}_2^c$

	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 0$	(1,2)	(1,2)	(1,2)	(1,2)	(1,1)
$x_1 = 1$	(1,2)	(1,2)	(1,2)	(1,2)	(1,1)
$x_1 = 2$	(1,2)	(1,2)	(1,2)	(1,2)	(1,0)
$x_1 = 3$	(2,2)	(2,2)	(0,2)	(0,2)	(0,0)

Social Policy
S_{θ^*}

	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 0$	(1,2)	(1,2)	(1,0)	(1,0)	(1,0)
$x_1 = 1$	(1,2)	(1,2)	(1,0)	(1,0)	(1,0)
$x_1 = 2$	(0,2)	(0,2)	(0,0)	(0,0)	(0,0)
$x_1 = 3$	(0,2)	(0,2)	(0,0)	(0,0)	(0,0)

Figure 5 Selfishly and socially optimal policies for Example 2.
Note: At each state $\mathbf{x} = (x_1, x_2)$, the corresponding decision vector $\mathbf{a} = (a_1, a_2)$ is shown.

Figure 5 shows that the new selfish state space $\tilde{S}$ may be represented diagrammatically as the smallest rectangle which encompasses both $\tilde{S}_1$ and $\tilde{S}_2$, where (for $i = 1, 2$) we have defined:

$$\tilde{S}_i := \left\{ (x_1, x_2, \ldots, x_N) : x_j \leqslant b_{ij},\ j = 1, 2, \ldots, N \right\}$$

It is somewhat interesting to observe that $\tilde{S}$ includes states in the 'intersection of complements' $\tilde{S}_1^c \cap \tilde{S}_2^c$. These states would not occur (under steady-state conditions) if the system operated with only a single class of customer of either type, but they do occur with both customer types present.

The policy θ^* depicted in the second table in Figure 5 has been obtained using value iteration, and illustrates the containment property for systems with heterogeneous customer classes. It may easily be seen that the socially optimal state space S_{θ^*} consists of only 9 states; under steady-state conditions, the system will remain within this smaller class of states. We also observe that, unlike the selfish decisions, the socially optimal decisions for a particular class of customer are affected by the decisions made by the other class of customer (as can be seen, in the case of the first customer class, by direct comparison with Figure 3 from Example 1). Indeed, under θ^*, customers of the first class never join Facility 2, and customers of the second class never join Facility 1.

It can be verified that the results of Lemmas 2–5, Theorems 1–2 and Lemma 6 apply to the model with heterogeneous customers, with only small modifications required to the proofs. For example, in Lemma 3 we prove the inequality $v_\gamma^*(\mathbf{x}^{j+}, \hat{r}) \leqslant v_\gamma^*(\mathbf{x}, \hat{r})$ by showing that, *for all classes* $i \in \{1, 2, \ldots, M\}$ and facilities $j \in \{1, 2, \ldots, N\}$, we have:

$$\max_{a_i} \left\{ \hat{r}_i\left(\mathbf{x}^{j+}, a_i\right) + \gamma\lambda_i v_{\gamma,k}^*\left(\left(\mathbf{x}^{j+}\right)^{a_i+}, \hat{r}\right) \right\}$$

$$\leqslant \max_{b_i} \left\{ \hat{r}_i(\mathbf{x}, b_i) + \gamma\lambda_i v_{\gamma,k}^*\left(\mathbf{x}^{b_i+}, \hat{r}\right) \right\}$$

for all $k \in \mathbb{N}_0$. In Lemma 4, we can define $\alpha_{\max} = \max_{i,j} \alpha_{ij}$ and establish a lower bound for $w_\gamma^*(\mathbf{0}^{j+}, \hat{r}) - w_\gamma^*(\mathbf{0}, \hat{r})$ similar to (22) by writing $\hat{r}(\mathbf{0}^{j+}, \mathbf{0})$ and $p_{\mathbf{0}^{j+},\mathbf{y}}(\mathbf{0})$ instead of $\hat{r}(\mathbf{0}^{j+}, 0)$ and $p_{\mathbf{0}^{j+},\mathbf{y}}(0)$ respectively (so that the action at state $\mathbf{0}^{j+}$ is the zero vector $\mathbf{0}$, ie all customer classes balk); the rest of the inductive proof goes through using similar adjustments. Theorem 2 holds because if S_{θ^*} was *not* contained in $\tilde{S}$, then the discount optimality equations would imply, for some $i \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, N\}$:

$$\hat{r}_i(\mathbf{x}, j) + \gamma\lambda_i v_\gamma^*(\mathbf{x}^{j+}, \hat{r}) \geqslant \gamma\lambda_i v_\gamma^*(\mathbf{x}, \hat{r})$$

with $\hat{r}_i(\mathbf{x}, j) < 0$, thus contradicting the result of (the modified) Lemma 3. The sample path argument in Lemma 6 can be applied to a customer of any class, with only trivial adjustments needed.

6. Conclusions

The principle that selfish users create busier systems is well-observed in the literature on behavioural queueing theory. While this principle is interesting in itself, we also believe that it has the potential to be utilized much more widely in applications. As we have demonstrated, the search for a socially optimal policy may be greatly simplified by reducing the search space according to the bounds of the corresponding ‘selfish’ policy, so that the methods of dynamic programming can be more easily employed.

Our results in this paper hold for an arbitrary number of facilities *N*, and (in addition) the results in Section 5 hold for an arbitrary number of customer classes *M*. This lack of restriction makes the results powerful from a theoretical point of view, but we must also point out that in practice, the ‘curse of dimensionality’ often prohibits the exact computation of optimal policies in large-scale systems, even when the state space can be assumed finite. This problem could be partially addressed if certain structural properties (eg monotonicity properties) of socially optimal policies could be proved with the same level of generality as our ‘containment’ results. It can be shown trivially that selfish policies are monotonic in various respects (eg, balking at the state $\mathbf{x}$ implies balking at state $\mathbf{x}^{j+}$, for any facility *j*) and, indeed, the optimality of monotone policies is a popular theme in the literature, although in our experience these properties are usually not trivial to prove for an arbitrary number of facilities. In future work, we intend to investigate how the search for socially optimal policies can be further simplified by exploiting their theoretical structure.

References

Argon NT, Ding L, Glazebrook KD and Ziya S (2009). Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences* **23**(2): 175–203.

Bell CE and Stidham S (1983). Individual versus social optimization in the allocation of customers to alternative servers. *Management Science* **29**(7): 831–839.

Bellman RE (1957). *Dynamic Programming*. Princeton University Press: New Jersey.

Cavazos-Cadena R (1989). Weak conditions for the existence of optimal stationary policies in average markov decision chains with unbounded costs. *Kybernetika* **25**(3): 145–156.

Cinlar E (1975). *Introduction to Stochastic Processes*. Prentice-Hall: Englewood Cliffs, NJ.

Economou A and Kanta S (2008). Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space. *Queueing Systems* **59**(3): 237–269.

Edelson NM and Hildebrand DK (1975). Congestion tolls for poisson queuing processes. *Econometrica* **43**(1): 81–92.

Glazebrook KD, Kirkbride C and Ouenniche J (2009). Index policies for the admission control and routing of impatient customers to heterogeneous service stations. *Operations Research* **57**(4): 975–989.

Grassmann W (1983). The convexity of the mean queue size of the M/M/c queue with respect to the traffic intensity. *Journal of Applied Probability* **20**(4): 916–919.

Gross D and Harris C (1998). *Fundamentals of Queueing Theory*. John Wiley & Sons: New York.

Guo P and Li Q (2013). Strategic behavior and social optimization in partially-observable Markovian vacation queues. *Operations Research Letters* **41**(3): 277–284.

Ha A (1997). Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research* **45**(1): 42–53.

Haviv M and Roughgarden T (2007). The price of anarchy in an exponential multi-server. *Operations Research Letters* **35**(4): 421–426.

Knight VA and Harper PR (2013). Selfish routing in public services. *European Journal of Operational Research* **230**(1): 122–132.

Knight VA, Williams JE and Reynolds I (2012). Modelling patient choice in healthcare systems: Development and application of a discrete event simulation with agent-based decision making. *Journal of Simulation* **6**(2): 92–102.

Knudsen NC (1972). Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* **40**(3): 515–528.

Lee HL and Cohen AM (1983). A note on the convexity of performance measures of *M/M/c* Queueing systems. *Journal of Applied Probability* **20**(4): 920–923.

Lippman SA (1975). Applying a new device in the optimisation of exponential queueing systems. *Operations Research* **23**(4): 687–710.

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

Littlechild SC (1974). Optimal arrival rate in a simple queueing system. *International Journal of Production Research* **12**(3): 391–397.

Naor P (1969). The regulation of queue size by levying tolls. *Econometrica* **37**(1): 15–24.

Puterman ML (1994). *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. Wiley & Sons: New York.

Ross SM (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press: New York.

Roughgarden T (2005). *Selfish Routing and the Price of Anarchy*. MIT Press: Cambridge, MA.

Sennott LI (1989). Average cost optimal stationary policies in infinite state markov decision processes with unbounded costs. *Operations Research* **37**(4): 626–633.

Sennott LI (1991). Value iteration in countable state average cost markov decision processes with unbounded costs. *Annals of Operations Research* **28**(1): 261–272.

Serfozo R (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research* **27**(3): 616–620.

Shone R, Knight VA and Williams JE (2013). Comparisons between observable and unobservable *M/M/*1 queues with respect to optimal customer behavior. *European Journal of Operational Research* **227**(1): 133–141.

Stidham S (1978). Socially and individually optimal control of arrivals to a GI/M/1 queue. *Management Science* **24**(15): 1598–1610.

Stidham S and Weber RR (1993). A survey of Markov decision models for control of networks of queues. *Queueing Systems* **13**(1): 291–314.

Sun W, Guo P and Tian N (2010). Equilibrium threshold strategies in observable queueing systems with setup/closedown times. *Central European Journal of Operations Research* **18**(3): 241–268.

Wang J and Zhang F (2011). Equilibrium analysis of the observable queues with balking and delayed repairs. *Applied Mathematics and Computation* **218**(6): 2716–2729.

Yechiali U (1971). On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research* **19**(2): 349–370.

Yechiali U (1972). Customers' optimal joining rules for the GI/M/s queue. *Management Science* **18**(7): 434–443.

Zijm H (1985). The optimality equations in multichain denumerable state markov decision proceses with the average cost criterion: The bounded cost case. *Statistics and Decisions* **3**(1): 143–165.

Received 25 April 2014;
accepted 13 October 2015 after two revisions

This work is licensed under a Creative Commons Attribution 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/