Mathematical Methods of Operations Research
https://doi.org/10.1007/s00186-020-00722-w

ORIGINAL ARTICLE

A conservative index heuristic for routing problems with multiple heterogeneous service facilities

Rob Shone[1] · Vincent A. Knight[2] · Paul R. Harper[2]

Received: 10 October 2019 / Revised: 22 June 2020

Abstract

We consider a queueing system with N heterogeneous service facilities, in which admission and routing decisions are made when customers arrive and the objective is to maximize long-run average net rewards. For this type of problem, it is well-known that structural properties of optimal policies are difficult to prove in general and dynamic programming methods are computationally infeasible unless N is small. In the absence of an optimal policy to refer to, the Whittle index heuristic (originating from the literature on multi-armed bandit problems) is one approach which might be used for decision-making. After establishing the required indexability property, we show that the Whittle heuristic possesses certain structural properties which do not extend to optimal policies, except in some special cases. We also present results from numerical experiments which demonstrate that, in addition to being consistently strong over all parameter sets, the Whittle heuristic tends to be more robust than other heuristics with respect to the number of service facilities and the amount of heterogeneity between the facilities.

Keywords Queueing systems · Dynamic programming · Whittle index heuristic

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00186-020-00722-w) contains supplementary material, which is available to authorized users.

✉ Rob Shone
r.shone@lancaster.ac.uk

Vincent A. Knight
knightva@cardiff.ac.uk

Paul R. Harper
harper@cardiff.ac.uk

[1] Lancaster University, Bailrigg, Lancaster, UK

[2] Cardiff University, Cardiff, UK

Published online: 22 July 2020

1 Introduction

Many routing problems involving multiple queueing facilities can be mathematically formulated as Markov decision processes (MDPs) and solved exactly using well-known dynamic programming (DP) techniques such as value iteration and policy improvement (Bellman 1957; Howard 1960; Puterman 1994). Unfortunately, these techniques are usually of no practical use for solving problems which are modelled on real-world applications. The size and complexity of the state space that one must consider has been cited as one of the *curses of dimensionality* which usually impede exact solution attempts (Sutton and Barto 1998; Powell 2007).

To address this problem, there are various strategies that one might employ. Depending on the particular features of the problem under consideration, it may be possible to simplify the search by proving that optimal policies must have certain characteristics. However, even if this is possible, one may need to resort to the use of *heuristics* which can be relied upon to find near-optimal policies in a time-efficient manner. In this paper we discuss the use of 'index-based' heuristics and consider their application to a problem involving a Markovian queueing system with heterogeneous service facilities arranged in parallel, each with its own queue and multiple servers available.

We consider systems which can be modelled as shown in Fig. 1. Customers arriving at the 'routing point' must either be sent to one of the N service facilities, in which case they wait in the queue for that facility until a server becomes available, or rejected without receiving service. A fixed, facility-dependent reward is earned each time a customer is served, but holding costs are also incurred and depend linearly on the waiting times of customers. A detailed formulation is provided in Sect. 2.

The fact that we consider heterogeneous facilities, each with the ability to serve multiple customers at once, makes public service systems a natural application area for our work. In the context of healthcare systems, for example, a patient seeking an elective knee operation might be faced with a choice of different treatment providers which differ from each other with respect to the quality of treatment provided, the number of beds available, the expected length of stay and other factors. It is well-known that for systems such as these, the performance of the system (measured, for example, by the long-run average net reward per unit time after deducting holding costs) is not optimized by allowing customers to make 'selfish' decisions based only on their own interests (Bell and Stidham 1983; Hassin and Haviv 2003; Haviv and Roughgarden 2007; Knight and Harper 2013; Shone et al. 2016; Knight et al. 2017). Instead, customers must be directed to follow an optimal (sometimes referred to as a 'socially optimal') policy which takes into account the effects of decisions on customers who have yet to arrive. It is the need to find, or approximate, a socially optimal policy that we focus on in this paper.

For routing problems involving multiple service facilities in parallel, optimal routing policies do not necessarily have simple characterizations. "Join the Shortest Queue" (JSQ) rules often apply to systems in which all facilities are identical (see Winston 1977; Weber 1978; Hordijk and Koole 1990; Menich and Serfozo 1991; Koole et al. 1999), but even in these cases an optimal policy may have counter-intuitive properties (Whitt 1986). In the case of heterogeneous facilities, some compelling results have been obtained for single-server queues (Hordijk and Koole 1992; Ha 1997), but

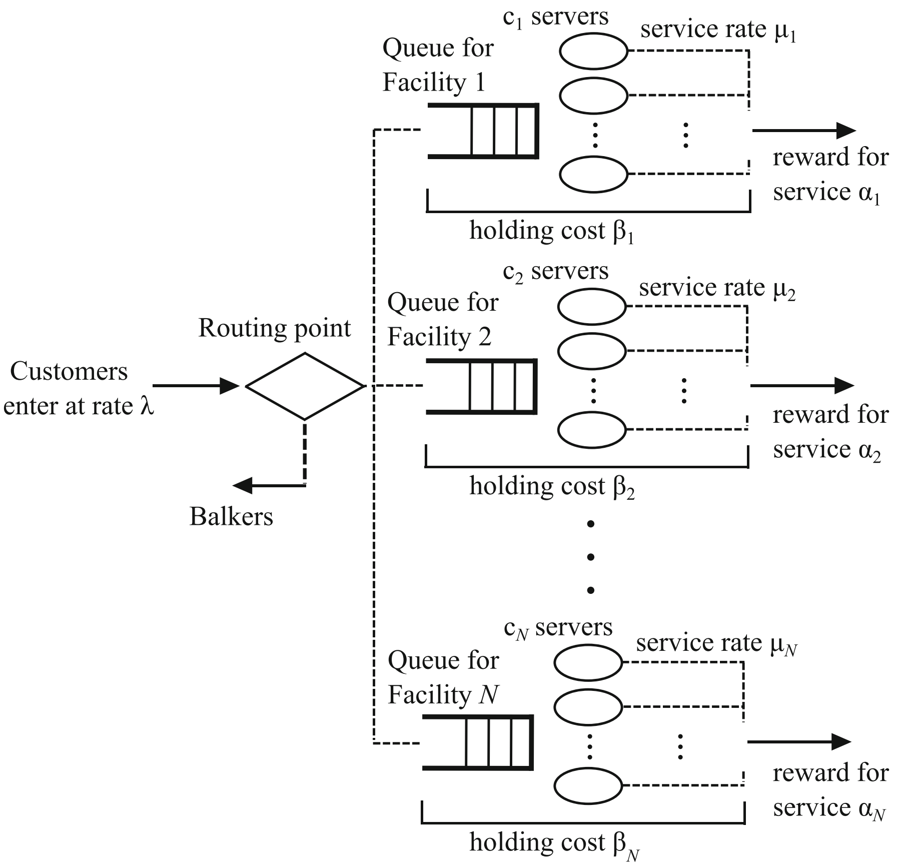

Fig. 1 A diagrammatic representation of the queueing system

in general the analysis is much more difficult. Consequently, previous research has tended to focus on developing heuristic (sub-optimal) routing policies. For example, policies based on applying one step of policy iteration to a 'Bernoulli splitting' policy have been shown to perform strongly in various contexts (Krishnan 1990; Sassen et al. 1997; Ansell et al. 2003a; Bhulai and Koole 2003; Argon et al. 2009; Hyytia et al. 2012).

The heuristic policy to be developed in this paper has its origins in the work of Gittins (1979) and Whittle (1988) on deriving 'dynamic allocation indices' for multi-armed bandit processes. Informally speaking, an *index-based* policy is a policy which associates a certain, easily-computable score or *index* to the various possible decision options in any given system state, and then chooses the option with the highest index. In order for an index-based policy to be applicable to a particular problem, the property of *indexability* must first be established. Although this property is not trivial to prove in general, it has been proven to hold in various problems involving queueing or inventory control (see Ansell et al. 2003b; Archibald et al. 2009; Glazebrook et al. 2009, 2011, 2014; Hodge and Glazebrook 2011; Nino-Mora 2012; Larranaga et al. 2016).

The *modus operandi* of the particular type of index policy that we focus on in this paper, which we shall refer to as the *Whittle index heuristic*, was first described in general terms by Whittle (1988) and has been shown to yield strong-performing policies in various types of problems involving dynamic resource allocation. Nino-Mora (2002) studied a similar problem to ours, with the evolution of 'service facilities' described by birth–death processes, and derived general expressions for the indices upon which decisions are based. Argon et al. (2009) also considered a routing problem in which customers are routed to single-server facilities. Their model does not include admission control, but does include more general cost structures and different customer types. Some recent applications for the Whittle heuristic in the literature include egalitarian processor sharing systems (Borkar and Pattathil 2017), scheduling for time-varying channels (Aalto et al. 2016), partially observed binary Markov chains (Borkar 2017), scheduling of information transmissions (Hsu 2018), allocation of scarce resources to a large number of jobs (Li et al. 2020) and allocation of assets prone to failure (Ford et al. 2020).

In the context of the routing problem that we consider in this paper, an advantage of using a Whittle index heuristic is that we are able to show that the resulting index policies possess certain intuitively 'nice' structural properties which usually cannot be proven to hold for optimal policies. Since the heuristic policies appear to perform very strongly in numerical experiments, this provides justification for searching for optimal policies within a reduced class of solutions which possess these 'nice' properties (this is a topic of ongoing work). The main contributions of our paper are as follows:

- We confirm (Sect. 3) that the service facilities in our problem are indexable and derive expressions for the Whittle indices in terms of the system parameters.
- We show (Sect. 4) that the positive recurrent state space under an arbitrary optimal stationary policy is bounded between two finite sets, one of which is derived using the Whittle indices.
- We show (Sect. 4) that the Whittle heuristic policy becomes asymptotically optimal in light-traffic and heavy-traffic limits.
- We prove (Sect. 4) that the Whittle heuristic policy possesses various other 'intuitive' structural properties and provide counter-examples to show that these may not hold in general for optimal policies.
- We introduce (Sect. 5) an alternative index-based heuristic based on one-step policy improvement, and show that it possesses asymptotic light-traffic (but not heavy-traffic) optimality.
- We present (Sect. 6) results from numerical experiments to compare the performance of the Whittle heuristic policy with those of optimal policies (where feasible) and alternative heuristic policies.

In the next section we provide a detailed formulation of the multiple-facility routing problem considered throughout this paper.

2 Problem formulation

We consider a queueing system with N service facilities, each of which has its own queue and a first-come–first-served (FCFS) queue discipline. Customers arrive from outside the system according to a Poisson process with demand rate $\lambda > 0$. Newly-arrived customers may proceed to any one of the N service facilities or, alternatively, exit the system immediately without incurring any cost or reward (referred to as *balking*). Thus, there are $N + 1$ possible destinations for any individual customer. An individual facility $i \in \{1, 2, \dots, N\}$ possesses c_i identical service channels, and service times at any channel of facility i are exponentially distributed with mean $\mu_i^{-1} > 0$.

The system earns a *fixed reward* $\alpha_i > 0$ for each customer who completes service at facility $i \in \{1, 2, \dots, N\}$, but also incurs a *linear holding cost* $\beta_i > 0$ per unit time for each customer waiting at the facility (whether in the queue or in service). We also assume that $\alpha_i > \beta_i/\mu_i$ for each $i \in \{1, 2, \dots, N\}$ in order to ensure that rewards adequately compensate customers for their own expected service costs (otherwise we would have redundancy in the system).

The system is *fully observable*, in the sense that the number of customers present at each facility is always known and can be used to inform decision-making. We do not make any assumptions as to whether decisions are made by customers themselves or whether they are directed by a central controller, since this depends on the physical context of the problem and does not affect our results in this paper. If customers make their own decisions, then an optimal policy can be regarded as that which arises from customers co-operating with each other to maximize their collective welfare.

Let $S = \{(x_1, x_2, \dots, x_N) : x_i \in \mathbb{N}_0 \text{ for each } i \in \{1, 2, \dots, N\}\}$ denote the state space of the system, where x_i is the number of customers present at facility i (including those being served). A system state is denoted by a vector $\mathbf{x} \in S$. We also use $\mathbf{x}^{i+}$ (resp. $\mathbf{x}^{i-}$) to denote the state which is identical to $\mathbf{x}$ except that one extra customer (resp. one fewer customer) is present at facility i. That is:

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

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

The sum of the infinitesimal transition rates under any state $\mathbf{x} \in S$ is bounded above by $\lambda + \sum_{i=1}^{N} c_i \mu_i$. We can therefore apply the process of uniformization (Lippman 1975; Serfozo 1979) to formulate the system as a discrete-time Markov Decision Process (MDP) in which the action space, transition probabilities and single-step reward function are defined as follows:

- Under any state $\mathbf{x} \in S$, the action space is

$$A = \{0, 1, 2, \dots, N\}.$$

  An action $a \in A$ represents the decision of the next customer to arrive in the system. If $a = 0$ then the customer balks, and if $a = i$ for some $i \in \{1, 2, \dots, N\}$ then the customer goes to facility i.

- The transition probability of going from state $\mathbf{x} \in S$ to state $\mathbf{y} \in S \setminus \{\mathbf{x}\}$ in a single discrete time step, given that action $a \in A$ has been chosen, is denoted by $p(\mathbf{x}, a, \mathbf{y})$, where

$$p(\mathbf{x}, a, \mathbf{y}) = \begin{cases} \lambda\Delta, & \text{if } a = i \text{ for some } i \in \{1, 2, \ldots, N\} \text{ and } \mathbf{y} = \mathbf{x}^{i+}, \\ \min(x_i, c_i)\mu_i\Delta, & \text{if } \mathbf{y} = \mathbf{x}^{i-} \text{for some } i \in \{1, 2, \ldots, N\}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

  and $\Delta = \left(\lambda + \sum_{i=1}^{N} c_i\mu_i\right)^{-1}$ is the discrete-time step size. Note that, following uniformization, we assume that at most one random event (either an arrival or a service completion) may occur in a single discrete time step. A 'self-transition' from state $\mathbf{x} \in S$ to itself may occur if no random event takes place, and the relevant transition probability is

$$p(\mathbf{x}, a, \mathbf{x}) = 1 - I(a \neq 0)\lambda\Delta - \sum_{i=1}^{N} \min(x_i, c_i)\mu_i\Delta,$$

  where I denotes the indicator function.
- Under state $\mathbf{x} \in S$, rewards are being earned at a total rate of $\sum_{i=1}^{N} \alpha_i \min(x_i, c_i)\mu_i$ and holding costs are being incurred at a rate $\sum_{i=1}^{N} \beta_i x_i$. We therefore formulate the single-step reward function as

$$r(\mathbf{x}) = \sum_{i=1}^{N} \big(\alpha_i \min(x_i, c_i)\mu_i - \beta_i x_i\big). \tag{2}$$

For simplicity we will assume throughout this paper that $\Delta = \left(\lambda + \sum_{i=1}^{N} c_i\mu_i\right)^{-1} = 1$, since the units of time are arbitrary.

We define an optimal policy as a decision-making rule θ which maximizes the long-run average net reward per unit time, defined as

$$g^{\theta}(\mathbf{x}) = \lim_{t\to\infty} t^{-1}\mathbb{E}_{\theta}\left[\sum_{n=0}^{t-1} r(\mathbf{x}_n)|\mathbf{x}_0 = \mathbf{x}\right], \tag{3}$$

where $\mathbf{x}_n$ denotes the state of the system at the nth discrete time step (with $\mathbf{x}_0$ as the initial state).

Let $w_a(\mathbf{x})$ denote an individual customer's expected net reward for choosing action $a \in \{0, 1, \ldots, N\}$ under state $\mathbf{x} \in S$. If the action chosen is to join some facility $i \in \{1, \ldots, N\}$ at which x_i customers are already present, then the expected waiting time is $1/\mu_i$ if $x_i < c_i$, and $(x_i - c_i + 1)/(c_i\mu_i) + 1/\mu_i = (x_i + 1)/(c_i\mu_i)$ if $x_i \geq c_i$. Hence:

$$w_a(\mathbf{x}) = \begin{cases} \alpha_a - \beta_a/\mu_a, & \text{if } a \in \{1, 2, \ldots, N\} \text{ and } x_a < c_a, \\ \alpha_a - \beta_a(x_a + 1)/(c_a\mu_a), & \text{if } a \in \{1, 2, \ldots, N\} \text{ and } x_a \geq c_a, \\ 0, & \text{if } a = 0. \end{cases} \tag{4}$$

Also, let $\tilde{\theta}$ denote the 'selfish' (myopic) policy which operates in such a way that any customer arriving in the system chooses the action $a \in \{0, 1, \ldots, N\}$ which maximizes $w_a(\mathbf{x})$, with ties broken arbitrarily except that we assume balking ($a = 0$) is chosen only if $w_i(\mathbf{x}) < 0$ for all $i \in \{1, 2, \ldots, N\}$. By considering the inequalities

$$\alpha_i - \frac{\beta_i x_i}{c_i \mu_i} \geq 0, \qquad \alpha_i - \frac{\beta_i(x_i + 1)}{c_i \mu_i} < 0,$$

we can show that the maximum number of customers at facility i under policy $\tilde{\theta}$ is $\lfloor \beta_i/(\alpha_i c_i \mu_i) \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Hence, the set of positive recurrent states under $\tilde{\theta}$ is $\tilde{S}$, where

$$\tilde{S} = \left\{ \mathbf{x} \in S : x_i \leq \left\lfloor \frac{\alpha_i c_i \mu_i}{\beta_i} \right\rfloor \ \forall i \in \{1, 2, \ldots, N\} \right\}. \tag{5}$$

It is proved in (Shone et al. 2016) (Theorem 2) that there always exists a stationary policy $\theta^* : S \to A$ which maximizes (3). Furthermore, if S_{θ^*} denotes the set of positive recurrent states under a particular optimal stationary policy θ^*, then

$$S_{\theta^*} \subseteq \tilde{S}. \tag{6}$$

This may be described as the 'containment property' of socially optimal policies.

3 The Whittle index heuristic

The main theoretical contributions of our paper (to follow in Sect. 4) rely upon the service facilities having a property referred to as *indexability*, and the resulting development of a heuristic routing policy based on optimal admission policies for individual facilities. Indexability is not necessarily a trivial property to prove in general settings, and sufficient conditions for this property to hold have been well-studied in the literature; see, for example, Bertsimas and Nino-Mora (1996) and Nino-Mora (2001, 2002) for the development of polyhedral methods. As noted by Glazebrook et al. (2009), however, it is often possible to provide simple, direct proofs of indexability in specific problems where the index for resource i (or, in our case, facility i) has a natural interpretation as a *fair charge* for utilization. Fortunately, in our model it is straightforward to establish the indexability property using a simple geometrical argument. Full details follow later in this section, but first we introduce the Lagrangian relaxation upon which the index heuristic is based.

Let Θ denote the class of *stationary* policies under which our N-facility queueing system is stable and has a stationary distribution; that is, if $\theta \in \Theta$ then the distribution

$\{\pi_\theta(\mathbf{x})\}_{\mathbf{x}\in S}$ exists, where $\pi_\theta(\mathbf{x})$ is the steady-state probability of the system being in state $\mathbf{x} \in S$ under θ and $\sum_{\mathbf{x}\in S} \pi_\theta(\mathbf{x}) = 1$. We note that although λ can be arbitrarily large, Θ is always non-empty since it includes the trivial policy which chooses to balk at all states in S.

For each policy $\theta \in \Theta$ and facility $i \in \{1, 2, \ldots, N\}$, let $\eta_i(\theta)$ denote the *effective* queue-joining rate per unit time at facility i under θ (i.e. the long-run average number of customers joining facility i per unit time), and let $L_i(\theta)$ denote the long-run average number of customers present at facility i under θ. Then the long-run average reward g^θ under policy θ is independent of the initial state $\mathbf{x}_0$ and may be expressed in the form

$$g^\theta = \sum_{i=1}^{N} (\alpha_i \eta_i(\theta) - \beta_i L_i(\theta)) \tag{7}$$

Closed-form expressions for $\eta_i(\theta)$ and $L_i(\theta)$ are unattainable in general when $N \geq 2$, but the expression on the right-hand side of (7) will nevertheless prove useful. We note that under any policy $\theta \in \Theta$, the sum of the effective queue-joining rates $\eta_i(\theta)$ at the various facilities must be bounded above by the system demand rate. That is:

$$\sum_{i=1}^{N} \eta_i(\theta) \leq \lambda. \tag{8}$$

Following Whittle (1988), we consider a Lagrangian relaxation of our original problem involving an expanded class of stationary policies Θ' which are at liberty to 'break' the natural physical restrictions of the system by sending a newly-arrived customer to any *subset* of the set of facilities $\{1, 2, \ldots, N\}$. That is, for each state $\mathbf{x} \in S$, the action $\theta(\mathbf{x})$ chosen by a policy $\theta \in \Theta'$ satisfies

$$\theta(\mathbf{x}) \in \mathcal{P}\left(\{1, 2, \ldots, N\}\right),$$

where $\mathcal{P}\left(\{1, 2, \ldots, N\}\right)$ is the *power set* (i.e. the set of all subsets, including the empty set) of $\{1, 2, \ldots, N\}$. Conceptually, one now considers a new optimization problem in which the option is available to produce 'copies' of each customer who arrives, and send these copies to any number of facilities (at most one copy per facility). For each state $\mathbf{x} \in S$, $\theta(\mathbf{x})$ is the set of facilities which, under the policy $\theta \in \Theta'$, receive (a copy of) a new customer if an arrival occurs under state $\mathbf{x}$.

We incorporate the constraint (8) in a Lagrangian fashion by letting $\hat{g}(W)$ denote the optimal expected long-run average reward for the new (relaxed) problem, defined as

$$\hat{g}(W) := \sup_{\theta\in\Theta'} \left(\sum_{i=1}^{N} (\alpha_i \eta_i(\theta) - \beta_i L_i(\theta)) + W \left(\lambda - \sum_{i=1}^{N} \eta_i(\theta) \right) \right), \tag{9}$$

where $W \in \mathbb{R}$ is a Lagrange multiplier. Clearly, any policy θ belonging to the class of policies Θ for the original problem may be represented by a policy θ' in the new class Θ' for which the cardinality of $\theta'(\mathbf{x})$ is either 1 or 0 at all states $\mathbf{x} \in S$. Hence, for $W \geq 0$,

$$g^* \leq \hat{g}(W),$$

where $g^* = \sup_{\theta \in \Theta} g^{\theta}$ is the optimal expected long-run average reward in the original problem. One can re-write (9) in an equivalent form:

$$\hat{g}(W) = \sup_{\theta \in \Theta'} \left(\sum_{i=1}^{N} \Big((\alpha_i - W)\, \eta_i(\theta) - \beta_i L_i(\theta) \Big) \right) + \lambda W. \tag{10}$$

Then, as in Glazebrook et al. (2009) (for example), one obtains a *facility-wise decomposition*:

$$\hat{g}(W) = \sum_{i=1}^{N} \hat{g}_i(W) + \lambda W, \tag{11}$$

where, for each facility $i \in \{1, 2, \ldots, N\}$,

$$\hat{g}_i(W) = \sup_{\theta \in \Theta_i{}'} \Big((\alpha_i - W)\, \eta_i(\theta) - \beta_i L_i(\theta) \Big).$$

Here, $\Theta_i{}'$ (for $i = 1, 2, \ldots, N$) is a class of stationary policies which choose either to accept a customer (denoted by 1) or reject (denoted by 0) at any given state. Since the relaxation of the problem allows newly-arrived customers to be sent to any subset of the N facilities, the decision of whether or not to admit a customer at some facility $i \in \{1, 2, \ldots, N\}$ can be made independently of the decisions made in regard to the other facilities $j \neq i$. It follows that an optimal solution to the relaxed N-facility problem can be found by solving N independent *single-facility* admission control problems. For each facility $i \in \{1, 2, \ldots, N\}$, the corresponding single-facility problem involves customers arriving according to a Poisson process with a demand rate λ (the same demand rate as for the N-facility problem), c_i service channels, and exponentially-distributed service times with mean μ_i^{-1}. The holding cost is β_i per customer per unit time, but importantly the reward for service is now $\alpha_i - W$ as opposed to α_i. Hence, it is natural to interpret W as an extra charge for admitting a customer; see Fig. 2.

The single-facility problem described above exactly fits the formulation of Sect. 2 (with $N = 1$), except that the reward $\alpha_i - W$ may not be positive. Interpreting Theorem 2 from Shone et al. (2016) in the context of a single-facility problem, we can be assured that there exists an average reward optimal *threshold* policy. This means that a customer arriving at the facility balks if and only if the system state x equals or exceeds the integer threshold n, i.e. $x \geq n$. We will refer to such a policy as an n-threshold policy.

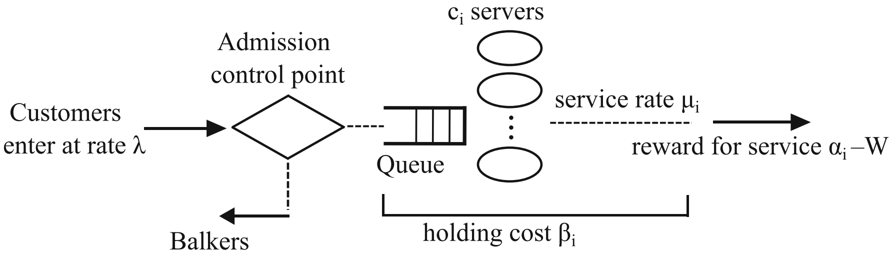

Fig. 2 An $M/M/c_i$ queue with an extra admission charge W

Let θ_i^* denote an optimal threshold policy at facility i, given an entry charge W. This means that θ_i^* chooses an action $a \in \{0, 1\}$ in response to an input $(x, W) \in \mathbb{N}_0 \times \mathbb{R}$, where x is the observed state and W is the entry charge. Re-interpreting the summary measures $\eta_i(\cdot)$ and $L_i(\cdot)$ so that they are now functions of policies θ_i belonging to the set $\Theta_i{}'$ associated with the *single-facility* problem, it follows that $(\alpha_i - W)\eta_i(\theta_i^*) - \beta_i L_i(\theta_i^*)$ is a valid expression for $\hat{g}_i(W)$, the optimal average reward for facility i.

Next, we recall that the facility $i \in \{1, 2, \ldots, N\}$ was arbitrary in this discussion and let $\theta_1^*, \theta_2^*, \ldots, \theta_N^*$ be optimal threshold policies at the various facilities. Also, let θ^* be a stationary policy belonging to the expanded class Θ' which operates in such a way that, for each state $\mathbf{x} \in S$,

$$\theta^*(\mathbf{x}, W) = \Big\{i \in \{1, 2, \ldots, N\} : \theta_i^*(x_i, W) = 1\Big\}. \tag{12}$$

That is, each time a new customer arrives, they are sent to *all* of the facilities $i \in \{1, 2, \ldots, N\}$ at which the optimal threshold policy θ_i^* would choose to accept a customer. By the previous arguments, θ^* attains average reward optimality in the relaxed version of the problem.

In order to derive the Whittle heuristic for the *original* N-facility problem, it remains to establish the connection between this heuristic and the optimal solutions for the relaxed version of the problem discussed thus far. The Whittle heuristic relies upon the notion of *indexability* of a service facility, which we define [in line with Whittle (1988)] as follows:

Definition 1 (*Indexability*) Facility $i \in \{1, 2, \ldots, N\}$ is said to be *indexable* if, given any state $x \in \mathbb{N}_0$, there exists $W_i(x) \in \mathbb{R}$ such that θ_i^* chooses to accept a new customer if and only if $W < W_i(x)$.

For each facility $i \in \{1, 2, \ldots, N\}$, let $T_i^*(W)$ denote the smallest integer n such that an n-threshold policy achieves average reward optimality in a single-facility problem with demand rate λ, c_i service channels, service rate μ_i, holding cost β_i and reward for service $\alpha_i - W$. Then we can show that the indexability property holds for facility i if and only if $T_i^*(W)$ satisfies the following properties:

1. $T_i^*(W)$ is monotonically decreasing with W.
2. For any $x \in \mathbb{N}_0$, there exists $W_i(x) \in \mathbb{R}$ such that $T_i^*(W) > x$ if and only if $W < W_i(x)$.

In the event that the indexability property holds, we refer to the critical value $W_i(x)$ as the *Whittle index* for facility i and state x. Although indexability is not trivial to prove in general, the property has been shown to hold in various problems involving queueing or inventory control (see Nino-Mora 2002; Ansell et al. 2003b; Archibald et al. 2009; Glazebrook et al. 2009; Argon et al. 2009; Hodge and Glazebrook 2011 and references therein). The next result confirms that the facilities in our problem are indexable, and also provides an expression for the Whittle index $W_i(x)$ in terms of the system parameters. Proof of the lemma can be found in Appendix A.

Lemma 1 *Each facility $i \in \{1, 2, \ldots, N\}$ is indexable. Furthermore, a valid expression for the Whittle index $W_i(x)$ is*

$$W_i(x) = \alpha_i - \frac{\beta_i \left(\sum_{y=0}^{x+1} y\pi_i(y, x+1) - \sum_{y=0}^{x} y\pi_i(y, x) \right)}{\lambda \left(\pi_i(x, x) - \pi_i(x+1, x+1) \right)}, \tag{13}$$

where $\pi_i(y, T)$ denotes the steady-state probability of facility i being in state $y \in \mathbb{N}_0$, given that a threshold of T is applied.

We note that convenient formulae for $\pi_i(y, T)$ are available from finite-buffer $M/M/c$ queueing theory (see, for example, Gross and Harris 1998):

$$\pi_i(y, T) = \begin{cases} [(\lambda/\mu_i)^y/(y!)]\, \pi_i(0, T), & \text{if } y \le c_i, \\ [(\lambda/\mu_i)^y/(c_i^{y-c_i} c_i!)]\, \pi_i(0, T), & \text{if } y \ge c_i, \end{cases}$$

where

$$\pi_i(0, T) = \left(\sum_{k=0}^{c_i-1} \frac{\lambda^k}{\mu_i^k k!} + \frac{\lambda^{c_i}}{\mu_i^{c_i} c_i!} \sum_{k=c_i}^{T} \frac{\lambda^{k-c_i}}{(c_i\mu_i)^{k-c_i}} \right)^{-1}.$$

Several remarks should be made at this point. Firstly, equation (13) can be found in a more general form in Corollary 7.1 of Nino-Mora (2012). Secondly, the expression $\sum_{y=0}^{x+1} y\pi_i(y, x+1) - \sum_{y=0}^{x} y\pi_i(y, x)$ which appears on the right-hand side of (13) is simply $L_i(x+1) - L_i(x)$, where $L_i(x)$ is the expected number of customers present at facility i given a threshold of x. In the special case where $x < c_i$, Little's formula yields $L_i(x+1) - L_i(x) = (\lambda/\mu_i)(\pi_i(x, x) - \pi_i(x+1, x+1))$, and hence we obtain

$$W_i(x) = \alpha_i - \frac{\beta_i}{\mu_i} \qquad (x < c_i). \tag{14}$$

Thus, the Whittle index at states $x < c_i$ is equal to a customer's expected net reward for joining facility i. As a further remark, suppose we have a single-server facility ($c_i = 1$). Then it is straightforward to apply results for finite-buffer $M/M/1$ queues in order to obtain

$$W_i(x) = \begin{cases} \alpha_i - \dfrac{\beta_i\left((x+1)(1-\rho_i) - \rho_i(1-\rho_i^{x+1})\right)}{\mu_i(1-\rho_i)^2}, & \text{if } \rho \neq 1, \\ \alpha_i - \dfrac{\beta_i(x+1)(x+2)}{2\mu_i}, & \text{if } \rho = 1, \end{cases}$$

where $\rho_i = \lambda/\mu_i$. This is analogous to the equation (7.3) given in Nino-Mora (2002), except that their result is given in the context of minimizing holding costs (without a reward for service). A similar result can also be found by considering equation (18) in Argon et al. (2009) and setting (in their notation) $\alpha = 0$, $\beta = \lambda/\mu = \rho$ and $c(i) = (i+1)h/\mu$.

In the light of Definition 1, we can obtain an optimal policy θ^* for the relaxed N-facility problem by specifying its decision at state $\mathbf{x} \in S$ as follows:

$$\theta^*(\mathbf{x}, W) = \left\{i \in \{1, 2, \ldots, N\} : W_i(x_i) > W\right\}.$$

As observed in Argon et al. (2009) and Glazebrook et al. (2009), the fact that an optimal solution to the relaxed problem may be described using the Whittle indices makes it logical to propose a *heuristic* policy for the original N-facility problem, which involves sending any new customer who arrives under state $\mathbf{x} \in S$ to a facility i which maximizes $W_i(x_i)$, or choosing to balk if none of the $W_i(x_i)$ values are positive. The optimality of such a policy cannot be guaranteed, but its intuitive justification lies in the fact that $W_i(x_i)$, when positive, is a measure of the amount by which the 'charge for admission' W would need to be increased before the optimal policy θ^* for the *relaxed* problem would choose not to admit a customer to facility i. Thus, $W_i(x_i)$ may be regarded somewhat crudely as a measure of the margin by which one would be 'in favor' of having an extra customer present at facility i. A similar interpretation is that $W_i(x_i)$ is a 'fair charge' for admitting a customer to facility i when there are x_i customers already present.

The *Whittle index heuristic policy* $\theta^{[W]}$ (hereafter referred to as the *Whittle policy*) for our original N-facility routing problem is defined below.

Definition 2 (*Whittle index policy*) At any given state $\mathbf{x} \in S$, the *Whittle index policy* $\theta^{[W]}$ chooses an action as follows:

$$\theta^{[W]}(\mathbf{x}) \in \begin{cases} \underset{i \in \{1,2,\ldots,N\}}{\arg\max}\ W_i(x_i), & \text{if } \exists\, i \in \{1, 2, \ldots, N\} \text{ such that } W_i(x_i) > 0, \\ \{0\}, & \text{otherwise}, \end{cases} \tag{15}$$

where $W_i(x)$ is defined in (13). In cases where two or more facilities attain the maximum in (15), it will be assumed that a decision is made according to some pre-determined ranking order of the N facilities.

We note that, for any state $\mathbf{x} = (x_1, \ldots, x_N) \in S$, there is an equivalence between the following three statements:

1. $W_i(x_i) > 0$ for some $i \in \{1, 2, \ldots, N\}$,
2. $\theta^{[W]}(\mathbf{x}) \neq 0$,

3. Any optimal stationary policy for a single-facility problem with parameters corresponding to those of facility $i \in \{1, 2, \ldots, N\}$, is a threshold policy with threshold greater than x_i.

We will make use of this equivalence in several of our later proofs.

In the next section we investigate the similarities and differences between the Whittle index policy and an optimal policy which maximizes (3).

4 Structural and asymptotic properties of the index heuristic

Suppose we have a stationary policy, θ^*, which is optimal under the average reward criterion. In this section we will present several counter-examples to show that θ^* may possess surprising and counter-intuitive structural properties. Indeed, there is little that can be proved about θ^* in general. However, it is possible to show that the positive recurrent state space under θ^* may be bounded by two finite sets. Let the sets S° and S_{θ^*} be defined as follows:

$$\begin{aligned} S^\circ &:= \{\mathbf{x} \in S : x_i \leq c_i \ \forall\, i \in \{1, 2, \ldots, N\}\}, \\ S_{\theta^*} &:= \{\mathbf{x} \in S : \mathbf{x} \text{ is positive recurrent under } \theta^*\}. \end{aligned} \tag{16}$$

Also, let $\tilde{S}$ be the 'selfish' state space defined in (5). The following relationship may be proved to hold for any optimal stationary policy θ^*:

$$S^\circ \subseteq S_{\theta^*} \subseteq \tilde{S}. \tag{17}$$

Indeed, the fact that $S_{\theta^*} \subseteq \tilde{S}$ has been proved in Shone et al. (2016) (Lemma 6). By using this result and also showing that optimal policies never choose to balk if there is an idle server available at one of the N facilities, the lower bound $S^\circ \subseteq S_{\theta^*}$ can be established. A full proof can be found in Appendix B. We refer to the property $S^\circ \subseteq S_{\theta^*}$ as the *non-idling* property of optimal policies.

We have not stated (17) as a theorem because it can be regarded as a corollary of a stronger result, which follows next. It is possible to use the structural properties of the Whittle index policy to obtain an improved lower bound for S_{θ^*}. Throughout the rest of this section we will use $\theta^{[W]}(\mathbf{x}) \in \{0, 1, \ldots, N\}$ to denote the action chosen by the Whittle policy $\theta^{[W]}$ in response to an observed state $\mathbf{x} \in S$, and we will also use S_W to denote the set of states in S which are positive recurrent under $\theta^{[W]}$. The following lemma is needed:

Lemma 2 *Let θ^* be an optimal stationary policy. Then, for any $\mathbf{x} \in S_{\theta^*}$,*

$$\theta^*(\mathbf{x}) = 0 \quad \Rightarrow \quad \theta^{[W]}(\mathbf{x}) = 0.$$

That is, the Whittle policy $\theta^{[W]}$ chooses to balk at any state $\mathbf{x} \in S_{\theta^*}$ where θ^* chooses to balk.

Proof of the lemma is established using dynamic programming recursions and can also be achieved via a sample path argument. The details can be found in Appendix C. Essentially, one can show that if balking is chosen at some state $\mathbf{x} \in S_{\theta^*}$ by the optimal policy θ^*, then balking would also be chosen by an optimal threshold policy in a single-facility problem involving any of the facilities $i \in \{1, 2, \ldots, N\}$ at the state with x_i customers present (where x_i is the i^{th} component of the state $\mathbf{x}$ in the N-facility problem). Since the Whittle policy $\theta^{[W]}$ makes decisions by considering each of the N facilities operating in isolation, this is sufficient to establish the result.

Our next theorem states that the Whittle index policy $\theta^{[W]}$ is *conservative* in comparison to an optimal stationary policy θ^*.

Theorem 1 (Conservativity of the Whittle policy) *For any optimal stationary policy θ^*, we have*

$$S^{\circ} \subseteq S_W \subseteq S_{\theta^*} \subseteq \tilde{S}. \tag{18}$$

Proof The containment property $S_{\theta^*} \subseteq \tilde{S}$ is already known. It follows that there must exist some state $\mathbf{x} \in S_{\theta^*}$ at which θ^* chooses to balk; otherwise, an unbroken sequence of customer arrivals (without any service completions) would cause the process to pass outside $\tilde{S}$ under θ^*. Let $\mathbf{z} \in S_{\theta^*}$ be a state at which θ^* chooses to balk, and let

$$S_{\mathbf{z}} := \left\{ \mathbf{x} \in \tilde{S} : x_i \leq z_i \quad \forall i \in \{1, 2, \ldots, N\} \right\}.$$

That is, $S_{\mathbf{z}}$ is the set of states in $\tilde{S}$ which satisfy the componentwise inequality $\mathbf{x} \leq \mathbf{z}$. Since $\mathbf{z} \in S_{\theta^*}$, it follows that all states in $S_{\mathbf{z}}$ are also included in S_{θ^*}, since they are accessible from $\mathbf{z}$ via service completions. Hence, $S_{\mathbf{z}} \subseteq S_{\theta^*}$. On the other hand, since balking is chosen by θ^* at $\mathbf{z}$, it follows from Lemma 2 that balking is also chosen at $\mathbf{z}$ by the Whittle policy $\theta^{[W]}$. By definition of the Whittle policy, this implies that $W_i(z_i) \leq 0$ for all $i \in \{1, 2, \ldots, N\}$. Therefore it is impossible for any state $\mathbf{x} \notin S_{\mathbf{z}}$ to be accessible from state $\mathbf{0}$ (the empty system state) under the Whittle policy, since this would require joining some facility $i \in \{1, 2, \ldots, N\}$ to be chosen at a state $\mathbf{y} \in S_{\mathbf{z}}$ with $y_i = z_i$ and hence $W_i(y_i) \leq 0$. It follows that $S_W \subseteq S_{\mathbf{z}} \subseteq S_{\theta^*}$.

To complete the proof, it remains only to show that $S^{\circ} \subseteq S_W$. In Sect. 3 it was shown that, for any facility $i \in \{1, \ldots, N\}$, we have $W_i(x) = \alpha_i - \beta_i/\mu_i$ for all states $x < c_i$ (see (14)). Recall that our model assumes $\alpha_i - \beta_i/\mu_i > 0$; otherwise, facility i would be redundant. Hence, $\theta^{[W]}$ cannot choose to balk at any state with $x_i < c_i$ for some $i \in \{1, 2, \ldots, N\}$. Since S_W is contained in S_{θ^*} (and hence finite), it then follows that there exists a state $\mathbf{x}$ with $x_i \geq c_i$ for all $i \in \{1, 2, \ldots, N\}$ which is positive recurrent under $\theta^{[W]}$ (indeed, such a state must be accessible from $\mathbf{0}$ via an unbroken sequence of customer arrivals). Hence, all states in S° are also positive recurrent under $\theta^{[W]}$. □

Next, we turn our attention to the asymptotic properties of the Whittle index policy as λ becomes either very small or very large. Since $\theta^{[W]}$ is a heuristic policy, its optimality cannot be proved in general, but the next theorem establishes that the Whittle policy achieves (asymptotic) optimality in a light-traffic limit, and also in a heavy-traffic limit.

Theorem 2 (Optimality of the Whittle policy in light-traffic and heavy-traffic limits) *Let $g^{[W]}(\lambda)$ be the long-run average reward attained by the Whittle policy $\theta^{[W]}(\lambda)$ given a demand rate $\lambda > 0$, and let $g^*(\lambda)$ be the corresponding value under an optimal policy. Then:*

1. $\theta^{[W]}$ *is asymptotically optimal in a light-traffic limit. That is:*

$$\lim_{\lambda \to 0} \frac{g^*(\lambda) - g^{[W]}(\lambda)}{g^*(\lambda)} = 0.$$

2. $\theta^{[W]}$ *is optimal in a heavy-traffic limit. That is:*

$$\lim_{\lambda \to \infty} \left(g^*(\lambda) - g^{[W]}(\lambda) \right) = 0.$$

Proof of Theorem 2 can be found in Appendix D. In the light-traffic case, it suffices to show that the Whittle policy makes optimal decisions at the state with no customers present, since the decisions chosen at other states essentially become unimportant in the limiting scenario. In the heavy-traffic case, the proof is accomplished by showing that the Whittle heuristic directs customers to balk if and only if all servers are busy at all facilities, and that this results in the system residing continuously in a state which maximizes the single-step reward function $r(\mathbf{x})$.

Theorem 2 relies upon the fact that optimal policies become quite simplistic in the limiting cases as $\lambda \to 0$ and $\lambda \to \infty$. For general values of λ, however, optimal policies can be quite intricate. The remaining results in this section identify structural properties of the Whittle policy $\theta^{[W]}$ which do not necessarily hold under an optimal policy.

First, we consider monotonicity. We will generalize our previous notation and use S_θ to denote the set of states which are positive recurrent under an arbitrary stationary policy θ.

Theorem 3 (Monotonicity) *Suppose $N \geq 2$, and let $\Theta^{[M]}$ denote the class of all stationary policies θ which satisfy the following three monotonicity properties:*

(a) *If $\theta(\mathbf{x}) = 0$ for some $\mathbf{x} \in S_\theta$, then $\theta(\mathbf{x}^{i+}) = 0$ for all $i \in \{1, 2, \ldots, N\}$ with $\mathbf{x}^{i+} \in S_\theta$.*
(b) *If $\theta(\mathbf{x}) = i$ for some $\mathbf{x} \in S_\theta$ and $i \in \{1, 2, \ldots, N\}$ with $x_i \geq 1$, then $\theta(\mathbf{x}^{i-}) = i$.*
(c) *If $\theta(\mathbf{x}) = i$ for some $\mathbf{x} \in S_\theta$ and $i \in \{1, 2, \ldots, N\}$, then $\theta(\mathbf{x}^{j+}) = i$ for any $j \in \{1, 2, \ldots, N\} \setminus \{i\}$ such that $\mathbf{x}^{j+} \in S_\theta$.*

Then:

1. $\theta^{[W]} \in \Theta^{[M]}$,
2. *If $N = 2$ and $c_1 = c_2 = 1$ then there exists an optimal policy in $\Theta^{[M]}$,*

3. *In general, $\Theta^{[M]}$ is not guaranteed to include an optimal policy.*

Proof It is trivial to show that the Whittle policy $\theta^{[W]}$ possesses the monotonicity properties (a)–(c), since this is a direct consequence of the index-based nature of the policy. We therefore begin with Statement 2, which relates to a special case of our model with only two facilities and a single server at each. In general, any optimal policy must be associated with a constant g^* and a function h satisfying the well-known average reward optimality equations:

$$g^* + h(\mathbf{x}) = \max_{a \in A} \left\{ r(\mathbf{x}) + \sum_{\mathbf{y} \in S} p(\mathbf{x}, a, \mathbf{y}) h(\mathbf{y}) \right\} \quad (\mathbf{x} \in S). \tag{19}$$

Importantly, the function h satisfying (19) is unique up to an additive constant (see Puterman 1994). The proof of Statement 2 depends on showing that, in the special case under consideration, h satisfies three properties defined as follows:

- $h((\mathbf{x}^{j+})^{j+}) - h(\mathbf{x}^{j+}) \leq h(\mathbf{x}^{j+}) - h(\mathbf{x})$ for all $\mathbf{x} \in S$ and $j \in \{1, 2\}$ **(concavity)**;
- $h((\mathbf{x}^{i+})^{j+}) - h(\mathbf{x}^{i+}) \leq h(\mathbf{x}^{j+}) - h(\mathbf{x}^{j+})$ for all $\mathbf{x} \in S$ and $i, j \in \{1, 2\}$ with $i \neq j$ **(submodularity)**;
- $h((\mathbf{x}^{j+})^{j+}) - h((\mathbf{x}^{i+})^{j+}) \leq h(\mathbf{x}^{j+}) - h(\mathbf{x}^{i+})$ for all $\mathbf{x} \in S$ and $i, j \in \{1, 2\}$ with $i \neq j$ **(diagonal submissiveness)**.

These properties can be established using inductive arguments based on value iteration, and the existence of an optimal policy in $\Theta^{[M]}$ then follows. For full details, please refer to Appendix E.

Unfortunately, the properties of concavity, submodularity and diagonal submissiveness cannot be proven to hold in the full generality of our model. We prove Statement 3 of the theorem using a counter-example. Consider a two-facility system with demand rate $\lambda = 12$ and the following parameters for the two facilities:

$$\begin{aligned} c_1 = 2,\ \mu_1 = 8,\ \beta_1 = 10,\ \alpha_1 = 2, \\ c_2 = 2,\ \mu_2 = 2,\ \beta_2 = 10,\ \alpha_2 = 6. \end{aligned}$$

We can use value iteration to confirm the existence of a unique optimal stationary policy θ^* for this system. The positive recurrent state space under this policy is $S_{\theta^*} = \{(x_1, x_2) \in \mathbb{N}_0^2 : x_1 \leq 2 \text{ and } x_2 \leq 2\}$, i.e. it includes 9 states. However, the decision of θ^* at state (0, 0) is to join Facility 2, whereas the decision at (1, 0) is to join Facility 1. This contravenes monotonicity property (b) stated in the theorem, so the proof is complete. □

Fellow researchers may be interested to know that we have been unable to find either a proof or a counter-example to show whether or not monotonicity property (a) is guaranteed to hold under an optimal policy θ^* (indeed, this property may be meaningless if it can be shown that $\theta^*(\mathbf{x}) = 0 \Rightarrow \mathbf{x}^{i+} \notin S_{\theta^*}$). In addition, we have been unable to find either a proof or a counter-example to show whether or not an optimal policy satisfying all three properties (a)–(c) is guaranteed to exist if $N \geq 3$ and $c_i = 1$ for all $i \in \{1, 2, \ldots, N\}$.

Next, we consider how the size of the positive recurrent state space changes as the demand rate λ varies. Intuitively, one might suppose that strong-performing policies should become more conservative as λ increases. The next theorem shows that this is indeed the case for the Whittle policy $\theta^{[W]}$, but not for optimal policies in general.

Theorem 4 (Conservativity with demand) *Let $S_W(\lambda)$ and $S^*(\lambda)$ be defined as follows:*

$$\begin{aligned} S_W(\lambda) &= \{\mathbf{x} \in S : \mathbf{x} \in S_W \textit{ under demand rate } \lambda\}, \\ S^*(\lambda) &= \{\mathbf{x} \in S : \textit{ there exists an optimal stationary policy } \theta^* \textit{ under demand} \\ &\qquad \textit{rate } \lambda \textit{ such that } \mathbf{x} \in S_{\theta^*}\}. \end{aligned}$$

Then, given any two demand rates λ_1, λ_2 with $\lambda_1 > \lambda_2 > 0$:

1. $S_W(\lambda_1) \subseteq S_W(\lambda_2)$,
2. *If $N = 1$, then $S^*(\lambda_1) \subseteq S^*(\lambda_2)$,*
3. *In general, $S^*(\lambda_1) \nsubseteq S^*(\lambda_2)$.*

Proof Since the Whittle index policy is derived from the properties of optimal admission policies for single-facility problems, Statement 1 is actually implied by Statement 2. However, Statement 2 is somewhat non-trivial to prove. We have used a dynamic programming argument to establish this result, and the details can be found in Appendix F.

We provide a counter-example to establish Statement 3. Consider a two-facility system in which both facilities have a single server available. The parameters for the facilities are:

$$\begin{aligned} c_1 = 1,\ \mu_1 = 14,\ \beta_1 = 5,\ \alpha_1 = 9, \\ c_2 = 1,\ \ \mu_2 = 5,\ \ \beta_2 = 3,\ \alpha_2 = 20. \end{aligned}$$

Consider two different demand rates, $\lambda_1 = 10$ and $\lambda_2 = 9.8$. Under the larger demand rate λ_1, the unique optimal policy θ_1^* found by value iteration has positive recurrent state space $S^*(\lambda_1) = \{(x_1, x_2) \in \mathbb{N}_0^2 : x_1 \leq 10$ and $x_2 \leq 14\}$, with a unique balking state $(10, 14)$. However, under the smaller demand rate λ_2, value iteration yields a unique optimal policy with $S^*(\lambda_2) = \{(x_1, x_2) \in \mathbb{N}_0^2 : x_1 \leq 11$ and $x_2 \leq 13\}$, with a unique balking state $(11, 13)$. Since $S^*(\lambda_1)$ includes states with $x_2 = 14$, the 'conservativity with demand' property does not hold. □

Next, we examine a property related to the distribution of balking states under a stationary policy. It is natural to suppose that, under a strong-performing policy θ, the positive recurrent state space S_θ should take the form of a cuboid in N dimensions. Indeed, if S_θ is finite, then the cuboid property is implied by the existence of a *unique* state in S_θ at which balking is chosen. The next result states that the Whittle policy $\theta^{[W]}$ must have a unique recurrent balking state, but this is not necessarily true for an optimal stationary policy.

Theorem 5 (Unique recurrent balking state) *Suppose $N \geq 2$, and let $\Theta^{[B]}$ denote the set of all stationary policies θ for which the set of positive recurrent states S_θ includes a unique state at which balking is chosen. Then:*

1. $\theta^{[W]} \in \Theta^{[B]}$*,*
2. *If $N = 2$ and $c_1 = c_2 = 1$ then there exists an optimal policy in $\Theta^{[B]}$,*
3. *In general, $\Theta^{[B]}$ is not guaranteed to include an optimal policy.*

Proof The proof of Statement 1 is trivial, since the only state $\mathbf{x} \in S_W$ at which $\theta^{[W]}$ chooses to balk is the state with $x_i = \min\{x \geq 0 : W_i(x) \leq 0\}$ for $i \in \{1, 2, \ldots, N\}$.

The proof of Statement 2 relies upon the properties of concavity, submodularity and diagonal submissiveness for the function h satisfying the Eq. 19 in a system with two single-server facilities. These properties were established (for the $N = 2, c_1 = c_2 = 1$ case) in the proof of Theorem 3. Details of how these properties imply a unique balking state can be found in Ha (1997) (Theorem 3). Ha's results are given in the context of a make-to-stock production system with two products and a single server. He defines a 'base stock policy' as a policy for which production is stopped if and only if all products have inventory at or above their specified base stock levels; this is analogous to a policy with a unique balking state in our model. We also note that Ha considers a minimization problem as opposed to a maximization problem, and the value function in his model has the converse properties of convexity, supermodularity and diagonal dominance. However, the arguments in his proof can be translated to our setting in an obvious way.

To prove Statement 3, we provide a counter-example which was found by a numerical search. Consider a system with 3 facilities and a demand rate $\lambda = 21.57$. The parameters for the facilities are:

$$\begin{array}{llll} c_1 = 2, & \mu_1 = 15.17, & \beta_1 = 12.01, & \alpha_1 = 5.65, \\ c_2 = 4, & \mu_2 = 10.09, & \beta_2 = 22.4, & \alpha_2 = 9.07, \\ c_3 = 3, & \mu_3 = 6.36, & \beta_3 = 7.16, & \alpha_3 = 5.46. \end{array}$$

For this system, the unique optimal policy θ^* found using value iteration chooses to balk at the states (13, 10, 14) and (12, 11, 14), both of which are positive recurrent under θ^*. Thus, the process operating under θ^* is able to access two different 'balking states', implying that $\theta^* \notin \Theta^{[B]}$. □

Our final result in this section concerns a special case in which all of the N facilities share the same parameters (c_i, μ_i, α_i and β_i). We refer to this as the 'homogeneous facilities' case. Like the previous three results, it highlights an intuitively 'sensible' structural property which is possessed by the Whittle policy $\theta^{[W]}$, but not by optimal policies in general.

Theorem 6 (Cube property for a homogeneous system) *Suppose $N \geq 2$ and the facilities are homogeneous, i.e. we have $c_i = c$, $\mu_i = \mu$, $\alpha_i = \alpha$, $\beta_i = \beta$ for $i \in \{1, 2, \ldots, N\}$. Let $\Theta^{[C]}$ denote the set of all stationary policies θ for which the set of positive recurrent states S_θ is of the form $\{\mathbf{x} \in S : x_i \leq M$ for all $i \in \{1, 2, \ldots, N\}\}$ for some $M \in \mathbb{N}_0$. Then:*

Table 1 An optimal decision-making structure for a system with homogeneous facilities

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

1. $\theta^{[W]} \in \Theta^{[C]}$,
2. *In general, $\Theta^{[C]}$ is not guaranteed to include an optimal policy.*

Proof As in Theorems 3 and 5, the proof of Statement 1 is trivial, since it is a direct consequence of the index-based nature of the Whittle index policy.

We provide a counter-example to establish Statement 2. Consider a system with demand rate $\lambda = 15$ and two single-server facilities which share an identical set of parameters as follows:

$$c = 1, \quad \mu = 4, \quad \beta = 1, \quad \alpha = 5.$$

With these parameters, it transpires that the set of positive recurrent states S_{θ^*} associated with any optimal stationary policy must be either of dimension 3×4 or 4×3. The optimal decision-making structure is shown in Table 1.

If we restrict attention to stationary policies, then it can be seen from Table 1 that there must be a unique balking state at either $(2, 3)$ or $(3, 2)$. Therefore an optimal stationary policy will allow *one* of the two facilities to have up to three customers present, but not both. The state $(3, 3)$ is not accessible from $\mathbf{0}$ (i.e. positive recurrent) under any optimal stationary policy. Since Table 1 accounts for all 8 optimal stationary policies in this system, we conclude that none of these are included in $\Theta^{[C]}$. □

The results in this section have shown that the Whittle policy $\theta^{[W]}$ belongs to a class of policies which possess certain intuitively 'sensible' structural properties. However, the counter-examples have shown that an optimal policy need not necessarily be included in the same class, and therefore $\theta^{[W]}$ must be sub-optimal in some cases. In Sect. 6 we present the results of numerical experiments to evaluate the performance of the Whittle policy. These numerical results include comparisons with an alternative heuristic policy, which is developed in the next section.

5 An alternative heuristic policy

In this section we describe an alternative heuristic policy which is derived from the application of a single step of policy iteration to a 'static routing' or 'Bernoulli splitting' policy. Similar approaches have been used for other routing problems in the literature; see Krishnan (1990), Ansell et al. (2003b) and Argon et al. (2009) and references therein. The heuristic shares some similarities with the Whittle heuristic, in the sense that it requires the calculation of indices for the N individual facilities; however, the indices themselves are calculated in a completely different way from those derived in Sect. 3.

To begin, consider a randomized policy under which routing decisions are made according to a fixed probability distribution $\{\sigma_a\}_{a=0}^N$, where a belongs to the same action set A described in Sect. 2; hence, $\sum_{a=0}^N \sigma_a = 1$. We refer to this type of policy as a *static* policy, since it does not have the ability to make decisions dynamically according to the system state. We will commit a slight abuse of notation and represent an arbitrary static policy by a vector $\Lambda = (\lambda_1, \ldots, \lambda_N)$, where $\lambda_i = \lambda\sigma_i$ is the arrival rate for facility i ($i \in \{1, \ldots, N\}$) and $\lambda_0 = \lambda\sigma_0$ is the rate at which customers balk. We can then write the expected long-run average reward under this policy as

$$g^{\Lambda} = \sum_{i=1}^{N} (\lambda_i \alpha_i - \beta_i L_i(\lambda_i)), \tag{20}$$

where $L_i(\lambda_i)$ is the expected number of customers present at facility i, given that arrivals occur according to a Poisson process with rate λ_i (here we are making use of the well-known ‘Poisson splitting’ property). It should be noted that $L_i(\lambda_i)$ is finite if and only if $\lambda_i < c_i\mu_i$, so we will define $g^{\Lambda} = -\infty$ for any policy Λ with $\lambda_i \geq c_i\mu_i$ for at least one facility i. Known results for $M/M/c$ queues imply that $L_i(\cdot)$ is a strictly convex function (see Grassmann 1983; Lee and Cohen 1983); hence, g^{Λ} is strictly concave and there must be a *unique* policy Λ which maximizes g^{Λ} over all static policies.

We will use $\Lambda^* := (\lambda_1^*, \ldots, \lambda_N^*)$ to denote the unique optimal static policy. Here, ‘optimal’ means ‘optimal among all static policies’, not ‘optimal over all policies’. In fact, it is easy to show that all static routing policies are sub-optimal if non-static policies which make state-dependent routing decisions are included as candidates. Nevertheless, it transpires that a strong-performing (heuristic) dynamic routing policy can be obtained by applying a single step of DP-style policy iteration to the optimal static policy Λ^*.

To assist our development, let us define $V^{(n)}(\mathbf{x}, \Lambda^*)$ as the expected total reward over n discrete time steps given that policy Λ^* is followed and the initial state is $\mathbf{x} \in S$. Note that, in our uniformized MDP described in Sect. 2, λ_i^* can be interpreted as the probability that a customer arrives and is sent to facility i at an arbitrary discrete time step under policy Λ^*. Under the usual paradigm of policy iteration, we aim to choose an action a under state $\mathbf{x}$ which maximizes

$$\delta(\mathbf{x}, a) := \lim_{n\to\infty} \left(\lambda V^{(n)}(\mathbf{x}^{a+}, \Lambda^*) - \sum_{b=0}^{N} \lambda_b^* V^{(n)}(\mathbf{x}^{b+}, \Lambda^*) \right). \tag{21}$$

That is, we aim to maximize the improvement in the long-run expected total net reward that would result from choosing action a under state $\mathbf{x}$ at an arbitrary time step and then following the optimal static policy Λ^* at all time steps thereafter, as opposed to simply following the policy Λ^* at all times. Given that the implementation of policy Λ^* results in individual facilities operating independently with their own Poisson arrival rates, it will be useful to write

$$V^{(n)}(\mathbf{x}, \Lambda^*) = \sum_{i=1}^{N} V_i^{(n)}(x_i, \lambda_i^*), \tag{22}$$

where $V_i^{(n)}(x, \lambda_i^*)$ is an expected finite-stage reward for facility i only, given x customers initially present and a Poisson demand rate λ_i^*. We will also define

$$h_i(x, \lambda_i^*) := \lim_{n\to\infty} \left(V_i^{(n)}(x, \lambda_i^*) - V_i^{(n)}(0, \lambda_i^*) \right), \tag{23}$$

$$D_i(x, \lambda_i^*) := h_i(x+1, \lambda_i^*) - h_i(x, \lambda_i^*), \tag{24}$$

for each facility i and state $x \in \mathbb{N}_0$. Then, after some manipulations using (21)–(24), it can be shown that

$$\delta(\mathbf{x}, a) = \begin{cases} \lambda D_i(x_i, \lambda_i^*) - \sum_{j=1}^{N} \lambda_j^* D_j(x_j, \lambda_j^*), & \text{if } a = i \text{ for some } i \in \{1, 2, \ldots, N\}, \\ -\sum_{j=1}^{N} \lambda_j^* D_j(x_j, \lambda_j^*), & \text{if } a = 0. \end{cases} \tag{25}$$

Hence, in order to obtain a dynamic routing policy via the application of a policy iteration step to an optimal static policy, we should make decisions in such a way that customers who arrive under a given state $\mathbf{x} = (x_1, \ldots, x_N)$ are directed to join the facility i which maximizes $D_i(x_i, \lambda_i^*)$ if this value is positive; otherwise, they should balk.

It can be seen from (23) that $h_i(x_i, \lambda_i^*)$ is equivalent to the well-known ‘relative value function’ which appears in the optimality equations and policy evaluation equations for average reward MDPs (see Puterman 1994). In our context, $h_i(x_i, \lambda_i^*)$ applies to facility i only and we can interpret the demand rate λ_i^* as the ‘policy’ for this facility. We have also defined state zero as the ‘reference state’ which the other states’ values are compared against. The policy evaluation equations for facility i can be written

$$g_i(\lambda_i^*) + h_i(x, \lambda_i^*) = r_i(x) + \sum_{y\in\mathbb{N}_0} p_i(x, y, \lambda_i^*) h_i(y, \lambda_i^*) \qquad (x \in \mathbb{N}_0), \tag{26}$$

where $r_i(x)$ and $p_i(x, y, \lambda_i^*)$ are the obvious single-facility analogues of the rewards and transition probabilities defined in Sect. 2 and $g_i(\lambda_i^*)$ is the long-run average reward for facility i. We can then calculate the $D_i(x, \lambda_i^*)$ values by using the Eq. (26). Before proceeding, we note that if $\lambda_i^* = 0$ for a particular facility i then $h_i(x, \lambda_i^*)$ and $D_i(x, \lambda_i^*)$ are trivially equal to zero for all $x \in \mathbb{N}_0$, so we will only consider facilities i for which $\lambda_i^* > 0$.

By setting $x = 0$ in the Eq. (26) and noting that $r_i(0) = 0$ and $h_i(0, \lambda_i^*) = 0$, we obtain:

$$D_i(0, \lambda_i^*) = h_i(1, \lambda_i^*) = g_i(\lambda_i^*)/\lambda_i^*.$$

In general, for integers $x \in \{1, \ldots, c_i - 1\}$, we have:

$$g_i(\lambda_i^*) + h_i(x, \lambda_i^*) = r_i(x) + \lambda_i^* h_i(x+1, \lambda_i^*) + x\mu_i h_i(x-1, \lambda_i^*) + (1 - \lambda_i^* - x\mu_i) h_i(x, \lambda_i^*).$$

Following simple manipulations, we obtain the recurrence relationship:

$$D_i(x, \lambda_i^*) = \frac{x\mu_i}{\lambda_i^*} D_i(x-1, \lambda_i^*) + \frac{g_i(\lambda_i^*) - r_i(x)}{\lambda_i^*}. \tag{27}$$

By making recursive substitutions in (27) we then obtain, for $x \in \{0, 1, \ldots, c_i - 1\}$:

$$D_i(x, \lambda_i^*) = \sum_{k=0}^{x} \frac{x!}{k!} \left(\frac{\mu_i}{\lambda_i^*}\right)^{x-k} \left(\frac{g_i(\lambda_i^*) - r_i(k)}{\lambda_i^*}\right). \tag{28}$$

Similarly, for integers $x \geq c_i$, the recurrence relationship is:

$$D_i(x, \lambda_i^*) = \frac{c_i\mu_i}{\lambda_i^*} D_i(x-1, \lambda_i^*) + \frac{g_i(\lambda_i^*) - r_i(x)}{\lambda_i^*}. \tag{29}$$

By using (28) and (29) and applying a simple inductive argument, one can then show that for $x \geq c_i$, we have:

$$D_i(x, \lambda_i^*) = \sum_{k=0}^{c_i - 1} \frac{c_i!\, c_i^{x-c_i}}{k!} \left(\frac{\mu_i}{\lambda_i^*}\right)^{x-k} \left(\frac{g_i(\lambda_i^*) - r_i(k)}{\lambda_i^*}\right) + \sum_{k=c_i}^{x} \left(\frac{c_i\mu_i}{\lambda_i^*}\right)^{x-k} \left(\frac{g_i(\lambda_i^*) - r_i(k)}{\lambda_i^*}\right). \tag{30}$$

Let $\theta^{[B]}$ denote the ‘Bernoulli improvement’ heuristic which is obtained by applying a step of policy iteration to the optimal static policy Λ^*. The conclusion of this section is that $\theta^{[B]}$ chooses actions as follows:

$$\theta^{[B]}(\mathbf{x}) \in \begin{cases} \underset{i \in \{1,2,\ldots,N\}}{\arg\max}\ D_i(x_i, \lambda_i^*), & \text{if } \exists\, i \in \{1, 2, \ldots, N\} \text{ such that } D_i(x_i, \lambda_i^*) > 0, \\ \{0\}, & \text{otherwise,} \end{cases} \tag{31}$$

where $D_i(x_i, \lambda_i^*)$ is defined in (28) (for $x_i \in \{0, 1, \ldots, c_i - 1\}$) and (30) (for $x \geq c_i$). We note that, from a practical point of view, implementation of $\theta^{[B]}$ requires the initial solution of a convex optimization problem in order to obtain the optimal static policy Λ^*. The values $g_i(\lambda_i^*)$ (required for the computation of $D_i(x, \lambda_i^*)$) are then readily obtained as functions of the λ_i^*.

For the special case $c_i = 1$, we note that (28) and (30) reduce to

$$D_i(x, \lambda_i^*) = \alpha_i - \frac{\beta_i(x+1)}{\mu_i - \lambda_i^*} \qquad (x \geq 0).$$

This implies that $\theta^{[B]}$ is more conservative than the selfish policy $\tilde{\theta}$ in a system with single servers at all facilities, since $\tilde{\theta}$ prefers joining facility i to balking at state $\mathbf{x}$ if and only if $\alpha_i - \beta_i(x_i + 1)/\mu_i \geq 0$ (see Sect. 2).

The next theorem states that the Bernoulli improvement policy possesses the property of asymptotic light-traffic optimality which (according to Theorem 2) is also a feature of the Whittle policy $\theta^{[W]}$.

Theorem 7 (Optimality of the Bernoulli improvement policy in a light-traffic limit) *Let $g^{\Lambda^*}(\lambda)$ and $g^{[B]}(\lambda)$ be the long-run average rewards attained by the optimal static policy Λ^* and the Bernoulli improvement policy $\theta^{[B]}$ respectively given a demand rate $\lambda > 0$, and let $g^*(\lambda)$ be the corresponding value under an optimal policy. Then Λ^* and $\theta^{[B]}$ are both asymptotically optimal in a light-traffic limit. That is:*

$$\lim_{\lambda \to 0} \frac{g^*(\lambda) - g^{\Lambda^*}(\lambda)}{g^*(\lambda)} = \lim_{\lambda \to 0} \frac{g^*(\lambda) - g^{[B]}(\lambda)}{g^*(\lambda)} = 0.$$

On the other hand, it is easy to find counter-examples to show that $\theta^{[B]}$ does *not* possess the property of heavy-traffic optimality described (in the context of the Whittle policy $\theta^{[W]}$) in Theorem 2. In Appendix G we have provided a proof of Theorem 7 and also a counter-example to show the lack of heavy-traffic optimality for $\theta^{[B]}$.

6 Numerical results

In this section we report the results of a series of experiments involving more than 37,000 randomly-generated sets of system parameters. In order to evaluate the *exact* sub-optimality of a heuristic policy, it is necessary to evaluate the expected long-run average reward earned by the relevant policy and compare this with the optimal value g^* associated with an *average reward optimal* policy. Usually, one would wish to carry out these tasks using dynamic programming algorithms, but this is only practical if the finite state space $\tilde{S}$ is of relatively modest size. Of course, the Whittle policy described in Sect. 3 can easily be applied to systems in which $\tilde{S}$ is extremely large, but it is generally not feasible to evaluate the optimal value g^* in such systems, and therefore the only comparisons of interest that can be made in 'large' systems are between the Whittle policy (whose performance must be *approximated*, using simulation) and with alternative heuristics such as the selfish policy $\tilde{\theta}$ and the Bernoulli improvement policy $\theta^{[B]}$ described in Sects. 2 and 5 respectively. As such, this section is divided into two subsections:

- In Sect. 6.1, systems of relatively modest size are considered. These are systems in which the size of $|\tilde{S}|$ facilitates the efficient computation of the optimal average

Table 2 95% confidence intervals for the percentage suboptimality of heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$ (columns 3–5) for different values of N (the number of facilities)

		Pct. Suboptimality		
N value	Count	$\theta^{[W]}$	$\theta^{[B]}$	$\tilde{\theta}$
All values	32,934	0.659 ± 0.008	1.099 ± 0.013	38.386 ± 0.387
$N = 2$	12,332	0.526 ± 0.012	0.872 ± 0.020	35.934 ± 0.641
$N = 3$	11,456	0.729 ± 0.014	1.242 ± 0.025	41.171 ± 0.668
$N = 4$	9,146	0.750 ± 0.016	1.226 ± 0.025	38.204 ± 0.697

reward g^* using DP algorithms, and also enables similar evaluations of the average rewards earned by the Whittle policy $\theta^{[W]}$, the Bernoulli improvement policy $\theta^{[B]}$ and the selfish policy $\tilde{\theta}$.

- In Sect. 6.2, ‘large’ systems are considered. These are systems in which the exact computation of g^* is assumed to be infeasible, and the average rewards earned by $\theta^{[W]}$ are compared with those associated with alternative heuristic policies via simulation experiments.

For purposes of distinction, a ‘modest-sized’ system is defined in this section as a system in which $2 \leq N \leq 4$ and the cardinality of $\tilde{S}$ is between 100 and 100,000. Although it is certainly possible to apply DP algorithms to systems of greater size than this, it is desirable to impose a relatively strict restriction on $|\tilde{S}|$ in order to allow a large number of experiments to be carried out in a reasonable amount of time. The remainder of this section will proceed to present the results obtained from numerical experiments.

6.1 ‘Modest-sized’ systems with $2 \leq N \leq 4$

We conducted a series of numerical experiments involving 32,934 randomly-generated sets of system parameters. Details of the methods used to generate the parameters can be found in Appendix H.

Table 2 shows 95% confidence intervals for the percentage suboptimality values recorded for each of the three heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$, (columns 3-5). The first row shows summary results for all 32,934 systems, and the next three rows show results for particular values of N. Both $\theta^{[W]}$ and $\theta^{[B]}$ are consistently strong, with $\theta^{[W]}$ tending to be slightly stronger overall (within 1% of optimality on average). Indeed, $\theta^{[W]}$ was the best-performing of the three heuristics in about 65% of experiments. Noticeably, all of the heuristics tend to do better in the $N = 2$ case than in the $N = 3$ and $N = 4$ cases. In Sect. 6.2, we will present results for larger values of N.

Next, let $\rho := \lambda \left(\sum_{i=1}^{N} c_i \mu_i\right)^{-1}$ be a measure of the relative traffic intensity for a particular system. In Table 3 we have presented results for $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$ in a similar format to that of Table 2, except with results categorized according to ρ rather than N. Our results indicate that $\theta^{[W]}$ tends to be strongest for very small (i.e. close to zero) or very large (i.e. significantly larger than 1) values of ρ - which is unsurprising, since

Table 3 95% confidence intervals for the percentage suboptimality of heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$ (columns 3-5) for different values of $\rho = \lambda \left(\sum_{i=1}^{N} c_i \mu_i\right)^{-1}$

		Pct. suboptimality		
ρ value	Count	$\theta^{[W]}$	$\theta^{[B]}$	$\tilde{\theta}$
All values	32,934	0.659 ± 0.008	1.099 ± 0.013	38.386 ± 0.387
$\rho \in [0, 0.1)$	2141	0.001 ± 0.001	0.043 ± 0.010	0.131 ± 0.038
$\rho \in [0.1, 0.2)$	2229	0.009 ± 0.003	0.199 ± 0.021	1.459 ± 0.217
$\rho \in [0.2, 0.3)$	2171	0.031 ± 0.005	0.263 ± 0.020	3.768 ± 0.364
$\rho \in [0.3, 0.4)$	2202	0.084 ± 0.009	0.356 ± 0.020	8.075 ± 0.566
$\rho \in [0.4, 0.5)$	2204	0.165 ± 0.013	0.450 ± 0.023	12.964 ± 0.686
$\rho \in [0.5, 0.6)$	2173	0.319 ± 0.017	0.532 ± 0.025	17.795 ± 0.773
$\rho \in [0.6, 0.7)$	2185	0.550 ± 0.021	0.696 ± 0.032	22.163 ± 0.825
$\rho \in [0.7, 0.8)$	2187	0.996 ± 0.028	0.957 ± 0.038	27.406 ± 0.902
$\rho \in [0.8, 0.9)$	2240	1.440 ± 0.031	1.210 ± 0.039	34.684 ± 0.905
$\rho \in [0.9, 1)$	2258	1.540 ± 0.035	1.259 ± 0.043	45.309 ± 0.885
$\rho \in [1, 1.1)$	2263	1.375 ± 0.033	1.391 ± 0.050	61.214 ± 0.751
$\rho \in [1.1, 1.2)$	2170	1.101 ± 0.029	1.662 ± 0.051	76.962 ± 0.567
$\rho \in [1.2, 1.3)$	2177	0.909 ± 0.026	2.031 ± 0.051	84.584 ± 0.465
$\rho \in [1.3, 1.4)$	2193	0.696 ± 0.022	2.477 ± 0.051	88.673 ± 0.356
$\rho \in [1.4, 1.5)$	2141	0.595 ± 0.021	2.980 ± 0.052	91.067 ± 0.289

it is known to be asymptotically optimal in light-traffic and heavy-traffic limits due to Theorem 2. Of greater interest, perhaps, are the comparisons with the alternative heuristic policies $\theta^{[W]}$ and $\tilde{\theta}$, and how these are affected by the value of ρ.

Table 3 shows that the suboptimality of $\theta^{[W]}$ is greatest when ρ is close to 1, although it remains within 1.6% of optimality (on average) in such cases. The Bernoulli improvement policy $\theta^{[B]}$ may be slightly stronger than $\theta^{[W]}$ when ρ is close to 1, but (unlike $\theta^{[W]}$) it performs worse as ρ increases beyond 1. This is consistent with the result of Theorem 7. The selfish policy $\tilde{\theta}$ performs well when ρ is very small, but is very poor in other cases.

We also investigated the effect of heterogeneity between service facilities on our results. Recall that a particular facility i in our model has four parameters: c_i, μ_i, α_i and β_i. For each of our 32,934 randomly-generated parameter sets we calculated the coefficient of variation (i.e. the ratio of the standard deviation to the mean) of the values $c_1, \ldots, c_N$ in order to obtain a measure, denoted by ϕ_c, of the variation between c_i values. We then repeated this process for the other parameter types in order to obtain the analogous statistics ϕ_μ, ϕ_α and ϕ_β, and calculated the average coefficient of variation as $\bar{\phi} := (\phi_c + \phi_\mu + \phi_\alpha + \phi_\beta)/4$. Table 4 shows comparisons between the performances and suboptimality values of the three heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$, with results categorized according to the value of $\bar{\phi}$.

Table 4 indicates that $\theta^{[W]}$ remains very strong for all values of $\bar{\phi}$. More interestingly, however, there is a clear trend for the other two heuristics ($\theta^{[B]}$ and $\tilde{\theta}$) to perform

Table 4 95% confidence intervals for the percentage suboptimality of heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$ (columns 3–5) for different values of $\bar{\phi} := (\phi_c + \phi_\mu + \phi_\alpha + \phi_\beta)/4$

		Pct. suboptimality		
$\bar{\phi}$ value	Count	$\theta^{[W]}$	$\theta^{[B]}$	$\tilde{\theta}$
All values	32,934	0.659 ± 0.008	1.099 ± 0.013	38.386 ± 0.387
$\bar{\phi} \in [0, 0.05)$	53	0.684 ± 0.201	0.664 ± 0.249	36.818 ± 10.463
$\bar{\phi} \in [0.05, 0.1)$	389	0.712 ± 0.084	0.526 ± 0.074	24.871 ± 3.289
$\bar{\phi} \in [0.1, 0.15)$	1127	0.725 ± 0.048	0.645 ± 0.048	30.055 ± 2.033
$\bar{\phi} \in [0.15, 0.2)$	2286	0.655 ± 0.033	0.699 ± 0.035	32.118 ± 1.452
$\bar{\phi} \in [0.2, 0.25)$	3838	0.673 ± 0.025	0.844 ± 0.031	34.752 ± 1.120
$\bar{\phi} \in [0.25, 0.3)$	6117	0.674 ± 0.019	0.984 ± 0.027	37.290 ± 0.897
$\bar{\phi} \in [0.3, 0.35)$	7323	0.674 ± 0.017	1.138 ± 0.027	39.677 ± 0.821
$\bar{\phi} \in [0.35, 0.4)$	6403	0.655 ± 0.018	1.263 ± 0.033	40.483 ± 0.874
$\bar{\phi} \in [0.4, 0.45)$	3602	0.638 ± 0.023	1.397 ± 0.048	42.816 ± 1.163
$\bar{\phi} \in [0.45, 0.5)$	1354	0.549 ± 0.036	1.562 ± 0.087	44.347 ± 1.894
$\bar{\phi} \in [0.5, 0.55)$	365	0.473 ± 0.061	1.805 ± 0.193	44.542 ± 3.676
$\bar{\phi} \geq 0.55$	77	0.274 ± 0.088	1.918 ± 0.482	45.588 ± 7.944

worse as $\bar{\phi}$ increases. Consequently, the improvements given by $\theta^{[W]}$ as opposed to $\theta^{[B]}$ and $\tilde{\theta}$ tend to increase with $\bar{\phi}$. Indeed, the Spearman's rank correlation coefficient between $\bar{\phi}$ and $(g^{[W]} - g^{[B]})/g^{[B]}$ is 0.178 when all 32,934 trials are considered, and between $\bar{\phi}$ and $(g^{[W]} - \tilde{g})/\tilde{g}$ (where $\tilde{g}$ is the selfish policy's performance) it is 0.102. These values are statistically highly significant, and it is not obvious why $\theta^{[W]}$ should be more robust to heterogeneity between facilities than the other heuristics. We will return to this subject in our conclusions.

6.2 'Large' systems with $N \geq 4$

We performed a further series of experiments, involving another 4660 randomly-generated sets of system parameters. The intention in this part was to investigate 'large' systems, in which the size of the selfish state space $\tilde{S}$ would preclude the use of dynamic programming algorithms. The median of $|\tilde{S}|$ over all of the 4660 trials performed was approximately 6.25 billion states, and the maximum was approximately 6.03×10^{20} (0.6 sextillion). Please see Appendix H for details of how the parameter sets were generated.

Without the use of DP algorithms, one cannot evaluate the optimal average reward g^* for a given system, nor is it possible to evaluate the performances of the heuristic policies $\theta^{[W]}$, $\theta^{[B]}$ and $\tilde{\theta}$ exactly. However, as discussed in previous sections, the indices used for decision-making by these policies are simple to obtain (regardless of the size of $|\tilde{S}|$), since they are determined by considering facilities individually. One can then use simulation to *estimate* the average rewards earned by the respective heuristic policies.

As Sect. 6.1, our interest lies mainly in assessing the strength of the Whittle policy $\theta^{[W]}$. Given that we cannot evaluate the exact suboptimality of $\theta^{[W]}$ in larger systems, we decided to compensate by expanding our set of alternative heuristic policies to be used for comparison purposes. The results in Sect. 6.1 have already shown that the selfish policy $\tilde{\theta}$ performs poorly in many cases, especially if the demand rate is high. However, we can derive other (possibly stronger) heuristics by considering a simple generalization of the selfish decision-making rule. For a given state $\mathbf{x} \in S$, action $a \in \{0, 1, \ldots, N\}$ and parameter $p \in [0, 1]$, let $w_a(\mathbf{x}, p)$ be defined as follows:

$$w_a(\mathbf{x}, p) = \begin{cases} p\alpha_a - \beta_a/\mu_a, & \text{if } a \in \{1, 2, \ldots, N\} \text{ and } x_a < c_a, \\ p\alpha_a - \beta_a(x_a + 1)/(c_a\mu_a), & \text{if } a \in \{1, 2, \ldots, N\} \text{ and } x_a \geq c_a, \\ 0, & \text{if } a = 0. \end{cases} \tag{32}$$

Thus, $w_a(\mathbf{x}, p)$ is equivalent to the expected net reward for an individual customer defined in (4) except that the rewards α_i for the various facilities are scaled by a multiplier p.

Let $\tilde{\theta}^p$ denote the policy which operates in such a way that the action chosen under state $\mathbf{x} \in S$ is the action which maximizes $w_a(\mathbf{x}, p)$, with ties broken arbitrarily (except that $a = 0$ is chosen only if $w_i(\mathbf{x}, p) < 0$ for all $i \in \{1, 2, \ldots, N\}$). Also, let $\tilde{g}^p$ denote the average reward under policy $\tilde{\theta}^p$. If $p = 1$ then $\tilde{\theta}^p$ is equivalent to the usual selfish policy, $\tilde{\theta}$. However, the value of p which maximizes $\tilde{g}^p$ is likely to be smaller than 1, especially if the demand rate is high.

Let $\mathcal{D}$ be a discretization of the interval [0, 1] and let us define $\tilde{g}^{\mathcal{D}}$ by

$$\tilde{g}^{\mathcal{D}} = \max_{p \in \mathcal{D}} \tilde{g}^p, \tag{33}$$

i.e. $\tilde{g}^{\mathcal{D}}$ is the maximum average reward attained over all possible policies $\tilde{\theta}^p$, subject to the constraint $p \in \mathcal{D}$. We will also use $\tilde{\theta}^{\mathcal{D}}$ to denote a policy in $\{\tilde{\theta}^p\}_{p\in\mathcal{D}}$ which attains the average reward $\tilde{g}^{\mathcal{D}}$. It should be noted that $\tilde{\theta}^{\mathcal{D}}$ is not an admissible policy itself; instead, it represents the strongest-performing of a set of policies for a particular system. We intend to use $\tilde{g}^{\mathcal{D}}$ as a benchmark in order to evaluate the strength of $\theta^{[W]}$ in larger systems.

In each of our 4660 randomly-generated scenarios we simulated the performances of all 100 policies in the set $\{\tilde{\theta}^p\}_{p\in\mathcal{D}}$, with the discretized set $\mathcal{D}$ given by $\mathcal{D} = \{0.01, 0.02, \ldots, 0.99, 1\}$, and estimated $\tilde{g}^{\mathcal{D}}$ by taking the maximum of these. We also simulated the performances of $\theta^{[W]}$ and $\theta^{[B]}$ using the same random number seed used to simulate the 100 policies in $\{\tilde{\theta}^p\}_{p\in\mathcal{D}}$. We note here that simulating the performances of 102 different stationary policies is a computationally intensive task, and this is why we have considered fewer random scenarios in Sect. 6.2 than in Sect. 6.1. The implementation and simulation of the Whittle policy itself is extremely fast even in very large systems, and does not pose any computational difficulty.

Table 5 summarizes the performance of $\theta^{[W]}$ against $\theta^{[B]}$ and $\tilde{\theta}^{\mathcal{D}}$ in these 4460 experiments, with results categorized according to the value of N (the number of

Table 5 95% confidence intervals for the percentage improvement given by $\theta^{[W]}$ against alternative heuristics (columns 3–4) and the percentage of experiments in which $\theta^{[W]}$ matched or exceeded the performances of alternative heuristics (columns 5–6) for different values of N

		Pct. improvement		Pct. of experiments	
N value	Count	$\theta^{[W]}$ vs. $\theta^{[B]}$	$\theta^{[W]}$ vs. $\tilde{\theta}^{\mathcal{D}}$	$g^{[W]} \geq g^{[B]}$	$g^{[W]} \geq \tilde{g}^{\mathcal{D}}$
All values	4660	1.481 ± 0.065	6.044 ± 0.181	72.77	86.33
$N = 4$	773	0.765 ± 0.119	4.444 ± 0.403	67.01	79.56
$N = 5$	772	1.181 ± 0.137	5.476 ± 0.448	70.21	81.99
$N = 6$	669	1.093 ± 0.155	5.501 ± 0.479	68.76	83.86
$N = 7$	569	1.585 ± 0.192	6.381 ± 0.511	72.76	87.52
$N = 8$	475	1.781 ± 0.213	6.871 ± 0.570	76.84	91.37
$N = 9$	443	1.853 ± 0.236	6.421 ± 0.578	74.72	88.04
$N = 10$	343	1.903 ± 0.261	6.779 ± 0.670	74.93	91.55
$N = 11$	319	2.136 ± 0.279	7.450 ± 0.693	81.82	93.10
$N = 12$	297	2.572 ± 0.315	8.015 ± 0.757	81.82	94.61

facilities). Columns 3–4 show the percentage improvements achieved by $\theta^{[W]}$ against the other heuristics for different N values. Column 5 (resp. 6) shows the percentages of experiments in which the (estimated) long-run average reward achieved by the Whittle policy, $g^{[W]}$, was at least as great as $g^{[B]}$ (resp. $\tilde{g}^{\mathcal{D}}$). There is a general trend for $\theta^{[W]}$ to increase its advantages against $\theta^{[B]}$ and $\tilde{\theta}^{\mathcal{D}}$ as N increases. Also, by comparing Tables 2 and 5, we may observe that the relative strength of $\theta^{[W]}$ versus the other heuristics appears to have increased significantly in these 'large system' experiments.

Table 6 shows additional comparisons between the three heuristic policies with results categorized according to the value of $\rho = \lambda \left(\sum_{i=1}^{N} c_i \mu_i\right)^{-1}$. As in Sect. 6.1 (see Table 3), $\theta^{[W]}$ becomes stronger relative to the alternative heuristics as ρ increases beyond 1. The Bernoulli improvement policy, $\theta^{[B]}$, also tends to be stronger than $\tilde{\theta}^{\mathcal{D}}$ (this can be seen by comparing columns 3 and 4, for example).

Finally, Table 7 shows the results of our experiments categorized according to the value of $\bar{\phi}$, where $\bar{\phi}$ is defined the same way as in Sect. 6.1; i.e. it is the average of the coefficients of variation for the four parameter types $(c_i, \mu_i, \alpha_i, \beta_i)$. As in Section 6.1, we observe that $\theta^{[W]}$ tends to increase its advantage over other heuristics when the heterogeneity between facilities is increased.

It seems clear from our results in Sect. 6.2 that, if we want to find an alternative heuristic policy which rivals the performance of the Whittle policy $\theta^{[W]}$, it is not sufficient to simply modify the selfish decision rule so that rewards are given relatively less importance compared to expected waiting costs (and hence the system becomes less busy). Indeed, the Whittle policy is based on *socially* optimal (i.e. average reward optimal) decisions at individual facilities, and this enables it to make smarter decisions than the policies in the set $\{\tilde{\theta}_p\}_{p \in \mathcal{D}}$. To illustrate this point, suppose we have two facilities i and j such that $w_i(\mathbf{x}, p) = w_j(\mathbf{x}, p)$ under a particular state $\mathbf{x} \in S$. Suppose also that the reward for service α_i is substantially larger than α_j, but the expected waiting costs at facility i are also larger (this may be due to a longer expected

Table 6 95% confidence intervals for the percentage improvement given by $\theta^{[W]}$ against alternative heuristics (columns 3–4) and the percentage of experiments in which $\theta^{[W]}$ matched or exceeded the performances of alternative heuristics (columns 5–6) for different values of $\rho = \lambda \left(\sum_{i=1}^{N} c_i \mu_i\right)^{-1}$

		Pct. improvement		Pct. of experiments	
ρ value	Count	$\theta^{[W]}$ vs. $\theta^{[B]}$	$\theta^{[W]}$ vs. $\tilde{\theta}^{\mathcal{D}}$	$g^{[W]} \geq g^{[B]}$	$g^{[W]} \geq \tilde{g}^{\mathcal{D}}$
All values	4660	1.481 ± 0.065	6.044 ± 0.181	72.77	86.33
$\rho \in [0, 0.1)$	315	0.096 ± 0.025	0.047 ± 0.033	89.21	52.38
$\rho \in [0.1, 0.2)$	301	0.122 ± 0.032	0.326 ± 0.098	73.75	53.16
$\rho \in [0.2, 0.3)$	325	0.066 ± 0.041	0.972 ± 0.222	61.85	66.46
$\rho \in [0.3, 0.4)$	335	0.010 ± 0.054	1.613 ± 0.224	49.25	77.31
$\rho \in [0.4, 0.5)$	299	0.042 ± 0.079	2.401 ± 0.337	48.83	80.6
$\rho \in [0.5, 0.6)$	341	0.208 ± 0.094	3.884 ± 0.367	55.72	91.79
$\rho \in [0.6, 0.7)$	308	0.262 ± 0.122	5.108 ± 0.490	53.57	94.48
$\rho \in [0.7, 0.8)$	295	0.286 ± 0.137	6.006 ± 0.559	55.93	93.56
$\rho \in [0.8, 0.9)$	308	0.251 ± 0.160	7.130 ± 0.579	55.52	95.78
$\rho \in [0.9, 1)$	311	0.954 ± 0.195	9.054 ± 0.714	67.52	95.5
$\rho \in [1, 1.1)$	314	2.083 ± 0.187	9.923 ± 0.615	87.9	98.73
$\rho \in [1.1, 1.2)$	278	3.328 ± 0.199	10.164 ± 0.706	97.12	98.92
$\rho \in [1.2, 1.3)$	315	4.239 ± 0.165	11.032 ± 0.681	100	99.05
$\rho \in [1.3, 1.4)$	300	5.042 ± 0.160	11.935 ± 0.750	99.67	99.67
$\rho \in [1.4, 1.5)$	315	5.585 ± 0.136	12.058 ± 0.702	100	99.68

Table 7 95% confidence intervals for the percentage improvement given by $\theta^{[W]}$ against alternative heuristics (columns 3–4) and the percentage of experiments in which $\theta^{[W]}$ equalled or exceeded the performances of alternative heuristics (columns 5–6) for different values of $\bar{\phi} = (\phi_c + \phi_\mu + \phi_\alpha + \phi_\beta)/4$

		Pct. improvement		Pct. of experiments	
$\bar{\phi}$ value	Count	$\theta^{[W]}$ vs. $\theta^{[B]}$	$\theta^{[W]}$ vs. $\tilde{\theta}^{\mathcal{D}}$	$g^{[W]} \geq g^{[B]}$	$g^{[W]} \geq \tilde{g}^{\mathcal{D}}$
All values	4660	1.481 ± 0.065	6.044 ± 0.181	72.77	86.33
$\bar{\phi} \in [0, 0.2)$	20	0.263 ± 0.745	1.787 ± 1.617	45.00	70.00
$\bar{\phi} \in [0.2, 0.25)$	101	0.433 ± 0.369	3.028 ± 0.758	53.47	80.20
$\bar{\phi} \in [0.25, 0.3)$	429	0.833 ± 0.178	4.347 ± 0.491	64.57	80.19
$\bar{\phi} \in [0.3, 0.35)$	1190	1.209 ± 0.124	5.053 ± 0.317	68.99	84.71
$\bar{\phi} \in [0.35, 0.4)$	1758	1.594 ± 0.107	6.286 ± 0.291	74.74	88.05
$\bar{\phi} \in [0.4, 0.45)$	967	1.944 ± 0.154	7.557 ± 0.447	77.56	89.04
$\bar{\phi} \in [0.45, 0.5)$	180	1.888 ± 0.315	7.887 ± 1.106	85.00	85.00
$\bar{\phi} \in [0.5, 0.55)$	15	2.262 ± 0.912	11.082 ± 4.333	86.67	93.33

waiting time, for example). In this situation, the generalized selfish policy $\tilde{\theta}_p$ is unable to distinguish between facilities i and j, but one might imagine that joining facility j should be a better choice in the context of average reward maximization, since it has a smaller impact on future congestion levels in the system. The index (13) employed by the Whittle policy is better-suited to taking such considerations into account.

The randomly-generated parameter sets and results of the numerical experiments reported in this paper have been archived and are available at http://doi.org/10.5281/zenodo.3775332.

7 Conclusions

Theorem 2 in Shone et al. (2016) has established that it is theoretically possible to find an average reward optimal policy for the MDP formulated in Sect. 2 by truncating the state space S, and applying a dynamic programming algorithm to an MDP with the *finite* state space $\tilde{S}$. Unfortunately, the finite set $\tilde{S}$ might itself be very large in many problem instances, and for this reason it is necessary to look for heuristic approaches which can be relied upon to yield *near-optimal* policies in a short amount of time.

As discussed in the introduction, the Whittle index heuristic is now well-established in the field of stochastic dynamic programming and we have shown (Lemma 1) that the indexability property, which can be difficult to prove in other settings, can be applied in our problem. A key finding of our paper is that the positive recurrent state space under an optimal stationary policy is not only bounded above by the selfish state space $\tilde{S}$, but also bounded below by the 'Whittle state space' S_W (Theorem 1). We have also proved certain structural properties of the Whittle policy, including its asymptotic optimality in light-traffic and heavy-traffic limits (Theorems 2–6). These results are useful since, in general, structural properties of optimal policies are difficult to prove for routing problems involving heterogeneous service facilities.

The empirical results in Sect. 6 have shown that the Whittle policy $\theta^{[W]}$ is very close to optimality in systems which are 'small enough' to allow the computation of an optimal policy. In larger systems, we have verified that it performs strongly against alternative heuristics, including the policy $\theta^{[B]}$ obtained by applying one step of policy improvement to a 'Bernoulli splitting' policy. Notably, its superiority over other heuristics appears to increase as

(i) the traffic intensity increases beyond 1;
(ii) the number of facilities increases;
(iii) the heterogeneity between service facilities increases.

The first of the above characteristics is implied by Theorem 2, but the reasons for the second and third characteristics are less obvious. Indeed, all of the heuristics that we have considered share some broad methodological similarities, in that they require the computation of indices for individual facilities—so it is not immediately clear why the Whittle policy's indices should be more robust than others with respect to dimensionality or heterogeneity. We intend to investigate this further in future work.

For any given set of system parameters, the indices which characterize the Whittle policy $\theta^{[W]}$ are calculated in a completely deterministic way. Thus, this heuristic does not rely on any iterative algorithm, nor does it involve any type of simulation or random sampling. One might regard the deterministic nature of this heuristic as both a strength and a weakness. On one hand, the simplicity makes it extremely easy to implement; on the other hand, if the heuristic is found to perform *poorly* in a particular system, then it is not necessarily easy to see how the decision-making indices might be adjusted in order to achieve closer proximity to an optimal policy. In future work, we intend to test the performance of the Whittle heuristic against policies obtained by approximate dynamic programming (ADP) methods, including those which have achieved popularity in the fields of neuro-dynamic programming and reinforcement learning (Bertsekas and Tsitsiklis 1996; Powell 2007; Sutton and Barto 1998). We also intend to use the Whittle policy in conjunction with ADP methods, by allowing it to act as a reference point within broader search algorithms.

Acknowledgements This research has been funded by the EPSRC (grant EP/J500197/1). The authors would like to express their gratitude to the associate editor and anonymous reviewers for their astute and constructive suggestions which have improved this paper significantly. They would also like to thank Kevin Glazebrook for useful discussions which have greatly assisted the development of this work. Thanks are also due to Will Shone and Christopher Campfield for their assistance in running numerical experiments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aalto S, Lassila P, Osti P (2016) Whittle index approach to size-aware scheduling for time-varying channels with multiple states. Queueing Syst 83:195–225

Ansell PS, Glazebrook KD, Kirkbride C (2003a) Generalised "join the shortest queue" policies for the dynamic routing of jobs to multi-class queues. J Oper Res Soc 54:379–389

Ansell PS, Glazebrook KD, Nino-Mora J, O'Keeffe M (2003b) Whittle's index policy for a multi-class queueing system with convex holding costs. Math Methods Oper Res 57:21–39

Archibald TW, Black DP, Glazebrook KD (2009) Indexability and index heuristics for a simple class of inventory routing problems. Oper Res 57:314–326

Argon NT, Ding L, Glazebrook KD, Ziya S (2009) Dynamic routing of customers with general delay costs in a multiserver queuing system. Probab Eng Inf Sci 23(2):175–203

Bell CE, Stidham S (1983) Individual versus social optimization in the allocation of customers to alternative servers. Manag Sci 29(7):831–839

Bellman RE (1957) Dynamic programming. Princeton University Press, Princeton

Bertsekas DP, Tsitsiklis JN (1996) Neuro-dynamic programming. Athena Scientific, Nashua

Bertsimas D, Nino-Mora J (1996) Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems. Math Oper Res 21:257–306

Bhulai S, Koole G (2003) On the structure of value functions for threshold policies in queueing models. J Appl Probab 40:613–622

Borkar VS (2017) Whittle index for partially observed binary Markov decision processes. IEEE Trans Autom Control 62:6614–6618

Borkar VS, Pattathil S (2017) Whittle indexability in egalitarian processor sharing systems (available online). Ann Oper Res. https://doi.org/10.1007/s10479-017-2622-0

Ford S, Atkinson MP, Glazebrook KD, Jacko P (2020) On the dynamic allocation of assets subject to failure. Eur J Oper Res 284:227–239

Gittins JC (1979) Bandit processes and dynamic allocation indices. J R Stat Soc B41:148–177

Glazebrook KD, Hodge DJ, Kirkbride C (2011) General notions of indexability for queueing control and asset management. Ann Appl Probab 21(3):876–907

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

Glazebrook KD, Hodge DJ, Kirkbride C, Minty RJ (2014) Stochastic scheduling: a short history of index policies and new approaches to index generation for dynamic resource allocation. J Sched 17(5):407–425

Grassmann W (1983) The convexity of the mean queue size of the M/M/c queue with respect to the traffic intensity. J Appl Probab 20:916–919

Gross D, Harris C (1998) Fundamentals of queueing theory. Wiley, New York

Ha A (1997) Optimal dynamic scheduling policy for a make-to-stock production system. Oper Res 45:42–53

Hassin R, Haviv M (2003) To queue or not to queue: equilibrium behavior in queueing systems. Kluwer Academic Publishers, Norwell

Haviv M, Roughgarden T (2007) The price of anarchy in an exponential multi-server. Oper Res Lett 35(4):421–426

Hodge DJ, Glazebrook KD (2011) Dynamic resource allocation in a multi-product make-to-stock production system. Queueing Syst 67(4):333–364

Hordijk A, Koole G (1990) On the optimality of the generalised shortest queue policy. Probab Eng Inf Sci 4(4):477–487

Hordijk A, Koole G (1992) On the assignment of customers to parallel queues. Probab Eng Inf Sci 6(4):495–511

Howard R (1960) Dynamic programming and markov processes. MIT Press, Cambridge

Hsu Y-P (2018) Age of information: whittle index for scheduling stochastic arrivals. In: Proceedings of the 2018 IEEE international symposium on information theory (ISIT), Vail, CO

Hyytia E, Aalto S, Penttinen A, Virtamo J (2012) On the value function of the $M/G/1$, FCFS and LCFS queues. J Appl Probab 49:1052–1071

Knight VA, Harper PR (2013) Selfish routing in public services. Eur J Oper Res 230(1):122–132

Knight VA, Komenda I, Griffiths JD (2017) Measuring the price of anarchy in critical care unit interactions. J Oper Res Soc 68:630–642

Koole G, Sparaggis PD, Towsley D (1999) Minimizing response times and queue lengths in systems of parallel queues. J Appl Probab 36:1185–1193

Krishnan KR (1990) Joining the right queue: a state-dependent decision rule. IEEE Trans Autom Control 35:104–108

Larranaga M, Ayesta U, Verloop IM (2016) Dynamic control of birth-and-death restless bandits: application to resource-allocation problems. IEEE/ACM Trans Netw 24(6):3812–3825

Lee HL, Cohen AM (1983) A note on the convexity of performance measures of M/M/c queueing systems. J Appl Probab 20:920–923

Li D, Ding L, Connor S (2020) When to switch? Index policies for resource scheduling in emergency response. Prod Oper Manag 29(2):241–262

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

Menich R, Serfozo R (1991) Optimality of shortest-queue routing for dependent service stations. Queueing Syst 9:403–418

Nino-Mora J (2001) Restless bandits, partial conservation laws and indexability. Adv Appl Probab 33:76–98

Nino-Mora J (2002) Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. Math Program 93(3):361–413

Nino-Mora J (2012) Towards minimum loss job routing to parallel heterogeneous multiserver queues via index policies. Eur J Oper Res 220:705–715

Powell WB (2007) Approximate dynamic programming: solving the curses of dimensionality. Wiley, New York

Puterman ML (1994) Markov decision processes—discrete stochastic dynamic programming. Wiley, New York

Sassen SAE, Tijms HC, Nobel RD (1997) A heuristic rule for routing customers to parallel servers. Stat Neerl 51:107–121

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

Shone R, Knight VA, Harper PR, Williams JE, Minty J (2016) Containment of socially optimal policies in multiple-facility Markovian queueing systems. J Oper Res Soc 67:629–643

Sutton R, Barto A (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Weber RR (1978) On the optimal assignment of customers to parallel servers. J Appl Probab 15:406–413

Whitt W (1986) Deciding which queue to join: some counterexamples. Oper Res 34(1):55–62

Whittle P (1988) Restless bandits: activity allocation in a changing world. J Appl Probab 25:287–298

Winston W (1977) Optimality of the shortest line discipline. J Appl Probab 14:181–189