The Annals of Applied Probability
2000, Vol. 10, No. 3, 877–896

INDEX-BASED POLICIES FOR DISCOUNTED MULTI-ARMED BANDITS ON PARALLEL MACHINES[1]

BY K. D. GLAZEBROOK AND D. J. WILKINSON

Newcastle University

We utilize and develop elements of the recent achievable region account of Gittins indexation by Bertsimas and Niño-Mora to design index-based policies for discounted multi-armed bandits on parallel machines. The policies analyzed have expected rewards which come within an $O(\alpha)$ quantity of optimality, where $\alpha > 0$ is a discount rate. In the main, the policies make an initial once for all allocation of bandits to machines, with each machine then handling its own workload optimally. This allocation must take careful account of the index structure of the bandits. The corresponding limit policies are average-overtaking optimal.

1. Introduction. Ever since Gittins and Jones (1974) proved the classical result establishing the optimality of Gittins index policies for multi-armed bandits with discounted rewards earned over an infinite horizon, it has been widely believed that policies based on such indices will perform very well when the single machine/server of the Gittins and Jones model is replaced by a collection of identical machines/servers working in parallel. Exploration of such issues goes back to Glazebrook (1976). It has become clear that parallel machine stochastic scheduling problems are much less tractable in general than their single machine counterparts. See, for example, Weber (1982), Weber, Varaiya and Walrand (1986) and Weiss (1990, 1992). Weiss (1995) has recently given an account of index policies for a problem involving the scheduling of a batch of stochastic jobs on parallel machines with a linear holding cost objective.

New approaches to the analysis of such problems have emerged from recent research on the so-called achievable region approach to stochastic optimization. This approach develops solutions to stochastic optimization problems by (1) characterising the space of all possible performances (the achievable region) of the system of interest, and (2) optimizing the overall system-wide objective over this space. Following foundational contributions by Coffman and Mitrani (1980) and Shanthikumar and Yao (1992), Bertsimas and Niño-Mora (1996) took this approach decisively further forward and gave an account of Gittins indices from this perspective.

Recently, Glazebrook and Garbe (1999) explained how the achievable region approach could be deployed to analyze Gittins index policies for systems in which the conditions sufficient to establish that such policies are fully op-

Received July 1999; revised November 1999.

[1]Supported by the Engineering and Physical Sciences Research Council Grant GR/MO9308.

AMS 1991 *subject classifications*. Primary 90B36; secondary 90C40.

Key words and phrases. Average-overtaking optimal, average-reward optimal, Gittins index, multi-armed bandit problem, parallel machines, suboptimality bound.

timal fail narrowly. They analysed a general discounted multi-armed bandit problem on parallel machines subject only to the condition that the Markov chain modelling the evolution of each arm is irreducible and has finite state space. In a continuous time formulation in which discounting is of the form $e^{-\alpha t}$ and hence total expected rewards from the implementation of any policy are $O(1/\alpha)$ where $\alpha > 0$ is a discount rate, Glazebrook and Garbe (1999) showed that a Gittins index policy implemented in the parallel machine environment earns a reward which comes within $O(1)$ of optimality. This in turn implies that for small enough discount rate, the corresponding index policy is average-reward optimal.

The current paper strengthens the available tools of analysis (see Section 2) in a way that facilitates stronger results for the parallel machine version of the discounted multi-armed bandit problem than those described in the previous paragraph. In Section 3 we are able to prove the remarkable fact that it is possible to design an initial once for all allocation of bandits to machines such that the policy which then ensures that each machine handles its own allocated workload optimally (i.e., uses a Gittins index policy) has a reward which comes within $O(\alpha)$ of optimality. This initial allocation must take account of the index structure of the bandits. Much practical and theoretical interest attaches to cases with small discount rate α and to the related limit $\alpha \to 0$. Plainly the $O(\alpha)$ result implies, inter alia, that as discounting disappears ($\alpha \to 0$) so does the suboptimality gap associated with the policy. The model structures elucidated by the analysis in Sections 2 and 3 are deployed in Section 4 to strengthen the Glazebrook and Garbe (1999) result cited above for certain classes of model. We obtain an $O(\alpha)$ suboptimality gap for a Gittins index policy implemented in the parallel machine environment. In Section 5 we show that these $O(\alpha)$ results imply that the corresponding limit policies (i.e., policies of equivalent structure but based upon limiting index values as $\alpha \to 0$) are 1-optimal and average-overtaking optimal for our multi-armed bandit problem with parallel machines. Finally, we discuss some instances of the index-based policies of Section 3 in which there may be many initial allocations of bandits to machines for which the $O(\alpha)$ result is available. We develop a load balancing problem in the form of an integer program whose solution will guide the choice of initial allocation.

2. The model and tools of analysis. M identical machines are available to process B projects or "bandits" as we shall call them, where $B > M \geq 1$. Bandits are of Q types and n_q bandits of type q are in the system, where $\sum_q n_q = B$. A bandit of type q has finite state space E_q. The *state* of the system is a B-vector whose b^{th} component is the state of bandit b. Hence the (finite) state space for the system is $\times_q (E_q)^{n_q}$. Write $E = \bigcup_q E_q$. Please note that we have allowed for the possibility of there being many bandits of a single type in order to facilitate the analysis later in the paper. See Lemma 4, Section 4 and the discussion of load balancing in Section 5.

At each decision epoch $t = 0, 1, 2, \ldots$ M bandits are chosen for processing, one on each machine. Should bandit b of type q be chosen for processing at

time t when in state $i \in E_q$ then with probability P_{ij}^q it will be in state $j \in E_q$ at time $t+1$. This type q transition law is Markovian and distinct bandits are assumed to evolve independently of each other. The $B - M$ bandits *not* chosen for processing at t remain stationary. The Markov chains determined by the Q one-step transition matrices P^q (such a chain would be observed in real time were a type q bandit to be processed without interruption) are assumed irreducible and hence positive recurrent. Standard results indicate that all first passage times $T_{ij},\ i, j \in E_q$ in the Markov chain with transition matrix P^q have a distribution with a tail no heavier than geometric and hence have all positive moments finite. Please note that this assumption of positive recurrence plays a central role in the analysis. Hence, for example, bandits which terminate a period of active evolution by entering an absorbing state are not covered by our results.

We study the classical multi-armed bandit problem with the discounted reward criterion. To simplify our computations, we shall assume that a reward of $r_i \int_t^{t+1} e^{-\alpha s} ds$ is earned when a bandit of type q is chosen for processing at time t while in state i. The constant $\alpha > 0$ is a discount rate. In this event, we shall use the terminology "a job of type i is processed at time t." Hence we use "job type" to indicate members of E. Rewards are additive across machines and over time. A (nonanticipative) policy π is a rule which specifies an allocation of M distinct bandits to the machines at each decision epoch as a function of the history of the process to date. We shall express the total expected reward earned under policy π from initial state $\mathbf{k} \in \times_q (E_q)^{n_q}$ as

$$R^{\pi}(\mathbf{k}) = \sum_{i \in E} r_i x_i^{\pi}(\mathbf{k}) \tag{1}$$

where

$$x_i^{\pi}(\mathbf{k}) = E_{\pi}\left\{ \int_0^{\infty} n_i(t) e^{-\alpha t} dt | \mathbf{k} \right\},\ i \in E. \tag{2}$$

In (2) we take $n_i(t)$ to be the number of jobs of type i being processed at t.

Our goal is the analysis of the stochastic optimization problem

$$R^{OPT}(\mathbf{k}) = \sup_{\pi} \sum_{i \in E} r_i x_i^{\pi}(\mathbf{k}). \tag{3}$$

We will drop initial state $\mathbf{k}$ from the notation on occasion when no confusion arises.

We now proceed to describe objects and ideas utilized by Bertsimas and Niño-Mora (1996) in their analysis of the $M = 1$ case. Fix $i \in E_q$ and subset $S \subseteq E$ containing i. Consider the evolution of a Markov chain with initial state i and one-step transition matrix P^q. Write $X_q(t)$ for the state of the Markov Chain at time t. We define

$$T_i^{S^c} \equiv \inf\left\{ t; t \geq 1 \text{ and} X_q(t) \in S \cap E_q \right\}$$

as the first time after $t = 1$ at which the chain leaves S^c. Plainly $T_i^{S^c} \leq T_{ii}$ a.s. and hence all positive moments of $T_i^{S^c}$ are finite. Following Bertsimas and

Niño-Mora (1996), we define the matrix $\mathbf{A} \equiv (A_i^S)_{i\in E, S\subseteq E}$ as follows:

$$A_i^S = \begin{cases} 0, & i \notin S \\ \left[1 - E\left\{e^{-\alpha T_i^{S^c}}\right\}\right] \Big/ (1 - e^{-\alpha}), & i \in S. \end{cases} \tag{4}$$

As will emerge in the course of the analysis, the quantities

$$A^\pi(S, \mathbf{k}) \equiv \sum_{i\in S} A_i^S x_i^\pi(\mathbf{k}) \tag{5}$$

play a central role. Plainly, for each subset S, (5) corresponds to an objective of the form (1) in which entries from the matrix $\mathbf{A}$ give the reward rates.

Bertsimas and Niño-Mora (1996) describe a so-called *adaptive greedy algorithm* $AG(\mathbf{A}, \mathbf{r})$ whose inputs are the matrix $\mathbf{A}$ and reward vector $\mathbf{r}$ and whose outputs include a set of non-negative reals $G_i,\ i \in E$ called *generalised Gittins indices*. In the current context, the Gittins index G_i for a job of type $i \in E_q$ is a measure of the rate at which rewards can accrue from a bandit of type q currently in state i. See Bertsimas and Niño-Mora (1996) for more details. We shall suppose that the members of E are labelled $\{1, 2, \ldots, |E|\}$ such that

$$G_{|E|} \geq G_{|E|-1} \geq \cdots \geq G_2 \geq G_1.$$

We write $S_j = \{j, j-1, \ldots, 1\}$ for the subset of E of cardinality j with the lowest Gittins indices. Note that in none of the results in the paper does it matter how ties are broken between job types of equal index when some policy based on the index values is used. While we do not need full details of the algorithm $AG(\mathbf{A}, \mathbf{r})$ here, we shall require the following key fact regarding its structure:

$$\begin{aligned} r_i &= G_{|E|} A_i^E - \sum_{j=i}^{|E|-1} (G_{j+1} - G_j) A_i^{S_j} \\ &= G_{|E|} - \sum_{j=i}^{|E|-1} (G_{j+1} - G_j) A_i^{S_j},\ i \in E. \end{aligned} \tag{6}$$

Equation (6) uses the fact [plain from (4)], that $A_i^E = 1, i \in E$. It now yields the following important calculation lemma which we shall use frequently.

LEMMA 1. *For all policies π and initial states $\mathbf{k}$,*

$$R^\pi(\mathbf{k}) = G_{|E|}\left(\frac{M}{\alpha}\right) - \sum_{j=1}^{|E|-1} (G_{j+1} - G_j) A^\pi(S_j, \mathbf{k}). \tag{7}$$

PROOF. We utilize (6) to obtain

$$\begin{aligned} R^\pi(\mathbf{k}) &= \sum_{i\in E} r_i x_i^\pi(\mathbf{k}) \\ &= G_{|E|} \sum_{i\in E} x_i^\pi(\mathbf{k}) - \sum_{i\in S_{|E|-1}} \sum_{j=i}^{|E|-1} (G_{j+1} - G_j) A_i^{S_j} x_i^\pi(\mathbf{k}). \end{aligned} \tag{8}$$

To obtain (7) note first from (2) that

$$\sum_{i \in E} x_i^{\pi}(\mathbf{k}) = \frac{M}{\alpha}$$

for all choices of π and $\mathbf{k}$, then interchange the order of summation in the second term on the r.h.s. of (8) and finally use (5). □

We now develop a class of stochastic optimization problems, one for each subset S, given by

$$A(S, \mathbf{k}) \equiv \inf_{\pi} A^{\pi}(S, \mathbf{k}). \tag{9}$$

The following is an immediate consequence of Lemma 1.

COROLLARY 1. *For all initial states* $\mathbf{k}$,

$$R^{OPT}(\mathbf{k}) \leq G_{|E|}\left(\frac{M}{\alpha}\right) - \sum_{j=1}^{|E|-1} (G_{j+1} - G_j) A(S_j, \mathbf{k}). \tag{10}$$

In the single machine case $M = 1$, (10) can be shown to be satisfied with equality. This is a consequence of the fact, established by Bertsimas and Niño-Mora (1996), that for each $S \subseteq E$, the infimum in (9) is achieved by any policy π which chooses bandits for processing in such a way that job types in S^c are always given priority over those in S. We write $\pi : S^c \to S$ for any such policy. But any Gittins index policy π_G which always chooses bandits consistent with the ordering $|E| \to |E| - 1 \to \cdots \to 2 \to 1$ is such that $\pi_G : S_j^c \to S_j$ for all j. Hence we have

$$A^{\pi_G}(S_j, \mathbf{k}) = A(S_j, \mathbf{k}), \quad 1 \leq j \leq |E| - 1, \tag{11}$$

and the optimality of π_G when $M = 1$ follows from Lemma 1, Corollary 1 and (11).

In the parallel machine case $M > 1$ we do not have (11) in general for the policy π_G which chooses at each decision epoch the M bandits whose current states have the highest indices among those present. However, as we shall see, the difference between the two quantities in (11) is often small, implying that $R^{OPT} - R^{\pi_G}$ is also small. We shall also see that there are other policies based on Gittins indices which come close to optimality. The suboptimality bound presented in Theorem 1 is our main tool for analysing the policies based on Gittins indices described in the next sections. It is a trivial consequence of Lemma 1 and Corollary 1.

THEOREM 1 (Suboptimality bound for policy π).

$$R^{OPT}(\mathbf{k}) - R^{\pi}(\mathbf{k}) \leq \sum_{j=1}^{|E|-1} (G_{j+1} - G_j) \left\{ A^{\pi}(S_j, \mathbf{k}) - A(S_j, \mathbf{k}) \right\}. \tag{12}$$

In order to use Theorem 1 to obtain suboptimality bounds for given policies we shall need access to the quantities $A(S_j, \mathbf{k})$ or to lower bounds for them. The lower bounds which will be utilized in our analyses of index-based policies described in Sections 3 and 4 will be given in Lemma 2 below. To prepare for this result we need some additional notation. We write

$$g_q = \min\{j;\ j \in E_q\}$$

for the state of lowest index in E_q. Now label the bandit types $\{1, 2, \ldots, Q\}$ such that

$$g_Q > g_{Q-1} > \cdots > g_1 = 1,$$

and define $\psi_j,\ j \in E$, by

$$\psi_j = \begin{cases} 0, & j \geq g_Q, \\ \sum_{r=q}^{Q} n_r, & g_q - 1 \geq j \geq g_{q-1},\ Q \geq q \geq 2. \end{cases} \tag{13}$$

It plainly follows from (13) that

$$0 = \psi_{|E|} \leq \psi_{|E|-1} \leq \cdots \leq \psi_1 = N - n_1$$

and ψ_j has the interpretation as the total number of bandits whose smallest state is indexed greater than j and therefore the total number of bandits whose state spaces have no intersection with S_j. Denote by B_j the collection of bandits with these properties, $j \in E$. In cases where there are no ties between index values, ψ_j is also the number of bandits all of whose associated Gittins indices exceed G_j.

Suppose that j is such that $\psi_j = |B_j| \geq M$. Under this condition there exist policies π which only process bandits in B_j and hence never process a job type in S_j. Plainly, for all such π,

$$A^{\pi}(S_j, \mathbf{k}) = 0,$$

which in turn implies that

$$A(S_j, \mathbf{k}) = 0.$$

Now consider the case $0 \leq \psi_j \leq M - 1$. Let bandit $b \notin B_j$ be of type q and have initial state k_b. Write $X_q(t)$ for the state at time t of a Markov Chain with initial state k_b and one-step transition matrix P^q. We define

$$T(k_b, S_j^c) \equiv \begin{cases} 0, & k_b \in S_j, \\ \inf\{t;\ t \geq 1 \text{ and } X_q(t) \in S_j\}, & k_b \notin S_j, \end{cases}$$

as the amount of processing required for bandit b to escape S_j^c for the first time. We then write

$$T(\mathbf{k}, S_j^c) = \sum_{b \notin B_j} T(k_b, S_j^c) \tag{14}$$

for the total processing required for all bandits outside of B_j to escape S_j^c. Note that the summands in (14) are independent random variables.

LEMMA 2. (i) *If* $0 \le \psi_j \le M-1$ *then*

$$A(S_j, \mathbf{k}) \ge \frac{(M-\psi_j)}{\alpha} E\big[\exp\{-\alpha T(\mathbf{k}, S_j^c)/(M-\psi_j)\}\big].$$

(ii) *If* $\psi_j \ge M$ *then*

$$A(S_j, \mathbf{k}) = 0.$$

PROOF. The proof of part (ii) is trivial and is contained in the above text. Hence we suppose that $0 \le \psi_j \le M-1$ and prove part (i). We fix policy π and denote by ν_i the number of times (which may be infinite) that a job of type i is chosen under π and by $\tau_{i,l}$, $1 \le l \le \nu_i$, the times at which this occurs. With each $\tau_{i,l}$ is associated a random variable $T_{i,l}^{S_j^c}$ which is the length of the subsequent excursion of the corresponding bandit into S_j^c. For a given i, the $T_{i,l}^{S_j^c}$ are independent and identically distributed. All share the distribution of $T_i^{S_j^c}$, defined in the preamble to (4). It is straightforward to show from the definitions of the quantities involved that

$$A^\pi(S_j, \mathbf{k}) = \sum_{i \in S_j} A_i^{S_j} x_i^\pi(\mathbf{k}) = \sum_{i \in S_j} E_\pi \left(\sum_{l=1}^{\nu_i} e^{-\alpha \tau_{i,l}} \int_0^{T_{i,l}^{S_j^c}} e^{-\alpha t} dt | \mathbf{k} \right). \tag{15}$$

We infer from (15) that

$$A(S_j, \mathbf{k}) \ge \inf_\pi \sum_{b=1}^{B} \sum_{m=1}^{M} E_\pi \left[\int_0^\infty I_{bm}(t) e^{-\alpha t} dt | \mathbf{k} \right] \tag{16}$$

where

$$I_{bm}(t) = \begin{cases} 1, & \text{if machine } m \text{ processes bandit } b \text{ at time } t, \text{ and where } b \text{ has} \\ & \text{paid its first visit to } S_j \text{ at some time } s \le t, \\ 0, & \text{otherwise.} \end{cases}$$

Inequality (16) follows from (15) by means of the following two observations: firstly, the processing which bandit b receives from some $\tau_{i,l}$, $i \in S_j$, $1 \le l \le \nu_i$, until b next enters S_j will be delivered during $[\tau_{i,l}, \tau_{i,l} + T_{i,l}^{S_j^c})$ at the earliest. Secondly, once a bandit has paid its first visit to S_j, subsequent visits to S_j will alternative with (possibly null) excursions into S_j^c.

We proceed to bound the r.h.s. of (16) below by considering a relaxation of the minimization problem as follows: denote by the pair (m, t), $1 \le m \le M$, $t = 0, 1, 2, \ldots$ the decision opportunity afforded on machine m at time t. Consider the mapping onto $\mathbb{N}$ given by

$$(m, t) \to \phi(m, t) \equiv Mt + (m-1)$$

in which this decision opportunity is thought to occur on a single machine at time $\phi(m,t)$. We relax the minimization in (16) by considering the class Π of "single machine" policies defined below. For this purpose we require identifiers for the bandits in B_j. (Note that in Section 2 we developed specified numberings for job types and for bandit types, but not for individual bandits.) For simplicity, we number the bandits in B_j from 1 to ψ_j. The reader should recall that $\psi_j < M$. We now develop our "single machine" relaxation as follows:

(i) a policy $\pi' \in \Pi$ can only schedule bandit b, $1 \le b \le \psi_j$, at times $\phi(b,t)$, $t \ge 0$;

(ii) subject only to (i) a policy $\pi' \in \Pi$ schedules a single bandit at each decision epoch in $\mathbb{N} = \{\phi(m,t); 1 \le m \le M, t \ge 0\}$;

(iii) the allocation at $\phi(m,t)$ under $\pi' \in \Pi$ attracts discounting equal to that at time t in the original parallel machine problem.

Note that (i)–(iii) are equivalent to building a policy for the parallel machine problem by allocating bandits to machines in numerical order at each successive time point with repeat allocations allowed, subject to the requirement that each bandit $b, 1 \le b \le \psi_j$, may only be allocated to machine b. Naturally, any such policy which does schedule a bandit more than once at some t is non-admissible for the parallel machine problem. Under this single machine relaxation it is clear from (16) that

$$A(S_j,\mathbf{k}) \ge \inf_{\pi'\in\Pi} \sum_{b=1}^{B}\sum_{m=1}^{M}\sum_{t=0}^{\infty} E_\pi\left[\left\{\int_t^{t+1} e^{-\alpha s}ds\right\}\left\{\int_{\phi(m,t)}^{\phi(m,t)+1} I_b(s)\,ds\right\}|\mathbf{k}\right] \tag{17}$$

where

$$I_b(s) = \begin{cases} 1, & \text{if bandit } b \text{ is processed at time } s, \text{ having paid its first visit to } S_j \text{ at some time } u \le s, \\ 0, & \text{otherwise.} \end{cases}$$

It is straightforward to obtain policies which achieve the infimum in the single machine problem (17). Firstly, note that since the bandits $b,\ 1 \le b \le \psi_j$, belong to B_j, they by definition can never visit S_j and hence cannot contribute to the objective on the r.h.s. of (17). From this it is trivial to demonstrate that any π' attaining the infimum in (17) must choose these bandits whenever possible—that is, it must choose bandit b at all times $\phi(b,t),\ t \ge 0,\ 1 \le b \le \psi_j$. Hence we can replace (17) by

$$A(S_j,\mathbf{k}) \ge \inf_{\pi'\in\Pi} \sum_{b=\psi_j+1}^{B}\sum_{m=\psi_j+1}^{M}\sum_{t=0}^{\infty} E_\pi\left[\left\{\int_t^{t+1} e^{-\alpha s}ds\right\} \times\left\{\int_{\phi(m,t)}^{\phi(m,t)+1} I_b(s)\,ds\right\}|\mathbf{k}\right] \tag{18}$$

to give us a minimization involving $B-\psi_j$ bandits, all of which can enter S_j. A simple interchange argument very similar to one used by Bertsimas and

Niño-Mora (1996) serves to show that a minimizing $\pi' \in \Pi$ will make choices at the remaining decision epochs $\{\phi(m,t);\ \psi_j + 1 \le m \le M;\ t \ge 0\}$ which enforce the priority $S_j^c \to S_j$. Write

$$T(\mathbf{k}, S_j^c) = \Big[T(\mathbf{k}, S_j^c)/(M-\psi_j)\Big](M-\psi_j) + R$$

where $0 \le R \le M - \psi_j - 1$ and $[x]$ is the integer part of x. The argument in the preceding paragraph enables us to compute the infimum in (18). A simple calculation yields

$$\begin{aligned} A(S_j, \mathbf{k}) &\ge \frac{\{\mathrm{Re}^{-\alpha} + (M-\psi_j - R)\}}{\alpha} E\Big(\exp\big\{-\alpha\big[T(\mathbf{k}, S_j^c)/(M-\psi_j)\big]\big\}\Big) \\ &\ge \frac{(M-\psi_j)}{\alpha} E\Big[\exp\big\{-\alpha T(\mathbf{k}, S_j^c)/(M-\psi_j)\big\}\Big] \end{aligned}$$

as required. □

3. Index policies based on an initial allocation of projects to machines. We use the machinery developed in Section 2 to analyze a simple class of index-based policies in which the collection of available bandits is divided at time zero into M sub-collections, with sub-collection m to be processed exclusively on machine m, $1 \le m \le M$. If this initial division is done appropriately, and if each machine operates an optimal policy (i.e., a Gittins index policy) for its own sub-collection then the total reward earned by this approach comes within an $O(\alpha)$ quantity of $R^{OPT}(\mathbf{k})$ for each initial state $\mathbf{k}$.

More formally, suppose that $M > 1$ and that the collection of B bandits is partitioned into M subsets β_m, $1 \le m \le M$. The bandits in β_m are to be processed (only) on machine m, $1 \le m \le M$. Each machine processes the bandits allocated to it according to a Gittins index policy, as described in Section 2. In the numbering of job types we have adopted, i is preferred to j on each machine if and only if $i > j$. We use the notation $\pi_G(\boldsymbol{\beta})$ for any such policy. Since Gittins index policies are optimal for the single machine problem ($M = 1$), it is transparent that $\pi_G(\boldsymbol{\beta})$ maximises the total reward available for a given partition $\boldsymbol{\beta}$. What is much less clear is that there exist partitions $\boldsymbol{\beta}$ for which the performance of $\pi_G(\boldsymbol{\beta})$ is very close to optimal for the parallel machine problem ($M > 1$) in the sense of the following result.

THEOREM 2.

$$R^{OPT}(\mathbf{k}) - \sup_{\boldsymbol{\beta}} R^{\pi_G(\boldsymbol{\beta})}(\mathbf{k}) \le O(\alpha) \tag{19}$$

where the supremum in (19) *is over all partitions and the* $O(\alpha)$ *bound is uniform in* $\mathbf{k}$.

We develop explicitly a partition $\tilde{\boldsymbol{\beta}}$ for which the properties required for Theorem 2 hold as follows: allocate one bandit to each of machines 1 through $M-1$ in increasing numerical order. In this initial allocation, the type Q

bandits are assigned first, then (if $n_Q < M-1$) the type $Q-1$ bandits, then (if $n_Q + n_{Q-1} < M-1$) type $Q-2$ and so on. Once these $M-1$ single bandit assignments have been made, all remaining bandits are assigned to machine M. Note that in the case $M=1$, $\pi_G(\tilde{\boldsymbol{\beta}})$ is the single machine Gittins index policy, and hence is optimal. The key to the proposed policy $\pi_G(\tilde{\boldsymbol{\beta}})$ is, loosely speaking, that it guarantees sufficient concentration on high index options. Sufficient guarantees are actually rather more difficult to secure in relation to, for example, a conventional Gittins index policy operating in the parallel machine environment, as discussed in Section 4. The reader should also note important similarities between the structure of the partition $\tilde{\boldsymbol{\beta}}$ and that of the single machine relaxation described in the proof of Lemma 2. We require the following result.

LEMMA 3.

$$R^{OPT}(\mathbf{k}) - R^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) \le O(\alpha), \tag{20}$$

and the $O(\alpha)$ bound is uniform in $\mathbf{k}$.

PROOF. We first consider the machines $1, 2, \ldots, M-1$ to which a single bandit is allocated under $\tilde{\boldsymbol{\beta}}$. Let $q(m)$ denote the bandit type allocated to machine m, $1 \le m \le M-1$, by $\tilde{\boldsymbol{\beta}}$. By construction we have

$$n_Q + n_{Q-1} + \cdots + n_{q(m)+1} < m \le n_Q + \cdots + n_{q(m)},$$

where we adopt the convention that $n_{Q+1} = 0$. We denote the bandit on machine m by $b(m)$ and call its initial state $k_{b(m)}$, $1 \le m \le M-1$. From Lemma 1, the contribution to the reward $R^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k})$ from the processing on machine m may be written

$$R_m^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) = G_{|E|}\left(\frac{1}{\alpha}\right) - \sum_{j=1}^{|E|-1} (G_{j+1} - G_j) A^{\pi_G}\left\{S_j, k_{b(m)}\right\}, \tag{21}$$

$$1 \le m \le M-1.$$

In (21) we use π_G to denote a single machine Gittins index policy. Now, since $b(m)$ is of type $q(m)$, the processing on machine m, $1 \le m \le M-1$, satisfies

$$x_i^{\pi_G}\{k_{b(m)}\} = 0, \ i \in S_j, \ j \le g_{q(m)} - 1$$

and hence

$$A^{\pi_G}\{S_j, k_{b(m)}\} = 0, \ j \le g_{q(m)} - 1. \tag{22}$$

Now consider the quantities

$$A^{\pi_G}\left\{S_j, k_{b(m)}\right\}, \quad j \ge g_{q(m)}, \ 1 \le m \le M-1.$$

For machine m, the first contribution to $A^{\pi_G}\left\{S_j, k_{b(m)}\right\}$ will occur from the processing on machine m when a job type in S_j is chosen for the first time. This will occur at time $T\{k_{b(m)}, S_j^c\}$. Following this epoch, excursions to S_j^c

will alternate with allocations to S_j. Hence, using the notation in (15), it is straightforward to see that

$$A^{\pi_G}\{S_j, k_{b(m)}\} = \sum_{i \in S_j \cap E_{q(m)}} E_{\pi_G}\left(\sum_{l=1}^{\nu_i} e^{-\alpha\tau_{i,l}} \int_0^{T_{i,l}^{S_j^c}} e^{-\alpha t} dt | \mathbf{k}\right)$$

$$= \frac{1}{\alpha} E\Big(\exp\big[-\alpha T\{k_{b(m)}, S_j^c\}\big]\Big), \quad j \geq g_{q(m)}, \tag{23}$$

$$1 \leq m \leq M-1.$$

Substituting from (22) and (23) into (21) we conclude that

$$R_m^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) = G_{|E|}\left(\frac{1}{\alpha}\right)$$

$$-\frac{1}{\alpha} \sum_{j=g_{q(m)}}^{|E|-1} (G_{j+1} - G_j)\, E\Big(\exp\big[-\alpha T\{k_{b(m)}, S_j^c\}\big]\Big), \tag{24}$$

$$1 \leq m \leq M-1.$$

Now consider machine M. In what follows we shall use $q(M)$ to denote the bandit of largest type allocated to M, $b(M)$ the collection of bandits on machine M and $\mathbf{k}_{b(M)}$ the associated initial state. By the structure of the single machine index policy π_G, no job type $j \leq g_{q(M)} - 1$ will ever be processed by machine M under policy $\pi_G(\tilde{\boldsymbol{\beta}})$ since a job type associated with a bandit of type $q(M)$ will always be preferred. Hence, as in (22) we have

$$A^{\pi_G}\{S_j, \mathbf{k}_{b(M)}\} = 0, \quad j \leq g_{q(M)} - 1.$$

Following a calculation similar to that which yielded (23), we conclude from Lemma 1 that

$$R_M^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) = G_{|E|}\left(\frac{1}{\alpha}\right)$$

$$-\frac{1}{\alpha} \sum_{j=g_{q(M)}}^{|E|-1} (G_{j+1} - G_j)\, E\Big(\exp\big[-\alpha T\{\mathbf{k}_{b(M)}, S_j^c\}\big]\Big). \tag{25}$$

From (24) and (25) we conclude that

$$
\begin{aligned}
R^{\pi_G(\tilde{\boldsymbol{\beta}})} &= \sum_{m=1}^{M} R_m^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) \\
&= G_{|E|}\left(\frac{M}{\alpha}\right) - \frac{1}{\alpha}\sum_{m=1}^{M}\sum_{j=g_{q(m)}}^{|E|-1}\left(G_{j+1}-G_j\right) \\
&\qquad \times E\Big(\exp\Big[-\alpha T\big\{k_{b(m)}, S_j^c\big\}\Big]\Big) \qquad (26)\\
&= G_{|E|}\left(\frac{M}{\alpha}\right) - \frac{1}{\alpha}\sum_{j=g_{q(M)}}^{|E|-1}\left(G_{j+1}-G_j\right) \\
&\qquad \times \sum_{m=\psi_j+1}^{M} E\Big(\exp\Big[-\alpha T\big\{k_{b(m)}, S_j^c\big\}\Big]\Big).
\end{aligned}
$$

Note that in (26) and in what follows we have written $k_{b(M)}$ for $\mathbf{k}_{b(M)}$ for ease of notation. We now utilize Lemma 2 within (10) to deduce that

$$
\begin{aligned}
R^{OPT}(\mathbf{k}) &\le G_{|E|}\left(\frac{M}{\alpha}\right) - \frac{1}{\alpha}\sum_{j=g_{q(M)}}^{|E|-1}\left(G_{j+1}-G_j\right)\left(M-\psi_j\right) \\
&\qquad \times E\Big[\exp\Big\{-\alpha T\big(\mathbf{k}, S_j^c\big)/\left(M-\psi_j\right)\Big\}\Big]. \qquad (27)
\end{aligned}
$$

Combining (26) and (27) we conclude that

$$
\begin{aligned}
&R^{OPT}(\mathbf{k}) - R^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) \\
&\le \frac{1}{\alpha}\sum_{j=g_{q(M)}}^{|E|-1}\left(G_{j+1}-G_j\right) \\
&\qquad \times \Bigg[\Bigg\{\sum_{m=\psi_j+1}^{M} E\Big(\exp\big[-\alpha T\{k_{b(m)}, S_j^c\}\big]\Big)\Bigg\} \qquad (28)\\
&\qquad\quad - \left(M-\psi_j\right) E\Big[\exp\Big\{-\alpha T\big(\mathbf{k}, S_j^c\big)/\left(M-\psi_j\right)\Big\}\Big]\Bigg].
\end{aligned}
$$

However, by the definitions of the quantities involved and the construction of the partition $\tilde{\boldsymbol{\beta}}$ we have that

$$
\sum_{m=\psi_j+1}^{M} T\big\{k_{b(m)}, S_j^c\big\} = T\big(\mathbf{k}, S_j^c\big), \quad g_{q(M)} \le j \le |E|-1, \qquad (29)
$$

where the summands in the l.h.s. of (29) are independent. See equation (14). Exploiting the existence of all positive moments of the random variables concerned, we can easily deduce from (28) and (29) that

$$
\begin{aligned}
&R^{OPT}(\mathbf{k}) - R^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) \\
&\qquad \le \frac{\alpha}{2}\sum_{j=g_{q(M)}}^{|E|-1}\left(G_{j+1}-G_j\right)\sum_{m=\psi_j+1}^{M} E\left(\Big[T\{k_{b(m)}, S_j^c\}\Big]^2\right), \qquad (30)
\end{aligned}
$$

where the second moments in (30) are guaranteed to exist. We noted in the preamble to (4) above that $T_i^{S^c} \leq T_{ii}$ a.s. when $i \in S$. It is a simple matter to utilize that to bound the r.h.s. of (30) and deduce that

$$R^{OPT}(\mathbf{k}) - R^{\pi_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) \leq \frac{\alpha}{2} \left\{ \sum_{j=g_{q(M)}}^{|E|-1} (G_{j+1} - G_j) \right\} \times \left\{ \sum_{q=1}^{Q} n_q \sigma_q^2 + \left(\sum_{q=1}^{Q} n_q \mu_q \right)^2 \right\}, \tag{31}$$

where

$$\mu_q \equiv \max_{i \in E_q} E(T_{ii}), \quad 1 \leq q \leq Q$$

and

$$\sigma_q^2 \equiv \max_{i \in E_q} \operatorname{var}(T_{ii}), \quad 1 \leq q \leq Q.$$

Note that the r.h.s. of (31) does not depend upon initial state $\mathbf{k}$. We conclude the proof by remarking that in consideration of the limit $\alpha \to 0$, the matrix $\mathbf{A}$ defined in (4) is $O(1)$. The operation of algorithm $AG(\mathbf{A}, \mathbf{r})$ guarantees that the indices G_i, $i \in E$, share this property. The result now follows from (31). □

Theorem 2 is now an immediate consequence of Lemma 3.

It is possible to strengthen the above analysis in the case $n_Q \geq M$ in which there are at least as many type Q bandits as machines. This includes the special case in which all the bandits are of a single type, when we have $B = n_Q = n_1 \geq M$. See the comments following the proof of Lemma 4. With this additional condition, we study all partitions $\hat{\boldsymbol{\beta}}$ which are such that the collection of bandits allocated to each machine contains at least one of type Q. The remaining bandits can be distributed between the machines in any fashion. The next result states that for any such $\hat{\boldsymbol{\beta}}$, the reward earned by $\pi_G(\hat{\boldsymbol{\beta}})$ comes within $O(\alpha)$ of optimality.

LEMMA 4. *If $n_Q \geq M$ then*

$$R^{OPT}(\mathbf{k}) - R^{\pi_G(\hat{\boldsymbol{\beta}})}(\mathbf{k}) \leq O(\alpha) \tag{32}$$

for any qualifying $\hat{\boldsymbol{\beta}}$, where the $O(\alpha)$ bound is uniform in $\mathbf{k}$.

SUMMARY PROOF. Retaining the notation of the earlier proof we have that

$$n_Q \geq M \Rightarrow q(M) = Q \quad \text{and} \quad \psi_j = 0, \ j \geq g_{q(M)}. \tag{33}$$

Hence (26) becomes, for any qualifying $\hat{\boldsymbol{\beta}}$,

$$R^{\pi_G(\hat{\boldsymbol{\beta}})}(\mathbf{k}) = G_{|E|}\left(\frac{M}{\alpha}\right) - \frac{1}{\alpha}\sum_{j=g_Q}^{|E|-1}\left(G_{j+1}-G_j\right)\sum_{m=1}^{M}E\Big(\exp\Big[-\alpha T\{\mathbf{k}_{b(m)}, S_j^c\}\Big]\Big) \tag{34}$$

where $b(m)$ now denotes the collection of bandits allocated by $\hat{\boldsymbol{\beta}}$ to machine m.

In addition to (33) we can also deduce that

$$\psi_j \geq M, \quad j \leq g_{q(M)} - 1 \tag{35}$$

and so from Corollary 1 and Lemma 2 we infer that

$$R^{OPT}(\mathbf{k}) \leq G_{|E|}\left(\frac{M}{\alpha}\right) - \frac{M}{\alpha}\sum_{j=g_Q}^{|E|-1}\left(G_{j+1}-G_j\right)E\Big(\exp\big[-\alpha T\{\mathbf{k}, S_j^c\}/M\big]\Big) \tag{36}$$

From (34) and (36) the remainder of the proof follows closely the calculations in the proof of Lemma 3 [from (26) and (27)]. □

Comment. As mentioned above, the stronger analysis of Lemma 4 is in particular available for problems in which all bandits are of a single type. Models of this kind are important, inter alia, as (finite state) approximations to parallel server versions of many of the classical multi-armed bandit problems. For example, in the Bernoulli reward process of Gittins (1989) the state (i, j) of a bandit corresponds to the parameters of a (posterior) beta distribution. A finite state approximation in which all bandits are of a single type (but with possibly different initial states) and with common state space $\{0, 1, 2, \ldots, n\}^2$ for some suitably chosen large integer n will be available, for example, when all the beta priors have parameters drawn from the integers. For such finite state approximations which meet the requirement of irreducibility, Lemma 4 holds.

4. Gittins index policies implemented in the parallel machine environment. In this section, the policy of interest will be the Gittins index policy implemented in the parallel machine environment, denoted π_G. Under any such policy, M bandits whose associated Gittins indices are maximal are chosen at each decision epoch. To be specific, ties are broken such that the priorities $|E| \to |E|-1 \to \cdots \to 2 \to 1$ are respected. This is as in earlier sections. We shall strengthen the result of Glazebrook and Garbe (1999) for those problems which satisfy the hypothesis of Lemma 4. As mentioned above, this includes the important case in which all bandits are of a single type.

THEOREM 3. *If* $n_Q \geq M$ *then*

$$R^{OPT}(\mathbf{k}) - R^{\pi_G}(\mathbf{k}) \leq O(\alpha) \tag{37}$$

where the $O(\alpha)$ *bound is uniform in* $\mathbf{k}$.

PROOF. We shall utilize the upper bound for $R^{OPT}(\mathbf{k})$ in (36). Further, the condition $n_Q \geq M$ guarantees that no job type j with $j \leq g_Q - 1$ will be chosen by π_G for processing. Hence it follows that

$$A^{\pi_G}(S_j, \mathbf{k}) = 0, \quad j \leq g_Q - 1,$$

and so from Lemma 1,

$$R^{\pi_G}(\mathbf{k}) = G_{|E|}\left(\frac{M}{\alpha}\right) - \sum_{j=g_Q}^{|E|-1} \left(G_{j+1} - G_j\right) A^{\pi_G}(S_j, \mathbf{k}). \tag{38}$$

Now fix j in the range $g_Q \leq j \leq |E| - 1$. Policy π_G implements the priority $S_j^c \to S_j$. From initial state $\mathbf{k}$, use $T(\pi_G)$ to denote the first decision epoch at which π_G chooses an S_j-job (i.e., a job type in S_j) for processing. It must be true that at $T(\pi_G)$, all of the $B - M$ bandits *not* scheduled for processing must have current state in S_j. It follows trivially that from $T(\pi_G)$ onwards all S_j^c-jobs in the system will be scheduled for processing by π_G. It will simplify the argument if we assume (without loss of generality) that from $T(\pi_G)$, scheduling is on a "minimal switching" basis—namely, that a bandit is only switched from processing on a particular machine at epochs at which π_G takes it out of processing altogether.

With the above in place we write $T_m(\pi_G)$ for the first time at which machine m processes an S_j-job under π_G. Hence

$$T(\pi_G) = \min_{1 \leq m \leq M} T_m(\pi_G).$$

It is further clear that

$$\sum_{m=1}^{M} T_m(\pi_G) = T(\mathbf{k}, S_j^c), \tag{39}$$

since both sides of (39) are the total processing required for all bandits to escape S_j^c for the first time.

Because of the "minimal switching" assumption, each interval $[T_m(\pi_G), \infty)$ may be expressed as a disjoint union of intervals of the form $[\tau_i, \tau_i + T_i^{S_j^c})$ for some $i \in S_j$ where τ_i is an epoch at which job type i is scheduled and $\tau_i + T_i^{S_j}$ is the first subsequent epoch at which the corresponding bandit escapes S_j^c.

It then follows simply, utilising the notation of (15), that

$$
\begin{aligned}
A^{\pi_G}(S_j, \mathbf{k}) &= \sum_{i \in S_j} A_i^{S_j} x_i^{\pi_G}(\mathbf{k}) \\
&= \sum_{i \in S_j} E_{\pi_G} \left(\sum_{l=1}^{\nu_i} e^{-\alpha \tau_{i,l}} \int_0^{T_{i,l}^{S_j^c}} e^{-\alpha t} dt | \mathbf{k} \right) \qquad (40) \\
&= \frac{1}{\alpha} \sum_{m=1}^{M} E\Big(\exp\big[- \{\alpha T_m(\pi_G)\}\big] \Big).
\end{aligned}
$$

We then have from (36), (38) and (40) that

$$
\begin{aligned}
&R^{OPT}(\mathbf{k}) - R^{\pi_G}(\mathbf{k}) \\
&\quad \le \frac{1}{\alpha} \sum_{j=g_Q}^{|E|-1} (G_{j+1} - G_j) \left\{ \sum_{m=1}^{M} E\Big(\exp\big[- \alpha T_m(\pi_G) \big] \Big) \right. \qquad (41) \\
&\qquad \left. - E\Big(M \exp\big[- \alpha T\{\mathbf{k}, S_j^c\}/M \big] \Big) \right\}
\end{aligned}
$$

We now invoke (39) and (41) together with calculations like those which conclude the proof of Lemma 3 to prove the result. □

5. Asymptotic optimality and load balancing. We shall now consider the problems described above in a limit as the discount rate $\alpha \to 0$. It will assist clarity if we include α in the notation at key points. Hence, in this section we shall write $\pi_G(\alpha)$, $\pi_G(\tilde{\boldsymbol{\beta}}, \alpha)$, $\pi_G(\hat{\boldsymbol{\beta}}, \alpha)$, $R^{OPT}(\mathbf{k}, \alpha)$, $R^{\pi}(\mathbf{k}, \alpha)$, $\mathbf{A}(\alpha)$ and $G_j(\alpha)$. First note from (4) that

$$
\lim_{\alpha \to 0} A_i^S(\alpha) = E\left(T_i^{S^c}\right) \equiv \bar{A}_i^S, \ i \in S.
$$

It is easy to deduce that the limits

$$
\lim_{\alpha \to 0} G_i(\alpha) \equiv \bar{G}_i, \ i \in E, \qquad (42)
$$

all exist and are finite. The limiting indices $\bar{G}_i$, $i \in E$, may be computed from the adaptive greedy algorithm $AG(\bar{\mathbf{A}}, \mathbf{r})$ whose inputs are the matrix $\bar{\mathbf{A}} \equiv (\bar{A}_i^S)_{i \in E, S \subseteq E}$ and reward vector $\mathbf{r}$.

Our concern will be to analyze the limiting forms of the policies discussed in Sections 3 and 4, all of which will be identified via the $\bar{}$ notation. For example, when $n_Q \ge M$ (with n_Q now defined with respect to the limiting indices $\bar{G}_i$, $i \in E$), $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$ is the limiting form of the policy analyzed in Lemma 4—that is, $\hat{\boldsymbol{\beta}}$ is such that each machine has (at least) one type Q bandit to process and $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$ operates a Gittins index policy on the individual machines. Similarly, $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$ is the limiting form of the policies discussed in Lemma 3. Following Section 4, we shall use $\bar{\pi}_G$ for an index policy (based on the $\bar{G}_i$, $i \in E$) implemented in the parallel machine environment.

Suppose now that the limiting indices $\bar{G}_i,\ i \in E$, are all distinct, namely that

$$\bar{G}_{|E|} > \bar{G}_{|E|-1} > \cdots > \bar{G}_2 > \bar{G}_1. \tag{43}$$

In this case it is trivial to establish the existence of $\alpha^* > 0$ s.t. the policies $\pi_G(\tilde{\boldsymbol{\beta}}, \alpha)$, $\pi_G(\hat{\boldsymbol{\beta}}, \alpha)$ and $\pi_G(\alpha)$ are identical for $\alpha \in (0, \alpha^*]$ in the sense that they may only differ on how they choose from among bandits of the same state and type. For $\alpha \in (0, \alpha^*]$ these policies will correspond to the limiting forms $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$, $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$ and $\bar{\pi}_G$. It follows trivially from Lemmas 3 and 4 and Theorem 3 that under appropriate conditions $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$, $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$ and $\bar{\pi}_G$ are 1-optimal in the sense of Veinott (1966). This states in the case of $\bar{\pi}_G$ that, for all $\mathbf{k}$,

$$\lim_{\alpha \to 0} \left\{ R^{OPT}(\mathbf{k}, \alpha) - R^{\bar{\pi}_G}(\mathbf{k}, \alpha) \right\} = 0$$

with corresponding statements for $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$ and $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$. In general (43) does not hold. However, it is a piece of straightforward analysis to establish that there must always exist a sequence $\{\alpha_n, n \in \mathbb{N}\}$ and a numbering of the members of E such that:

(i) $\lim_{n \to \infty} \alpha_n = 0$;

(ii) for all $n \in \mathbb{N}$,

$$G_{|E|}(\alpha_n) \geq G_{|E|-1}(\alpha_n) \geq \cdots \geq G_2(\alpha_n) \geq G_1(\alpha_n);$$

(iii)

$$\bar{G}_{|E|} \geq \bar{G}_{|E|-1} \geq \cdots \geq \bar{G}_2 \geq \bar{G}_1.$$

The limit policies discussed below are all assumed to make choices in terms of the indices $\bar{G}_i,\ i \in E$, via the ordering $|E| \to |E| - 1 \to \cdots \to 1$ in (ii), (iii) above. Theorem 4 is an easy consequence of the results in Sections 3 and 4, (i)–(iii) above and classical results of Blackwell (1962), Veinott (1966) and Denardo and Miller (1968). We sketch the main ideas in the proof.

THEOREM 4 (Asymptotic optimality of limit policies).

(a) *Policies $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$ are 1-optimal and average-overtaking optimal.*

(b) *If $n_Q \geq M$, policies $\bar{\pi}_G(\hat{\boldsymbol{\beta}})$ and $\bar{\pi}_G$ are 1-optimal and average-overtaking optimal.*

SKETCH PROOF. For definiteness, we discuss (a); the proof of (b) is along similar lines. Suppose that we have (i)–(iii) above. In that event

$$\bar{\pi}(\tilde{\boldsymbol{\beta}}, \alpha_n) \equiv \bar{\pi}_G(\tilde{\boldsymbol{\beta}}), n \in \mathbb{N}. \tag{44}$$

From Lemma 3 and (44) we plainly have, for all $\mathbf{k}$,

$$\lim_{n \to \infty} \left\{ R^{OPT}(\mathbf{k}, \alpha_n) - R^{\bar{\pi}_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}, \alpha_n) \right\} = 0. \tag{45}$$

However, for discounted MDPs with finite state and action-space, we can write

$$R^{\pi}(\mathbf{k}, \alpha) = \frac{R_1^{\pi}(\mathbf{k})}{\alpha} + R_2^{\pi}(\mathbf{k}) + O(\alpha) \tag{46}$$

for stationary policy π, where R_1^π is the usual average return per unit time for π. See Blackwell (1962). It follows easily from (45) and (46) that for all $\mathbf{k}$

$$R_1^{\bar{\pi}_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) = \sup_\pi R_1^\pi(\mathbf{k}) \tag{47}$$

where the supremum is over all stationary policies and

$$R_2^{\bar{\pi}_G(\tilde{\boldsymbol{\beta}})}(\mathbf{k}) = \sup_\pi R_2^\pi(\mathbf{k}) \tag{48}$$

where this supremum is over all those policies attaining the supremum in (47) for all $\mathbf{k}$. Note that (47) states that $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$ is average-reward optimal. We now conclude the 1-optimality of $\bar{\pi}_G(\tilde{\boldsymbol{\beta}})$ from (46)–(48). That it is average-overtaking optimal follows from (47), (48) and results of Veinott (1966) and Denardo and Miller (1968). This concludes the proof. □

Comment. If we use $R^\pi(\mathbf{k}, t)$ to denote the (undiscounted) reward obtained at time t when policy π is applied from initial state $\mathbf{k}$, then stationary policy $\bar{\pi}$ is average-overtaking optimal if

$$\liminf_{T\to\infty} \frac{1}{T+1}\left[\sum_{t=0}^{T}\sum_{s=0}^{t} E\{R^{\bar{\pi}}(\mathbf{k}, \mathbf{s})\} - \sum_{t=0}^{T}\sum_{s=0}^{t} E\{R^{\pi}(\mathbf{k}, \mathbf{s})\}\right] \geq 0$$

for all policies π and initial states $\mathbf{k}$. As is implied in the proof of Theorem 4, the claim that our policies are average-overtaking optimal implies a claim that they are also average-reward optimal.

In the case of the result in Lemma 4 we pursue one final issue. According to that result, if $n_Q \geq M$ then the reward from $\pi_G(\hat{\boldsymbol{\beta}})$ comes within $O(\alpha)$ of the optimal reward for any partition $\hat{\boldsymbol{\beta}}$ of bandits to machines which guarantees that at least one type Q bandit is allocated to each machine. We explore the question of which partitions satisfying this requirement might perform particularly well by seeking to make the leading (α) term in the bound on $R^{OPT} - R^{\pi_G(\hat{\boldsymbol{\beta}})}$ developed in the proof Lemma 4 as small as possible. As we shall see, a natural problem for balancing the initial load between the machines results.

From (34), (36) and (42) we can deduce that

$$\begin{aligned} &R^{OPT}(\mathbf{k}) - R^{\pi_G(\hat{\boldsymbol{\beta}})}(\mathbf{k}) \\ &\quad\leq \frac{\alpha}{2}\sum_{j=q_Q}^{|E|-1}(\bar{G}_{j+1} - \bar{G}_j)\left\{\sum_{m=1}^{M} E\left(\left[T\{\mathbf{k}_{b(m)}, S_j^c\}\right] - \frac{T(\mathbf{k}, S_j^c)}{M}\right)^2\right\} + O(\alpha^2) \qquad (49) \\ &\quad= \frac{\alpha}{2}\sum_{j=q_Q}^{|E|-1}(\bar{G}_{j+1} - \bar{G}_j)\left(\sum_{m=1}^{M}\left(E\left[T\{\mathbf{k}_{b(m)}, S_j^c\}\right]\right)^2\right. \\ &\qquad\left. + \operatorname{var}\left\{T(\mathbf{k}, S_j^c)\right\} - \frac{E[\{T(\mathbf{k}, S_j^c)\}^2]}{M}\right) + O(\alpha^2). \qquad (50) \end{aligned}$$

To obtain (50) from (49) we utilize the independence of the random variables $T\{\mathbf{k}_{b(m)}, S_j^c\}$, $1 \le m \le M$, and the fact that

$$\sum_{m=1}^{M} T\{\mathbf{k}_{b(m)}, S_j^c\} = T(\mathbf{k}, S_j^c), \quad g_Q \le j \le |E| - 1.$$

However, note in (50) that the random variables $T(\mathbf{k}, S_j^c)$, $g_Q \le j \le |E| - 1$ do not depend stochastically upon the partition $\hat{\boldsymbol{\beta}}$ of the bandits. Hence it is natural to seek a partition of the bandit set which minimizes the quantity

$$\sum_{j=g_Q}^{|E|-1} (\bar{G}_{j+1} - \bar{G}_j) \sum_{m=1}^{M} \left(E\left[T\{\mathbf{k}_{b(m)}, S_j^c\}\right]\right)^2. \tag{51}$$

Consider an initial state $\mathbf{k}$ in which there are η_i jobs of type i in the system. Some allocation $\boldsymbol{\beta}$ sends $\eta_{i,m}$ of these (or, equivalently, the bandits to which they belong) for processing to machine m, $1 \le m \le M$. The minimization of the expression in (51) can then be formulated as the nonlinear integer program

$$\text{minimize} \sum_{j=g_Q}^{|E|-1} (\bar{G}_{j+1} - \bar{G}_j) \sum_{m=1}^{M} \left(\sum_{i \in S_j^c} \eta_{i,m} \mu_{ij} \right)^2 \tag{52}$$

$$\text{such that} \sum_{m=1}^{M} \eta_{i,m} = \eta_i, \quad g_Q \le i \le |E|,$$

$$\sum_{i \in E_Q} \eta_{i,m} \ge 1, \quad 1 \le m \le M, \tag{53}$$

$$\eta_{i,m} \in \mathbb{N}, \quad g_Q \le i \le |E|, \quad 1 \le m \le M.$$

Note that we have written μ_{ij} for $E\{T(i, S_j^c)\}$ in (52) and (53) expresses the requirement that at least one type Q bandit must be allocated to each machine.

The above optimization problem has a natural interpretation as a load balancing problem. There are simple solutions in some special cases. For example, if each η_i is a multiple of M, then it will be optimal to give identical initial allocations to the machines, such that $\eta_{im} = \eta_{im'}$ for all $m \ne m'$ and for all i. Consider also a problem in which

$$\bar{G}_{|E|} = \bar{G}_{|E|-1} = \cdots = \bar{G}_{j+1} > \bar{G}_j = \cdots \bar{G}_{g_Q}. \tag{54}$$

With the condition (54), (52) now becomes

$$\text{minimize} \sum_{m=1}^{M} \left(\sum_{i \in S_j^c} \eta_{i,m} \mu_{ij} \right)^2. \tag{55}$$

When $j = |E| - 1$ in (54) and (55), the optimal initial allocation will share the $|E|$-jobs between machines as equally as possible, while guaranteeing that each machine has at least one bandit of type Q allocated to it. For general j, an allocation minimizing the objective in (55) will attempt to equalise the M quantities $\sum_{i \in S_j^c} \eta_{i,m} \mu_{ij}$ subject to the constraints (53).

REFERENCES

BERTSIMAS, D. and NIÑO-MORA J. (1996). Conservation laws, extended polymatroids and multi-armed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.* **21** 257–306.

BLACKWELL D. (1962). Discrete dynamic programming. *Ann. Math. Stat.* **33** 719–726.

COFFMAN, E. and MITRANI, I. (1980). A characterization of waiting time performance realizable by single server queues. *Oper. Res.* **28** 810–821.

DENARDO, E. V. and MILLER, B. L. (1968). An optimality criterion for discrete dynamic programming with no discounting, *Ann. Math. Stat.* **39** 1220–1227.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics: European Meeting of Statisticians, Budapest, 1972* (J. Gani, K. Sarkadi and I. Vince, eds.) 241–266. North-Holland, Amsterdam.

GLAZEBROOK, K. D. (1976) Stochastic scheduling. Ph.D. thesis, Cambridge Univ.

GLAZEBROOK, K. D. and GARBE, R. (1999). Almost optimal policies for stochastic systems which almost satisfy conservation laws. *Ann. Oper. Res.* **92** 19–43.

SHANTHIKUMAR, J. G. and YAO, D. D. (1992). Multi-class queueing systems: polymatroidal structure and optimal scheduling control. *Oper. Res.* **40** 293–299.

VEINOTT, JR, A. F. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Stat.* **37** 1284–1294.

WEBER, R. R. (1982) Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime. *J. Appl. Probab.* **19** 167–182.

WEBER, R. R., VARAIYA, P. and WALRAND, J. (1986) Scheduling jobs with stochastically ordered processing times on parallel machines to minimize expected flowtime. *J. Appl. Probab.* **23** 841–847.

WEISS, G. (1990) Approximation results in parallel machines stochastic scheduling. *Ann. Oper. Res. Special Volume on Production Planning and Scheduling* (M. Queyranne, ed.) **26** 195–242.

WEISS, G. (1992) Turnpike optimality of Smith's rule in parallel machines stochastic scheduling. *Math. Oper. Res.* **17** 255–270.

WEISS, G. (1995) On almost optimal priority rules for preemptive scheduling of stochastic jobs on parallel machines. *Adv. Appl. Probab.* **27** 821–839.

DEPARTMENT OF STATISTICS
NEWCASTLE UNIVERSITY
NEWCASTLE UPON TYNE, NE1 7RU
UNITED KINGDOM
E-MAIL: kevin.glazebrook@ncl.ac.uk
d.j.wilkinson@ncl.ac.uk