The Annals of Applied Probability
1996, Vol. 6, No. 3, 1024–1034

FINITE STATE MULTI-ARMED BANDIT PROBLEMS: SENSITIVE-DISCOUNT, AVERAGE-REWARD AND AVERAGE-OVERTAKING OPTIMALITY

BY MICHAEL N. KATEHAKIS AND URIEL G. ROTHBLUM

Rutgers University and Technion–Israel Institute of Technology

We express Gittins indices for multi-armed bandit problems as Laurent expansions around discount factor 1. The coefficients of these expansions are then used to characterize stationary optimal policies when the optimality criteria are sensitive-discount optimality (otherwise known as Blackwell optimality), average-reward optimality and average-overtaking optimality. We also obtain bounds and derive optimality conditions for policies of a type that continue playing the same bandit as long as the state of that bandit remains in prescribed sets.

1. Introduction. Multi-armed bandit problems have traditionally been studied under a total-discounted-reward optimality criterion with a fixed interest rate. In the current paper, discrete time, finite state multi-armed bandit problems are studied under alternative optimality criteria, namely, sensitive-discount optimality (Blackwell optimality), average-reward optimality and average-overtaking optimality. Related work for specific instances of the problem was done by Kelly [(1981), Bayes treatment of Bernoulli bandits with unknown success probabilities] and by Lai and Ying [(1988), average optimality for a particular queuing model].

Sensitive-discount optimality concerns simultaneous maximization of total-discounted-reward under all sufficiently small positive interest rates. We show that the Gittins indices have representations as (computable) Laurent series in the (sufficiently small positive) interest rate; hence, a generalized index rule based on lexicographic maximization of the sequence of coefficients of the Laurent expansions can be used to obtain stationary index policies which are sensitive-discount optimal. The lexicographic comparisons require the computation of infinitely many coefficients. However, in the spirit of results of Miller and Veinott (1969) for Markov decision chains, we prove that the lexicographic comparisons can be truncated to rely only on a finite (prescribed) number of terms, yielding a finite algorithm for computing stationary index policies which are sensitive-discount optimal. As computation is applied to the projects independently, our results preserve the classic decomposition structure of the optimal policies for bandit problems with fixed interest rate.

Received February 1994; revised January 1996.

AMS 1991 *subject classifications*. 90C47, 90C31, 90C39, 60G40.

Key words and phrases. Bandit problems, optimality criteria, Markov decision chains, Gittins index, Laurent expansions.

We consider two additional optimality criteria, namely, average-reward optimality and average-overtaking optimality (see Sections 2 and 3 for formal definitions). Known results about Markov decision chains show that every stationary sensitive-discount optimal policy is both average-reward optimal and average-overtaking optimal. However, we obtain algorithms for computing stationary generalized index policies that are, respectively, average-reward optimal and average-overtaking optimal which are more efficient than the one that we developed for finding stationary generalized index policies which are sensitive-discount optimal. These algorithms use, respectively, only two or three coefficients of the corresponding Laurent series of the Gittins indices.

In constructing and implementing policies for multi-armed bandit problems, it is reasonable to activate selected projects for more than a single period. *Holding policies* are procedures that use first exit times of particular sets of states to determine the time for reevaluating the selection of projects. We also construct optimal holding policies for each of the three criteria we consider. At decision epochs, these policies maximize lexicographically coefficients of the Laurent expansions of the indices, but one fewer term is needed than for optimal stationary policies; in particular, average-reward optimality requires a single coefficient and average-overtaking optimality requires two.

Our approach extends to problems with infinitely many projects and states. However, we do not consider such extensions here because additional technical requirements are needed and the resulting algorithms do not reduce to finite calculation.

Results about Markov decision chains and multi-armed bandit problems are reviewed in Sections 2 and 3, respectively. Laurent expansions of the Gittins indices are developed and are used in Section 4 to construct optimal index policies for each of the three criteria we consider. Finally, optimal holding policies are constructed in Section 5.

2. Optimality criteria for Markov decision chains. Consider a *Markov decision chain* (MDC) with finite space S and finite action space A. For $s, u \in S$ and $a \in A$, let $R_a(s)$ be the *one-step reward* received when action a is taken in state s and let $P_a(u|s)$ be the *transition probability* from state s into state u under action a. Policies are functions which map history paths into actions. Depending on the initial state s, a policy π determines a reward stream denoted $\{R^t_\pi(s)\}_{t=0,1,\dots}$ and for $0 < \alpha < 1$ the *expected α-discounted reward associated* with π is then given by $W_\pi(s, \alpha) \equiv \sum_{t=0}^\infty \alpha^t \mathrm{E}[R^t_\pi(s)]$. The supremum of these quantities over all policies π is denoted $V(s, \alpha)$. Throughout we use the index α (the *discount factor*) interchangeably with the index $\rho \equiv (1 - \alpha)/\alpha$ (the *interest rate*); for example, we write $V(s, \rho)$ for $V(s, \alpha)$.

A policy is called *α-discount optimal* if $W_\pi(s, \alpha) = V(s, \alpha)$ for each $s \in S$. A policy π is called *stationary* if the action associated to each history path depends only on its last state, say s, and in this case we denote that action by

$\pi(s)$. Blackwell (1962) showed that for each $0 < \alpha < 1$, there exists a stationary policy which is α-discount optimal.

Miller and Veinott (1969) showed that for some $\rho^* > 0$ there are Laurent expansions

$$W_\pi(s,\rho) = \sum_{n=-1}^{\infty} \rho^n w_\pi^{(n)}(s) \text{ for each stationary policy } \pi, \quad s \in S \text{ and } 0 < \rho < \rho^*, \tag{2.1}$$

and

$$V(s,\rho) = \sum_{n=-1}^{\infty} \rho^n v^{(n)}(s) \quad \text{for each } s \in S \text{ and } 0 < \rho < \rho^*, \tag{2.2}$$

and that the coefficients $w_\pi^{(n)}(s)$ and $v^{(n)}(s)$ of these expansions can each be computed with finitely many arithmetic operations. It turns out that useful conditions for a stationary policy π are to match the first $k + 2$ coefficients of the expansion (2.1) with that of (2.2), that is, to satisfy

$$\left(w_{\pi^*}^{(-1)}(s), w_{\pi^*}^{(0)}(s), \dots, w_{\pi^*}^{(k)}(s)\right) = \left(v^{(-1)}(s), v^{(0)}(s), \dots, v^{(k)}(s)\right) \quad \text{for each } s \in S. \tag{2.3}$$

A policy π is called *sensitive-discount optimal* or *Blackwell optimal* if for some $0 < \alpha^* < 1$ it is α-discount optimal for all $0 < \alpha < \alpha^*$. Veinott (1969) showed that a stationary policy is sensitive-discount optimal if and only if satisfies (2.3) with $k = |S|$. Furthermore, he obtained an algorithm, requiring finite computation, that identifies a stationary policy that satisfies this condition. Thus, Veinott obtained a constructive proof for the existence of stationary sensitive-discount-optimal policies, a result proved earlier (nonconstructively) in Blackwell (1962).

A policy π^* is called *average-reward optimal* if

$$\liminf_{t\to\infty} \frac{1}{t+1}\left\{\sum_{k=0}^{t} \mathrm{E}\left[R_{\pi^*}^k(s)\right] - \sum_{k=0}^{t} \mathrm{E}\left[R_\pi^k(s)\right]\right\} \ge 0 \quad \text{for each policy } \pi \text{ and each state } s \in S. \tag{2.4}$$

and *average-overtaking-optimal* if

$$\liminf_{t\to\infty} \frac{1}{T+1}\left\{\sum_{t=0}^{T}\sum_{k=0}^{t} \mathrm{E}\left[R_{\pi^*}^k(s)\right] - \sum_{t=0}^{T}\sum_{k=0}^{t} \mathrm{E}\left[R_\pi^k(s)\right]\right\} \ge 0 \quad \text{for each policy } \pi \text{ and each state } s \in S. \tag{2.5}$$

It turns out that a stationary policy π^* is average-reward optimal if and only if it satisfies (2.3) with $k = -1$ and average-overtaking optimal if and only if it satisfies (2.3) with $k = 0$; see Veinott (1966, 1974) and Denardo and Miller (1968). In particular, for stationary policies sensitive-discount optimality implies average-overtaking optimality, which implies average-reward optimality.

Additional optimality criteria are obtained from (2.5) by replacing the double summation with any finite number of consecutive summations. In particular, if $(k + 2)$-order summations are used, the corresponding optimal stationary policies are characterized by (2.3); see Veinott (1974) and Rothblum and Veinott (1992). In all cases results continue to hold when optimization takes place within the class of randomized policies, and for stationary randomized policies the associated expected discounted reward have Laurent expansions as in (2.1).

3. Preliminaries in multi-armed bandit problems.

We next consider a *multi-armed bandit problems* (MABP) with a finite set of projects N, where each project i has a finite state space S_i. For $i \in N$ and $x, y \in S_i$, let $r_i(x)$ be the one-step reward received when project i is selected while in state x and let $p_i(x, y)$ be the transition probability of project i from state x into state y when i is active. As usual, this MABP is identified with a MDC having state space $S \equiv \prod_{i \in N} S_i$, and we use the terminology of policies and optimality summarized in Section 2 to that MDC.

Let $J \equiv \bigcup_{i \in N} (\{i\} \times S_i)$. An *index* for the MABP is a real-valued function μ: $J \to R$. We say that a stationary policy π is *consistent with* index μ if

$$\mu\big(\pi(s), s_{\pi(s)}\big) = \max_{i \in N} \mu(i, s_i) \quad \text{for each } s \in S. \tag{3.1}$$

A stopping time τ on the process $\{Y_i^t\}_{t=0,1,\dots}$ generated when project i is activated indefinitely is called a *stopping time* for project i. For such τ, consider the indices defined by

$$m_\tau(i, x, \alpha) \equiv \frac{\mathrm{E}\big[\sum_{t=0}^{\tau-1} \alpha^t r_i\big(Y_i^t\big) | Y_i^0 = x\big]}{1 - \mathrm{E}\big[\alpha^\tau | Y_i^0 = x\big]} \quad \text{for each } (i, x) \in J. \tag{3.2}$$

The *Gittins index* is obtained by taking the suprema of these quantities over all stopping times τ for project i. Herein, we denote the Gittins index associated with $(i, x) \in J$ by $m(i, x, \alpha)$; the parameter α is included to express the dependence (which we shall explore) on the discount factor α. Gittins and Jones (1974) proved that these suprema are well defined and attained, that each stationary policy which is consistent with the Gittins index is α-discount optimal and that such a stationary policy exists; see Gittins (1989), Ross (1983), Whittle (1982) and references therein.

Glazebrook (1982, 1990) used Gittins indices to bound the performance of stationary policies. We next use his results to bound the performance of stationary index policies via their indices.

Proposition 3.1 (Bounding the performance of stationary index policies). *Let $L, K \geq 0$ and let μ be an index satisfying*

$$-L \leq \mu(i, x) - m(i, x, \alpha) \leq K \quad \textit{for each } (i, x) \in J. \tag{3.3}$$

Then each stationary policy π that is consistent with μ satisfies

$$W_\pi(s, \alpha) \geq V(s, \alpha) - (1 - \alpha)^{-1}(L + K) \quad \textit{for each state } s \in S. \tag{3.4}$$

Proof. Let π be a stationary policy which is consistent with μ. The consistency of π with μ and two applications of (3.3) show that for $s \in S$ and for $i = 1, \dots, N$,

$$(3.5)\quad m\big(\pi(s), s_{\pi(s)}, \alpha\big) + K \geq \mu\big(\pi(s), s_{\pi(s)}\big) \geq \mu(i, s_i) \geq m(i, s_i, \alpha) - L.$$

So, $m(\pi(s), s_{\pi(s)}, \alpha) \geq \max_{i \in N} m(i, s_i, \alpha) - (L + K)$ for each $s \in S$, and the inequalities of (3.4) follow from Glazebrook [(1982), Theorem 2]. □

For related bounds for policies under which selected projects are activated for a number of periods determined by stopping times, see Katehakis and Veinott (1987) and Glazebrook (1991).

Katehakis and Veinott (1987) obtained a representation of the Gittins indices by considering Markov decision chains MDC^{ix} for each pair $(i, x) \in J$; MDC^{ix} has state space S_i and two actions—one which continues to activate the project and the other which instantly restarts the process at state x. A stationary policy δ for MDC^{ix} corresponds to a subset $C(\delta)$ of S_i that contains x and consists of the states at which the policy continues (rather than restarts at x). Also, a stationary policy δ of MDC^{ix} induces a stopping time $\tau(\delta)$ which is the first time the restart option is taken. Let $W_\delta^{ix}(y, \alpha)$ be the expected α-discounted reward associated with δ when y is the initial state and let $V^{ix}(y, \alpha)$ be the corresponding optimal expected α-discounted reward. Katehakis and Veinott [(1987), Proposition 2] showed that, with Δ^{ix} as the set of stationary policies for MDC^{ix}, for $0 < \alpha < 1$ and $(i, x) \in J$,

$$(3.6)\qquad m_{\tau(\delta)}(i, x, \alpha) = W_\delta^{ix}(x, \alpha) \quad \text{for each } \delta \in \Delta^{ix}$$

and

$$(3.7)\qquad m(i, x, \alpha) = \max_{\delta \in \Delta^{ix}} m_{\tau(\delta)}(i, x, \alpha) = V^{ix}(x, \alpha).$$

An alternative representation of Gittins indices was obtained by Whittle (1980) by considering parametric MDC's for each project, depending on a parameter m but not on the states of the projects. Two actions are available in Whittle's construction—one which continues to activate the project and the other which calls for retirement with (the parametric) payoff m. The above construction differs in that retirement is not allowed; rather the option of restarting project i in state x is available.

4. Stationary optimal policies for MABP. The (nonconstructive) arguments of Blackwell (1962) and the α-discount optimality of stationary index policies for each fixed α imply the existence of stationary index policies which are sensitive-discount optimal, hence, average-reward and average-overtaking optimal; see Section 2. In the current section we show how such policies can be computed.

From (3.7) and (2.2) we get the following Laurent expansions of Gittins indices.

THEOREM 4.1 (Laurent expansions of Gittins Indices). *For some $\rho^* > 0$, there are Laurent expansions*

$$(4.1)\quad m(i, x, \rho) = \sum_{n=-1}^{\infty} \rho^n m^{(n)}(i, x) \quad \text{for each } (i, x) \in J \text{ and } 0 < \rho < \rho^*,$$

and the coefficients $m^{(-1)}(i, x), m^{(0)}(i, x), \dots$ of these expansions equal the coefficients of the expansions of the $V^{ix}(x, \rho)$'s.

As in (2.2), each of the coefficients $m^{(-1)}(i, x), m^{(0)}(i, x), \dots$ of (4.1) can be computed with finitely many arithmetic operations. Also, the arguments of Katehakis and Veinott [(1987), Proposition 2] combine with standard renewal arguments to show that, with the Y_i^t's as in (3.1), $m^{-1}(i, x)$ has the representation (pointed out to us by Glazebrook)

$$(4.2)\qquad m^{-1}(i, x) = \sup_{\tau \geq 1} \frac{\mathrm{E}\left[\sum_{t=0}^{\tau-1} r_i(Y_i^t) | Y_i^0 = x\right]}{\mathrm{E}\left[\tau | Y_i^0 = x\right]}.$$

Given two real sequences $a = (a_{-1}, a_0, \dots)$ and $(b = (b_{-1}, b_0, \dots)$, we say that a *dominates* b *lexicographically*, written $a \gg_{\text{lex}} b$, if for some $k \in \{-1, 0, \dots\}$, $a_n = b_n$ for all $-1 \leq n \leq k-1$ and $a_k > b_k$. Also, we write $a \ggeq_{\text{lex}} b$ if either $a \gg_{\text{lex}} b$ or $a = b$. As $\gg_{\text{lex}}$ is a complete order on the set of infinite sequences, every finite set of such sequences, say $a^1, \dots, a^L$, has a *lexicographically maximal* element with respect to $\gg_{\text{lex}}$, which we denote $\operatorname{lex\,max}_{m \in L} a^m$. These definitions and observations extend to finite sequences in the obvious way.

Given two power series $a(\varepsilon) = \sum_{n=-1}^{\infty} a_n \varepsilon^n$ and $b(\varepsilon) = \sum_{n=-1}^{\infty} b_n \varepsilon^n$ which converge absolutely for all sufficiently small positive ε, $(a_{-1}, a_0, \dots) \ggeq_{\text{lex}} (b_{-1}, b_0, \dots)$ if and only if $a(\varepsilon) \geq b(\varepsilon)$ for all sufficiently small positive ε. Furthermore, if $(a_{-1}, a_0, \dots, a_k) = (b_{-1}, b_0, \dots, b_k)$ for some $k = -1, \dots$, then there exists a real number K such that $|a(\varepsilon) - b(\varepsilon)| \leq K\varepsilon^{k+1}$ for all sufficiently small positive ε. Similar conclusions hold for power series with finitely many terms. The above observations and Theorem 4.1 imply that if a stationary policy π satisfies

$$(4.3)\qquad \begin{aligned} &\left(m^{(-1)}(\pi(s), s_{\pi(s)}), m^{(0)}(\pi(s), s_{\pi(s)}), \dots\right) \\ &\quad = \operatorname*{lex\,max}_{i \in N} \left(m^{(-1)}(i, s_i), m^{(0)}(i, s_i), \dots\right) \quad \text{for each } s \in S, \end{aligned}$$

then for sufficiently small positive ρ,

$$(4.4)\qquad m(\pi(s), s_{\pi(s)}, \rho) = \max_{i \in N} m(i, s_i, \rho) \quad \text{for each } s \in S.$$

That is, π is consistent with the Gittins index and is therefore ρ-discount optimal. So, (4.3) is an (attainable) sufficient condition for a stationary policy to be sensitive-discount optimal. This condition is separable and is based on parameters $m^{(-1)}(i, x), m^{(0)}(i, x), \dots$ that are determined independently for each project. Though each of these coefficients is computable with finitely many arithmetic operations, the computation of the complete sequences requires infinite computation. The next result establishes truncated variants of the implication (4.3) ⇒ (4.4).

Laurent expansions of $W_\pi(\cdot)$ for each stationary policy π and for $V(\cdot)$ are given in (2.1) and (2.2), and we shall use the notation $w_\pi^{(0)}(s), w_\pi^{(1)}(s), \dots$ and $v^{(-1)}, v^{(0)}, \dots$ to denote the corresponding coefficients.

THEOREM 4.2. *Let $k = 0, 1, \dots$ and let π be a stationary policy that satisfies*

$$\begin{aligned}\left(m^{(-1)}\left(\pi(s), s_{\pi(s)}\right), \dots, m^{(k)}\left(\pi(s), s_{\pi(s)}\right)\right)& \\ = \operatorname{lex}\max_{i \in N}\left(m^{(-1)}(i, s_i), \dots, m^{(k)}(i, s_i)\right) &\quad \text{for each } s \in S.\end{aligned} \tag{4.5}$$

Then

$$w_\pi^{(n)}(s) = v^{(n)}(s) \quad \text{for each } s \in S \text{ and } n = -1, \dots, k-1. \tag{4.6}$$

PROOF. For each $\rho > 0$ consider the index μ_k^ρ defined by

$$\mu_k^\rho(i, x) \equiv \sum_{n=-1}^{k} \rho^n m^{(n)}(i, x) \quad \text{for each } (i, x) \in J. \tag{4.7}$$

For pairs $(i, x), (j, y) \in J$, $(m_{ix}^{(-1)}, m_{ix}^{(0)}, \dots, m_{ix}^{(k)}) \gg_{\text{lex}} (m_{jy}^{(-1)}, m_{jy}^{(0)}, \dots, m_{jy}^{(k)})$ if and only if $\mu_k^\rho(i, x) > \mu_k^\rho(j, y)$ for all sufficiently small positive ρ. As $\gg_{\text{lex}}$ is a complete order on the finite set $\{(m^{(-1)}(i, x), \dots, m^{(k)}(i, x)): (i, x) \in J\}$, (4.5) implies that π is consistent with each of the indices μ_k^ρ for $0 < \rho < \rho^*$.

Theorem 4.1 and standard arguments about power series show that there exist positive constants $\rho^\#$ and K such that $|m(i, x, \rho) - \mu_k^\rho(i, x)| \le K\rho^{k+1}$ for each $0 < \rho < \rho^\#$ and $(i, x) \in J$. As π is consistent with each of the indices μ_k^ρ for $0 < \rho < \rho^*$, Proposition 3.1 implies that

$$\begin{aligned}0 \le V(s, \rho) - W_\pi(s, \rho) \le 2K\rho^k(1 + \rho)& \\ \text{for each } 0 < \rho < \min\{\rho^*, \rho^\#\} \text{ and } s \in S.&\end{aligned} \tag{4.8}$$

Using the expansions of $W_\pi(s, \rho)$ and $V(s, \rho)$ given in (2.1) and (2.2), (4.8) implies that the first $k - 1$ coefficients of the two expansions coincide; that is, (4.6) has been verified. □

Theorem 4.2 is next combined with the characterizations of optimal stationary policies for MDC's through (2.3) to obtain sufficient conditions for these optimality criterion for MABP's.

THEOREM 4.3 (Sufficient conditions for optimality of stationary policies). *If π is a stationary policy satisfying* (4.5) *with $k = |S| + 1$, $k = 0$ or $k = 1$, then π is, respectively, sensitive-discount, average-reward or average-overtaking optimal.*

For each nonnegative integer k, the construction of a stationary policy that satisfies (4.5) requires the computation of the coefficients $m^{(n)}(i, x)$ for each pair $(i, x) \in J$ and each $n \in \{-1, \dots, k\}$. Each of these coefficients can be computed with finitely many arithmetic operations. Thus, Theorem 4.3 yields a finite algorithm for computing stationary sensitive-discount-optimal policies. Such policies are both average-reward optimal and average-overtaking optimal. Verification of (4.5) with $k = |S|$ may require extensive computation when S is large, but Theorem 4.3 also provides succinct sufficient conditions for a stationary policy to be average-reward optimal or average-overtaking optimal, respectively. On-line implementation of algorithms that apply policies that satisfy (4.5) will compute the corresponding coefficients $m^{(n)}(i, x)$ for pairs (i, x) only as they are encountered.

As is the case for index policies, the computation required for verifying (4.5) considers each of the projects independently. In fact, stationary policies that satisfy conditions (4.5) are index policies with index μ_k^ρ given by (4.7) for some positive (small) ρ. Still, (4.5) has the advantage of avoiding the need to determine an appropriate value of ρ which may be difficult.

One can construct stationary policies that satisfy (4.5) for any specified nonnegative integer k. By Theorem 4.1, such policies are then optimal with respect to the optimality criteria mentioned at the end of Section 2.

5. Holding optimal policies for MABP.

A *holding policy* is determined by a strict *ranking* $\gg$ of J and a *continuation function* $C(\cdot)$, which maps each pair $(i, x) \in J$ into a subset $C(i, x)$ of S_i that contains x. The implementation of the holding policy is then as follows:

STEP 0. Set s^1 to be the initial state of the system and enter Step 1.

STEP k. A project i_k $\gg$-maximizing (j, s_j^k) over $j \in N$ is selected and activated. Project i_k remains active while its state is in $C(i_k, s_j^k)$. Once the state of i_k leaves $C(i_k, s_j^k)$, set s^{k+1} to be the state of the system at that point and enter Step $k + 1$.

We refer to entrances into the evaluation step as *decision epochs*. Periods between consecutive decision epochs are stopping times; thus, holding policies are instances of the stopping policies considered in Katehakis and Veinott (1987) (where more complicated stopping rules are allowed).

THEOREM 5.1 (Bounding the performance of holding policies). *Let* $0 < \alpha < 1$, *let* A *and* B *be positive numbers and let* π *be a holding policy with ranking* $\gg$ *and continuation function* $C(\cdot, \cdot)$. *Suppose that* $m(i, x, \alpha) \geq m(j, y, \alpha) - A$ *for all pairs* $(i, x), (j, y) \in J$ *satisfying* $(i, x) \gg (j, y)$, *and further suppose that for each pair* $(i, x) \in J$ *the stationary policy* δ *for* MDC^{ix} *corresponding to* $C(i, x)$ *satisfies* $W_\delta^{ix}(x, \alpha) \geq V^{ix}(x, \alpha) - B$. *Then* $W_\pi(s, \alpha) \geq V(s, \alpha) - (A + B)$ *for each* $s \in S$.

PROOF. Suppose state s is observed and project i is selected at a particular decision epoch. Let δ be the stationary policy of MDC^{is_i} that corresponds to $C(i, s_i)$. Then the first exit time of $C(i, s_i)$ is the stopping time $\tau(\delta)$ as defined in Section 3. In particular, (3.6) and (3.7) combine with the assumptions about π to show that

$$\begin{aligned} m_{\tau(\delta)}(i, s_i, \alpha) &= W_\delta^{is_i}(s_i, \alpha) \geq V^{is_i}(s_i, \alpha) - B \\ &= m(i, s_i, \alpha) - B \geq \max_{j \in N} m(j, s_j, \alpha) - A - B. \end{aligned} \tag{5.1}$$

The asserted inequalities now follow from Katehakis and Veinott [(1987), Theorem 1], where we already observed that holding policies are included in the set of stopping policies they consider. □

A holding policy need not be stationary because the selected action in a given state may depend on the occupied project and on its state when selected. However, holding policies are, in essence, stationary in a MDC with an extended state space. Consequently, for each holding policy π, $W_\pi(\cdot, \rho)$ has a Laurent expansion as in (2.1); furthermore, the characterizations of the various optimality criteria through (2.3) extend from stationary to holding policies.

For MDC^{ix}, we denote the coefficients of the Laurent expansions of $(W_\delta^{ix})(x, \rho)$ for a stationary policy δ and of $V^{ix}(x, \rho)$, respectively, by $(w_\delta^{ix})^{(n)}(x)$ and $(v^{ix})^{(n)}(x)$ for $n = -1, 0, \dots$. We recall from Theorem 4.1 that $(v^{ix})^{(n)}(x) = m^{(n)}(i, x)$.

Theorem 5.2. *Let k be a positive integer and let π be a holding policy with ranking $\gg$ and continuation function $C(\cdot, \cdot)$. Suppose*

$$\begin{aligned} (i, x) \gg (j, y) \quad \Rightarrow \quad &\left(m^{(-1)}(i, x), \dots, m^{(k)}(i, x)\right) \\ &\gtrsim_{\text{lex}} \left(m^{(-1)}(j, y), \dots, m^{(k)}(j, y)\right) \\ &\qquad \textit{for all } (i, x), (j, y) \in J, \end{aligned} \tag{5.2}$$

and further suppose that for each pair $(i, x) \in J$ the stationary policy δ for MDC^{ix} *corresponding to $C(i, x)$ satisfies*

$$\left(w_\delta^{ix}\right)^{(n)}(x) = \left(v^{ix}\right)^{(n)}(x) \quad \textit{for all } n = -1, \dots, k. \tag{5.3}$$

Then $w_\pi^{(n)}(s) = v^{(n)}(s)$ for each $s \in S$ and $n = -1, \dots, k$.

Proof. Applying the expansion (4.1) of Theorem 4.1 to pairs (i, x), $(j, y) \in j$, (5.2) implies that if $(i, x) \gg (j, y)$, then for some $\rho' > 0$ and $K' > 0$,

$$m(i, x, \alpha) \geq m(j, y, \alpha) - K'\rho^{k+1} \quad \text{for all } 0 < \rho \leq \rho'.$$

Also, by (2.1) and (2.2), for each $(i, x) \in J$, (5.3) implies that for some $\rho'' > 0$ and $K'' > 0$,

$$\left(W_\delta^{ix}\right)(x, \alpha) \geq V^{ix}(x, \alpha) - K''\rho^{k+1} \quad \text{for all } 0 < \rho \leq \rho''.$$

Hence, for $0 < \rho \leq \rho^* \equiv \min\{\rho', \rho''\}$, the conditions of Theorem 5.1 are satisfied with $A = K'\rho^{k+1}$ and $B = K''\rho^{k+1}$, and the conclusion of that theorem shows that, with $K \equiv K' + K''$, $W_\pi(s, \rho) \geq V(s, \rho) - K\rho^{k+1}$ for all $s \in S$ and $0 < \rho < \rho^*$. For all $s \in S$ and $\rho > 0$, we also have that

$$\sum_{n=-1}^{\infty} \rho^n v^{(n)}(s) = V(s, \rho) \geq W_\pi(s, \rho) = \sum_{n=-1}^{\infty} \rho^n w_\pi^{(n)}(s).$$

Thus,

$$\left| \sum_{n=-1}^{\infty} \rho^n v^{(n)}(s) - \sum_{n=-1}^{\infty} \rho^n w_\pi^{(n)}(s) \right| \leq K\rho^{k+1},$$

immediately implying the conclusion of the theorem. □

Theorem 5.2 is next combined with the characterizations of optimal holding policies through (2.3) to obtain sufficient optimality conditions for holding policies. The proof follows the arguments used to deduce Theorem 4.3 from Theorem 4.2 and is left to the reader.

THEOREM 5.3 (Sufficient conditions for optimality of holding policies). *If π is a holding policy with ranking $\gg$ and continuation function $C(\cdot)$ such that* (5.2) *holds with $k = |S| + 1$, $k = 0$ or $k = 1$ and, respectively, for each $(i, x) \in J$ the stationary policy of MDC^{ix} corresponding to $C(i, x)$ is sensitive-discount, average-reward or average-overtaking optimal, then π is, respectively, sensitive-discount, average-reward or average-overtaking optimal.*

Theorem 5.3 suggests the following implementation for holding policies that are sensitive-discount optimal, average-reward optimal and average-overtaking optimal, respectively. Suppose state s is observed at a decision epoch. For each $i \in N$, determine the corresponding coefficients of the expansion of $V^{is_i}(s_i, \rho)$; in fact, past initialization, new coefficients have to be computed only for the single project that has been selected in the previous decision epoch (while the coefficients of the other projects do not change). Next, select a project i^* that lexicographically maximizes the corresponding coefficients, compute a corresponding stationary optimal policy $\delta^{i^* s_{i^*}}$ for MDC$^{i^* s_{i^*}}$ and use the continuation set determined by $\delta^{i^* s_{i^*}}$.

One can construct stationary policies that satisfy (5.2) for any specified nonnegative integer k. By Theorem 5.2, such policies are then optimal with respect to the optimality criteria mentioned at the end of Section 2.

Acknowledgments. We acknowledge useful comments of A. Mandelbaum, H. Kaspi and an anonymous referee. Also, K. D. Glazebrook referred us to his 1982 paper which allowed us to sharpen the original version of Proposition 3.1 and to eliminate a lengthy argument used to prove a weaker bound than the one established in Theorem 2 of his paper; in addition, the characterization of $m^{-1}(i, x)$ given in (4.2) was observed by Glazebrook.

REFERENCES

BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **32** 719–726.

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

DERMAN, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

GITTINS, J. C. (1989). *Multi Armed Bandit Allocation Indices*. Wiley-Interscience, New York.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design experiments. In *Progress in Statistics. European Meeting of Statisticians* (J. Gani, K. Sarkadi and I. Vince, eds.) **1** 241–266. North-Holland, Amsterdam.

GLAZEBROOK, K. D. (1982). On the evaluation of suboptimal strategies for families of alternative bandit processes. *J. Appl. Probab.* **19** 716–722.

GLAZEBROOK, K. D. (1990). Procedures for the evaluation of strategies for resource allocation in a stochastic environment. *J. Appl. Probab.* **27** 215–220.

GLAZEBROOK, K. D. (1991). Bounds for discounted stochastic scheduling problems. *J. Appl. Probab.* **28** 791–801.

KATEHAKIS, M. N. and VEINOTT, A. F., JR. (1987). The multi-armed bandit problem: decomposition and computation. *Math. Oper. Res.* **12** 262–268.

KELLY, F. P. (1981). Multi-armed bandits with discount factor near one: the Bernoulli case. *Ann. Statist.* **9** 987–1001.

LAI, T. S. and YING, Z. (1988). Open bandit processes and optimal scheduling of queuing networks. *Adv. in Appl. Probab.* **20** 447–472.

MILLER, B. L. and VEINOTT, A. F., JR. (1969). Discrete dynamic programming with a small interest rate discounting. *Ann. Math. Statist.* **40** 366–370.

ROSS, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York.

ROTHBLUM, U. J. and VEINOTT, A. F., JR. (1992). Branching Markov decision chains: immigration induced optimality. Unpublished manuscript.

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

VEINOTT, A. F., JR. (1969). Discrete dynamic programming with sensitive optimality criteria. *Ann. Math. Statist.* **40** 1635–1660.

VEINOTT, A. F., JR. (1974). Markov decision chains. In *Studies in Optimization* (G. B. Dantzig and B. C. Eaves, eds.) 124–159. Math. Association of America, Washington, DC.

WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42** 143–149.

WHITTLE, P. (1982). *Optimization over Time* **1**. Wiley, New York.

GRADUATE SCHOOL OF MANAGEMENT
RUTGERS UNIVERSITY
NEWARK, NEW JERSEY 07102
E-MAIL: mnk@andromeda.rutgers.edu

FACULTY OF INDUSTRIAL ENGINEERING
MANAGEMENT
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
TECHNION CITY, HAIFA 32000
ISRAEL
E-MAIL: rothblum@ie.technion.ac.il