The Annals of Applied Probability
1993, Vol. 3, No. 4, 1013–1032

INDICES FOR FAMILIES OF COMPETING MARKOV DECISION PROCESSES WITH INFLUENCE

BY K. D. GLAZEBROOK

University of Newcastle upon Tyne

Nash obtained an important extension to the classical theory of Gittins indexation when he demonstrated that index policies were optimal for a class of multiarmed bandit problems with a multiplicatively separable reward structure. We characterise the relevant indices (herein referred to as Nash indices) as equivalent retirement rewards/penalties for appropriately defined maximisation/minimisation problems. We also give a condition which is sufficient to guarantee the optimality of index policies for a Nash-type model in which each constituent bandit has its own decision structure.

1. Introduction. Gittins (1979) defines a *simple family of alternative bandit processes* (SFABP) as a discounted Markov decision process (MDP) with vector-valued state $X(t) = \{X_1(t), X_2(t), \ldots, X_N(t)\}$ at time $t \in \mathbb{N}$. Here $X_j(t)$ denotes the state of bandit j at t. Actions $a_1, a_2, \ldots, a_N$ are available at each decision epoch $t \in \mathbb{N}$, where a_j denotes the choice of bandit j. Upon choosing action a_j at t, the state of bandit j changes in a Markovian fashion, whereas $X_i(t+1) = X_i(t)$, $i \neq j$. A discounted reward $\alpha^t R_j\{X_j(t)\}$ is earned also. It may help the reader to think of this process in terms of N Markovian reward streams, each with an on/off switch. The rule is that only one switch is allowed to be on at any time. Bergman and Gittins (1985) describe applications of such processes to the planning of pharmaceutical research. Here bandit j models one of N projects or routes to research success and action a_j is the allocation of research effort to project j. Such an allocation is assumed to lead to an updating of the status of project j *only*. The goal is to choose a policy (i.e., a rule for selecting bandits) which maximises total reward.

In what has proved to be one of the most significant recent contributions to our understanding of policy structure in stochastic dynamic programming, Gittins (1979) demonstrated the existence of index-based policies which are optimal. Hence with bandit j is associated a real-valued function on its state space, denoted G_j, such that action a_j is optimal at t if and only if

$$G_j\{X_j(t)\} = \max_{1 \leq i \leq N} G_i\{X_i(t)\}$$

Received May 1992; revised March 1993.
AMS 1991 *subject classification.* 90C40.
Key words and phrases. Gittins index, Markov decision process, optimal policy, stopping time.

and ties can be broken in any way. Interpret $G_j(x_j)$ as the best reward rate achievable from further choices of a_j when bandit j is in state x_j.

Whittle (1980) coined the term *Gittins index* for these functions and made several important contributions to the subsequent development of this theory. In addition to his dynamic programming proof of Gittins' classical result, these include:

1. The characterisation of Gittins indices as *equivalent retirement rewards*. Hence, in a sense which needs careful definition, $G_j(x_j)$ may be thought of as the smallest reward which would be accepted in exchange for further opportunities to choose bandit j when it is in state x_j.
2. The enunciation of a condition sufficient to ensure the existence of an index policy which is optimal for a given *family of competing Markov decision processes*. Here, the foregoing model is elaborated such that each bandit has its own decision structure. This work has been subsequently utilised and developed by Gittins (1989), Glazebrook (1982, 1988, 1991) and Varaiya, Walrand and Buyukkoc (1985) and has facilitated the analysis of a rich class of stochastic scheduling problems. See Gittins (1989).

Nash (1980) developed the work of Gittins (1979) in an interesting direction by introducing a multiplicatively separable reward structure, in which the choice of a_j at t earns a reward $\alpha^t[\prod_{i\neq j} Q_i\{X_i(t)\}]R_j\{X_j(t)\}$. Otherwise his model was as in the preceding description. He called the resultant decision processes *generalised bandit problems*. These are described in detail in Section 3 where they are referred to as *families of competing Markovian reward processes with influence*. The major point of interest in Nash's model is that bandits are allowed to influence the returns from their competitors via *influence functions* Q_j. The modelling of influence in this way has been found to be very helpful in the analysis of a range of problems in research planning and stochastic scheduling; see Fay and Glazebrook (1987, 1989), Glazebrook and Owen (1991) and Glazebrook and Greatrix (1993). Nash (1980) demonstrated the existence of index-based policies which are optimal for this class of processes. The relevant indices (which we shall call Nash indices) modify Gittins indices to take account of the nature of a bandit's influence.

Theoretical development of Nash's work has in the main been restricted to a special case, namely, where the foregoing indices are always positive. See Gittins (1989), Glazebrook and Owen (1991) and Fay and Walrand (1991). As Gittins (1989) and Glazebrook and Owen (1991) point out, this special case is equivalent to a semi-Markov extension of Gittins' original framework. Little is understood about the general case of Nash's model beyond his original index result.

This paper contributes to the development of our understanding of Nash-type models with influence along the lines of Whittle's contribution to the Gittins framework, as outlined in contributions 1 and 2 listed previously. To be specific:

1. We characterise Nash indices as equivalent retirement rewards or penalties. See Section 2. An interesting feature here is that to accommodate

negative Nash indices (as in the general case we must) we need to develop a notion of an *equivalent retirement penalty* which is related to a decision problem in which expected rewards are minimised.

2. We enunciate a condition which is sufficient to ensure the existence of an index policy which is optimal for a Nash-type model in which each bandit has its own decision structure. We call such a process a *family of competing Markov decision processes with influence*. This work is the subject of Section 3.

The ideas of the paper are illustrated in Section 4 by an account of how the results of Section 3 relate to one of the simplest kinds of families of competing MDPs of real interest.

2. Nash indices. We now describe in more detail the individual objects of choice (referred to previously as bandits) in the Nash model and how indices are associated with them. We shall use $[\{X(t),\ t \in \mathbb{N}\}, R, Q, \alpha]$ to denote what we shall term a *Markovian reward process with influence* where the component parts of this process are:

1. $\{X(t),\ t \in \mathbb{N}\}$ a Markov chain whose state space Ω may be finite, countable or continuous.
2. R: $\Omega \to \mathbb{R}^+$, an associated *reward function* which is bounded.
3. Q: $\Omega \to \mathbb{R}^+$, an associated *influence function* which is bounded.
4. $\alpha \in [0, 1)$, a *discount rate*.

Following Whittle (1980), we introduce the *retirement reward* $M \in \mathbb{R}$, interpreted as a *penalty* if $M \leq 0$. If $M \geq 0$, we denote by $P^1(x, M)$ the stopping (or retirement) problem: Choose τ, a stopping time on $\{X(t),\ t \in \mathbb{N}\}$, to maximise

$$(2.1)\quad R(x, \tau, M) \triangleq E_\tau\left[\sum_{t=0}^{\tau-1} \alpha^t R\{X(t)\} + \alpha^\tau Q\{X(\tau)\}M \mid X(0) = x\right].$$

Hence, prior to retirement, rewards are accumulated by the process in a way determined by the function R. At retirement, the reward M is subject to influence through the function Q. Standard results in the theory of Markov decision processes [see, e.g., Ross (1970)] imply the existence of a stopping time τ attaining the supremum of $R(x, \tau, M)$, which is deterministic, stationary and Markov. Call the maximised value $R^1(x, M)$.

If $M \leq 0$, denote by $P^2(x, M)$ the retirement problem: Choose τ, a stopping time on $\{X(t),\ t \in \mathbb{N}\}$ to minimise $R(x, \tau, M)$. We can again assert the existence of a deterministic, stationary and Markov stopping time which attains the infimum. Call the minimised value $R^2(x, M)$.

If we set $Q(\cdot) \equiv 1$ in $P^1(\cdot, M)$, then we obtain precisely the set of retirement problems considered by Whittle (1980) in the context of Gittins indexation. The novelty here is the introduction of the minimisation problems $P^2(\cdot, M)$ as a means of accommodating negative Nash indices.

To take the analysis further we require the following definition.

DEFINITION 1. The *influential set of states* $\Theta \subseteq \Omega$ is given by

$$\Theta = \{x \in \Omega;\ \exists\, \tau, \text{ positive-valued, such that } E_\tau[\alpha^\tau Q\{X(\tau)\}|X(0) = x] > Q(x)\}.$$

Further, for any $x \in \Theta$ we write $T(x)$ for the set of positive-valued, deterministic, stationary and Markov stopping times satisfying $E_\tau[\alpha^\tau Q\{X(\tau)\}|X(0) = x] > Q(x)$.

Hence from any state $x \in \Theta$, some further continuation of the process may be made in such a way that the retirement term in (2.1) exceeds the equivalent present value $MQ(x)$ when $M \geq 0$ and is below it for $M \leq 0$. It should be clear that in the problem $P^1(x, M)$, $M \geq 0$, any $x \in \Theta$ must be a continuation state, that is, an optimal stopping time need never choose to retire in x. We also see that in Nash's model any process currently in a state lying in its influential set plainly has a special status, because appropriate additional choices of that process will lead to an enhancement (in expectation) of the returns from other processes. Note that if Q is a constant function, then $\Theta = \phi$. Note also that because all stopping times discussed henceforth will be deterministic, stationary and Markov we shall omit this qualifying phrase throughout the paper.

We develop the notion of Nash indexation of a Markovian reward process with influence as follows.

DEFINITION 2. The Nash index N: $\Omega \to \{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$ for process $[\{X(t),\ t \in \mathbb{N}\}, R, Q, \alpha]$ is defined as follows:

(i) If $x \notin \Theta$, then

$$N(x) = \inf\{M;\ M \geq 0 \text{ and } R^1(x, M) = MQ(x)\}. \tag{2.2}$$

(ii) If $x \in \Theta$, then

$$N(x) = \inf\{M;\ M \leq 0 \text{ and } R^2(x, M) = MQ(x)\}. \tag{2.3}$$

If the condition specified on the r.h.s. of (2.2) is satisfied for no $M \geq 0$, we write $N(x) = \infty$ and if for all $M \geq 0$, we write $N(x) = 0^+$. If the condition in (2.3) is satisfied for no $M \leq 0$, we write $N(x) = 0^-$. No other cases are possible.

Hence for noninfluential states, the Nash index is an equivalent retirement reward. It is (uniquely) that retirement reward for which *both* retirement *and* nonretirement are optimal for the maximisation problem P^1. For influential states, the index is an equivalent retirement penalty, defined now with respect to the minimisation problem P^2.

From Definition 2 it is not difficult to recover the characterisation of Nash indices given by Nash (1980) himself. This is Lemma 1.

LEMMA 1. *The Nash index* $N\colon \Omega \to \{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$ *for* $[\{X(t),\ t \in \mathbb{N}\}, R, Q, \alpha]$ *is as follows*:

(i) *If* $x \notin \Theta$, *then*

$$N(x) = \sup_{\tau > 0} \left\{ R(x, \tau, 0)\big(Q(x) - E_\tau[\alpha^\tau Q\{X(\tau)\} | X(0) = x]\big)^{-1} \right\}, \tag{2.4}$$

the supremum in (2.4) *being taken over the set of positive-valued stopping times.*

(ii) *If* $x \in \Theta$, *then*

$$N(x) = \sup_{\tau \in T(x)} \left\{ R(x, \tau, 0)\big(Q(x) - E_\tau[\alpha^\tau Q\{X(\tau)\} | X(0) = x]\big)^{-1} \right\}. \tag{2.5}$$

In (2.4) *and* (2.5), *we take* $+/0$, $0/+$, $0/-$ *and* $0/0$ *to be* ∞, 0^+, 0^- *and* 0^+, *respectively.*

PROOF OF LEMMA 1(ii). Suppose that $x \in \Theta$ with $0 > N(x) > -\infty$. Following Ross (1983), it is not difficult to show that over $M \le 0$, $R^2(x, M) - MQ(x)$ is decreasing in M. Hence from (2.3), $0 > M \ge N(x) \Rightarrow MQ(x) = R^2(x, M)$ and so for all positive stopping times $\tau \in T(x)$, $MQ(x) \le R(x, \tau, M)$. Following some simple algebra we conclude that

$$\begin{aligned} &0 > M \ge N(x) \Rightarrow M \\ &\quad \ge R(x, \tau, 0)\big(Q(x) - E_\tau[\alpha^\tau Q\{X(\tau)\} | X(0) = x]\big)^{-1}, \qquad \tau \in T(x). \end{aligned} \tag{2.6}$$

Also from (2.3), $M < N(x) \Rightarrow MQ(x) > R^2(x, M)$. Hence there exists a positive stopping time τ satisfying $MQ(x) > R(x, \tau, M)$. Because M is negative, from (2.1) we deduce that any such τ must satisfy $E_\tau[\alpha^\tau Q\{X(\tau)\} | X(0) = x] > Q(x)$, that is, must be a member of $T(x)$. Following some simple algebra we conclude that for some $\tau \in T(x)$,

$$M < N(x) \Rightarrow M < R(x, \tau, 0)\big(Q(x) - E_\tau[\alpha^\tau Q\{X(\tau)\} | X(0) = x]\big)^{-1}. \tag{2.7}$$

Lemma 1(ii) follows immediately from (2.6) and (2.7) for the case being studied. $N(x) = 0^-$ and all cases which fall under Lemma 1(i) may be dealt with similarly. □

To summarise, Nash indices for influential states are negative. For noninfluential states with positive indices the Nash index is a maximised ratio of reward earned up to some stopping time to the extent to which (discounted) influence declines. In the simple case with Q constant, the index (now a Gittins index) has a simpler interpretation as a maximised reward rate.

EXAMPLES. It may assist the reader to refer to Fay and Glazebrook (1987, 1989) and Glazebrook and Greatrix (1993) for some examples of Markovian reward processes with influence and their associated indices. Fay and

Glazebrook (1987, 1989) discuss models relevant to research planning for which the associated Nash indices are always positive. In the stochastic scheduling models of Glazebrook and Greatrix (1993), the nature of influence is such that negative indices do occur.

We now generalise the notion of Nash indexation to the case where the Markov chain $\{X(t), t \in \mathbb{N}\}$ in process part 1 has an associated decision structure. We introduce $\{[X(t); A\{X(t)\}, t \in \mathbb{N}], R, Q, \alpha\}$ as a *Markov decision process* (MDP) *with influence* as follows:

1′. $X(t)$ is the state of the process at $t \in \mathbb{N}$. The state space Ω may be finite, countable or continuous.
2′. $A(x)$ is the finite *action set* for state $x \in \Omega$. In the MDP with influence, at every decision epoch $t \in \mathbb{N}$, an action a from $A\{X(t)\}$ is taken. A *policy* π is any rule for choosing actions which is a function of the history of the process to date.
3′. R: $\Omega \times A(\cdot) \to \mathbb{R}^+$ is a bounded reward function. Should action $a \in A\{X(t)\}$ be taken at decision epoch $t \in \mathbb{N}$, a reward $\alpha^t R\{X(t), a\}$ is earned.
4′. If action $a \in A\{X(t)\}$ is taken at time $t \in \mathbb{N}$, then the Markovian transition law $P(\cdot|x, a)$ yields the distribution of $X(t+1)$ conditional upon the event $X(t) = x$.
5′. Q: $\Omega \to \mathbb{R}^+$ is a bounded influence function.
6′. $\alpha \in [0, 1)$ is a discount rate.

To develop Nash indices for such a MDP with influence, we again introduce the retirement reward $M \in \mathbb{R}$. If $M \geq 0$, we denote by $P^3(x, M)$ the retirement problem: choose π, a policy for the MDP and τ, a stopping time on the MDP under π, to maximise

$$R(x, \pi, \tau, M) \triangleq E_{\pi,\tau}\left(\sum_{t=0}^{\tau-1} \alpha^t R[X(t), \pi\{X(t)\}] + \alpha^\tau Q\{X(\tau)\} M | X(0) = x\right).$$

As before, we may assert the existence of deterministic, stationary and Markov π and τ attaining the supremum of $R(x, \pi, \tau, M)$. Call the maximised value $R^3(x, M)$. As with stopping times, we shall henceforth assume that policies are always deterministic, stationary and Markov and shall drop the qualifying phrase.

If $M \leq 0$, denote by $P^4(x, M)$ the retirement problem: Choose π, a policy for the MDP and τ, a stopping time on the MDP under π, to minimise $R(x, \pi, \tau, M)$. Again, we may assert that the infimum is attained and call the minimised value $R^4(x, M)$. We develop Definition 1 as follows:

DEFINITION 3. The *influential set of states* $\Theta \subseteq \Omega$ is given by

$$\Theta = \{x \in \Omega; \exists \text{ policy } \pi \text{ and } \tau, \text{ a positive stopping time on the MDP under } \pi, \text{ such that } E_{\pi,\tau}[\alpha^\tau Q\{X(\tau)\} | X(0) = x] > Q(x)\}.$$

Further, for any $x \in \Theta$ and policy π, write $T_\pi(x)$ for the set of positive stopping times on the MDP under π (if any) satisfying $E_{\pi,\tau}[\alpha^\tau Q\{X(\tau)\}| X(0) = x] > Q(x)$.

We may now develop the notion of Nash indexation of a MDP with influence as follows:

DEFINITION 4. The Nash index N: $\Omega \to \{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$ for process $([X(t); A\{X(t)\}, t \in \mathbb{N}], R, Q, \alpha)$ is defined as follows:

(i) If $x \notin \Theta$, then

$$N(x) = \inf\{M;\ M \geq 0 \text{ and } R^3(x, M) = MQ(x)\}. \tag{2.8}$$

(ii) If $x \in \Theta$, then

$$N(x) = \inf\{M;\ M \leq 0 \text{ and } R^4(x, M) = MQ(x)\}. \tag{2.9}$$

As in Definition 2, we may obtain the values 0^+ and ∞ when $x \notin \Theta$ and 0^- when $x \in \Theta$.

The proof of Lemma 2 is along the lines of that of Lemma 1 and will not be given.

LEMMA 2. *The Nash index* N: $\Omega \to \{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$ *for* $([X(t); A\{X(t)\}, t \in \mathbb{N}], R, Q, \alpha)$ *is as follows*:

(i) *If* $x \notin \Theta$, *then*

$$N(x) = \sup_{\pi,\tau} \Big\{ R(x, \pi, \tau, 0) \times \big(Q(x) - E_{\pi,\tau}[\alpha^\tau Q\{X(\tau)\}|X(0) = x]\big)^{-1} \Big\}, \tag{2.10}$$

the supremum in (2.10) *being taken over all pairs* (π, τ), *where* π *is a policy and* τ *is a positive-valued stopping time on the MDP under* π.

(ii) *If* $x \in \Theta$, *then*

$$N(x) = \sup_{\pi,\tau \in T_\pi(x)} \Big\{ R(x, \pi, \tau, 0) \times \big(Q(x) - E_{\pi,\tau}[\alpha^\tau Q\{X(\tau)\}|X(0) = x]\big)^{-1} \Big\}, \tag{2.11}$$

the supremum being over all pairs (π, τ) *with* $\tau \in T_\pi(x)$. *The same conventions regarding zeros in the numerator and/or denominator of the expressions in* (2.10) *and* (2.11) *apply as before.*

Before concluding this preliminary discussion of Nash indices for Markovian reward and decision processes with influence, we need to make some observations concerning the stopping times and/or policies which attain

the suprema in Lemmas 1 and 2. To do that in a natural way, we need to introduce a total ordering on the range of the Nash index function, namely $\{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$. This ordering, described in Definition 5, reflects how preferences between Markovian reward processes with influence are determined by ther Nash indices. To summarise, Nash indices with negative indices (i.e., in influential states) are always preferred to processes with positive indices. When a collection of indices have the same sign, larger values are preferred.

DEFINITION 5. The *Nash index preference ordering* $\succcurlyeq$ is a total ordering on the set $\{\mathbb{R} \setminus 0\} \cup \{0^+, 0^-, \infty\}$ such that if x_1, x_2, y_1, y_2 are any real numbers satisfying $x_1 \leq x_2 < 0 < y_1 \leq y_2$, then

$$0^- \succcurlyeq x_2 \succcurlyeq x_1 \succcurlyeq \infty \succcurlyeq y_2 \succcurlyeq y_1 \succcurlyeq 0^+.$$

The strict version $\succ$ of the ordering is such that for any reals x_1, x_2, y_1, y_2 satisfying $x_1 < x_2 < 0 < y_1 < y_2$, then

$$0^- \succ x_2 \succ x_1 \succ \infty \succ y_2 \succ y_1 \succ 0^+.$$

Lemma 3 describes the stopping times on $\{X(t), t \in \mathbb{N}\}$ attaining the suprema in (2.4) and (2.5) in terms of the Nash index preference ordering.

LEMMA 3. (i) *If* $M > 0$, *then* $R^1(x, M) = R\{x, \hat{\tau}(x, M), M\}$, *where*

$$\hat{\tau}(x, M) = \inf[t; t \geq 0 \text{ and } N\{X(t)\} \preccurlyeq M]. \tag{2.12}$$

(ii) *If* $M < 0$, *then* $R^2(x, M) = R\{x, \hat{\tau}(x, M), M\}$, *where*

$$\hat{\tau}(x, M) = \inf[t; t \geq 0 \text{ and } N\{X(t)\} \preccurlyeq M]. \tag{2.13}$$

(iii) *If* $M = 0$, *then* $R^1(x, 0) = R\{x, \hat{\tau}(x, 0^+), 0\}$, *where* $\hat{\tau}(x, 0^+)$ *is as in* (2.12) *with* $M = 0^+$ *and* $R^2(x, 0) = R\{x, \hat{\tau}(x, 0^-), 0\}$, *where* $\hat{\tau}(x, 0^-)$ $(= 0$ a.s.$)$ *is as in* (2.13) *with* $M = 0^-$.

(iv) *The suprema in* (2.4) *and* (2.5) *are satisfied by* $\tilde{\tau}(x)$, *where*

$$\tilde{\tau}(x) = \inf[t; t > 0 \text{ and } N\{X(t)\} \preccurlyeq N(x)]. \tag{2.14}$$

(v) *Results* (i)–(iv) *all hold with* $\preccurlyeq$ *replaced by* $\prec$ *throughout. These are the only choices for the stopping times concerned.*

PROOF. (ii) Refer back to the proof of Lemma 1. We saw there that in $P^2(\cdot, M)$ with $M < 0$ it is optimal to stop in some state y if $N(y) \preccurlyeq M$. If $X(0) = x$, then $\hat{\tau}(x, M)$ is simply the first decision epoch at which the process is in such a state. The proofs of (i) and (iii) proceed similarly.

(iv) Suppose that $x \notin \Theta$ and $N(x) > 0$. From (i), $R^1\{x, N(x)\} = R[x, \hat{\tau}\{x, N(x)\}, N(x)] = R\{x, \tilde{\tau}(x), N(x)\}$. The latter equality follows from (2.12) and (2.14) by noting that $\hat{\tau}\{x, N(x)\} \equiv \tilde{\tau}(x)$ and that in $P^1\{x, N(x)\}$ both stopping and continuing are optimal at $t = 0$. Invoking the continuity of

$R^1(x, M)$ over $M \geq 0$ we now deduce from the characterisation of $N(x)$ given in Definition 2 that

$$
\begin{aligned}
Q(x)N(x) &= R^1\{x, N(x)\} = R\{x, \tilde{\tau}(x), N(x)\} \\
&= R\{x, \tilde{\tau}(x), 0\} \\
&\quad + E_{\tilde{\tau}(x)}\big(\alpha^{\tilde{\tau}(x)} Q[X\{\tilde{\tau}(x)\}] | X(0) = x\big) N(x).
\end{aligned} \tag{2.15}
$$

Solving for $N(x)$ in (2.15) we are able to infer that $\tilde{\tau}(x)$ attains the supremum in (2.4). Other cases are dealt with similarly.

(v) Close examination of the proof of Lemma 1 yields the conclusion that $\preccurlyeq$ may be replaced by $\prec$ throughout and that any other choice of the respective stopping times is suboptimal. □

In proceeding from Lemma 3 to an equivalent analysis for MDPs with influence it is plain that we have a much more substantial problem because now all suprema (infima) are over choices of policies for the MDP as well as over stopping times. A crucial simplifying step in the development of the theory is to consider cases where in the retirement problems $P^3(\cdot, M)$ and $P^4(\cdot, M)$ there exists a single policy which attains the relevant supremum (infimum) for all choices of M. Because such a policy will be optimal for $P^3(\cdot, 0)$, it will inter alia be optimal for the MDP with respect to the usual discounted criterion over an infinite horizon. "Optimal" is used in such a sense in Definition 6.

DEFINITION 6. Optimal policy $\hat{\pi}$ is *dominating* for $([X(t); A\{X(t)\}, t \in \mathbb{N}], R, Q, \alpha)$ if:

(i) $R^3(x, M) = \sup_\tau R(x, \hat{\pi}, \tau, M)$, $x \in \Omega$, $M \geq 0$, and
(ii) $R^4(x, M) = \inf_\tau R(x, \hat{\pi}, \tau, M)$, $x \in \Omega$, $M \leq 0$.

If a dominating policy exists we say that the *MDP* with influence satisfies Condition D.

The simpler idea of a dominating policy for families of competing MDPs (i.e., without influence) was first given expression by Whittle (1980). Note that Condition D is by no means guaranteed. See Example 3.13 in Gittins (1989) for an instance where it fails in a simple Gittins index case. However, we shall see later in the paper important examples where Condition D is satisfied. Lemma 4 follows trivially from Lemma 3 and the observation that an MDP with influence *together with* a prespecified policy π for choosing actions is simply a Markovian reward process with influence. Use the notation $N(\pi, x)$ for the resulting Nash index as a means of emphasising the dependence upon the choice of policy.

LEMMA 4. *If $\hat{\pi}$ is dominating for $([X(t); A\{X(t)\}, t \in \mathbb{N}\}, R, Q, \alpha)$, then we have the following*:

(i) *If $M > 0$, then $R^3(x, M) = R\{x, \hat{\pi}, \hat{\tau}(x, M), M\}$, where $\hat{\tau}(x, M)$ is a stopping time on the MDP under $\hat{\pi}$ defined by*

$$\hat{\tau}(x, M) = \inf[t; t \geq 0 \text{ and } N\{\hat{\pi}, X(t)\} \preccurlyeq M]. \tag{2.16}$$

(ii) *If $M < 0$, then $R^4(x, M) = R\{x, \hat{\pi}, \hat{\tau}(x, M), M\}$, where $\hat{\tau}(x, M)$ is a stopping time on the MDP under $\hat{\pi}$ defined by*

$$\hat{\tau}(x, M) = \inf[t; t \geq 0 \text{ and } N\{\hat{\pi}, X(t)\} \preccurlyeq M]. \tag{2.17}$$

(iii) *If $M = 0$, then $R^3(x, 0) = R(x, \hat{\pi}, \hat{\tau}(x, 0^+), 0\}$, where $\hat{\tau}(x, 0^+)$ is as in* (2.16) *with $M = 0^+$ and $R^4(x, 0) = R\{x, \hat{\pi}, \hat{\tau}(x, 0^-), 0\}$, where $\hat{\tau}(x, 0^-)$ (= 0 a.s.) is as in* (2.17) *with $M = 0^-$.*

(iv) *The suprema in* (2.10) *and* (2.11) *are satisfied by $\{\hat{\pi}, \hat{\tau}(x)\}$, where $\hat{\tau}(x)$ is a stopping time on the MDP under $\hat{\pi}$ defined by*

$$\hat{\tau}(x) = \inf[t; t > 0 \text{ and } N\{\hat{\pi}, X(t)\} \preccurlyeq N(x)]. \tag{2.18}$$

(v) *Results* (i)–(iv) *all hold with $\preccurlyeq$ replaced by $\prec$ throughout. These are the only choices for the stopping times concerned.*

(vi) *$N(x) = N(\hat{\pi}, x) \succcurlyeq N(\pi, x)$ $\forall$ π, $x \in \Omega$. If no dominating policy exists, we still have $N(x) \succcurlyeq N(\pi, x)$ $\forall$ π, $x \in \Omega$.*

If we combine Lemma 4(vi) with (i) and (ii) we see that the stopping times $\hat{\tau}(x, M)$ in (2.16) and (2.17) can be characterised as

$$\hat{\tau}(x, M) = \inf[t; t \geq 0 \text{ and } N\{X(t)\} \preccurlyeq M].$$

Similarly the stopping time $\hat{\tau}(x)$ in (2.18) is given by

$$\hat{\tau}(x) = \inf[t; t \geq 0 \text{ and } N\{X(t)\} \prec N(x)].$$

3. Families of competing MDPs with influence. In Section 2 we described the individual objects of choice in a Nash-type model and the indices associated with them. We now describe in detail the model obtained when N MDPs with influence are brought together in the way outlined briefly in the introduction. The idea is that to form a family of competing MDPs, each MDP is provided with an on/off switch, the rule being that only one can be switched on at a time. A policy for the family will have to specify *both* how to choose which MDP to switch on *and* how to choose an action for that MDP. We use the expanded notation $\{([X_i(t); A_i\{X_i(t)\}, t \in \mathbb{N}], R_i, Q_i, \alpha), 1 \leq i \leq N\}$ or, more usually, the abbreviated form $\{(\text{MDP}_i, Q_i), 1 \leq i \leq N\}$ to denote a *family of competing MDPs with influence*, this being a discounted MDP with the following special features:

1. Its state at time $t \in \mathbb{N}$ is $X(t) = \{X_1(t), X_2(t), \ldots, X_N(t)\}$, where $X_j(t)$ denotes the state of MDP_j at time t, which must lie in the general state space Ω_j, $1 \leq j \leq N$. Hence the process state is simply the vector of the states of the constituent MDPs.

2. $A(x)$ is the action set for $x \in \times_{j=1}^{N} \Omega_j$ and is given by
$$A(x) = \bigcup_{j=1}^{N} A_j(x_j),$$
where $A_j(x_j)$ is the action set (assumed finite) for MDP_j in state x_j, $1 \le j \le N$. In summary, at each decision epoch a single action is taken for one of the constituent MDPs. A policy is a rule for choosing actions which depend only upon the history of the process to date.
3. The decision epochs are the natural numbers $\mathbb{N}$. Should action $a \in A_j\{X_j(t)\}$ be taken at $t \in \mathbb{N}$, then the Markovian transition law for MDP_j, $P_j(\cdot|x, a)$ say, yields the distribution of $X_j(t+1)$ conditional upon the event $X_j(t) = x$ independently of the states and histories of MDP_i, $i \ne j$, at t. Under this scenario, $X_i(t+1) = X_i(t)$, $i \ne j$, that is, *only* the state of MDP_j changes.
4. The expected reward earned should action $a \in A_j\{X_j(t)\}$ be taken at t is given by
$$\alpha^t \Big[\prod_{i \ne j} Q_i\{X_i(t)\}\Big] R_j\{X_j(t), a\}, \tag{3.1}$$
where each R_j: $\Omega_j \times A_j(\cdot) \to \mathbb{R}^+$ is a bounded reward function, each Q_j: $\Omega_j \to \mathbb{R}^+$ is a bounded influence function and $\alpha \in [0, 1)$ is a discount rate.
5. An optimal policy maximises the total expected reward earned during $[0, \infty)$.

Note that the constituent MDPs interact only through the influence functions; see (3.1). Hence the rewards received upon activating an MDP are influenced by the states of the other MDPs in a multiplicative manner.

An important special case both in its own right and for the development of the theory is where
$$\big|A_j(x)\big| = 1, \qquad x \in \Omega_j, \qquad 1 \le j \le N. \tag{3.2}$$
Hence there are no choices of action *within* MDPs, only *between* them. This is plainly equivalent to bringing together a collection of Markovian reward processes with influence (as described in Section 2) into a family by providing each one with an on/off switch as before. Nash (1980) only studied this class of models and called them *generalised bandit problems*. To be consistent with our other terminology, we shall refer to a *family of competing Markovian reward processes with influence*. We shall use the expanded notation $([\{X_i(t), t \in \mathbb{N}\}, R_i, Q_i, \alpha], 1 \le i \le N)$ or, more usually, the abbreviated form $\{(\text{MRP}_i, Q_i), 1 \le i \le N\}$ in the obvious way. The action set for such a family is $(a_1, a_2, \ldots, a_N)$, say, in all states, where a_j denotes the choice of MRP_j.

The central theme of the paper is the status of policies for families of competing MDPs (MRPs) with influence which choose between the constituent MDPs (MRPs) offered on the basis of their Nash indices. The following definition introduces a central concept. Note that N_i denotes the Nash index for MRP_i with influence function Q_i, $1 \le i \le N$.

DEFINITION 7. Any policy π for $\{(\text{MRP}_i, Q_i), 1 \le i \le N\}$ satisfying

$$\pi(x) = a_i \Rightarrow N_i(x_i) \succcurlyeq N_j(x_j), \qquad j \neq i, \qquad x \in \bigtimes_{j=1}^{N} \Omega_j,$$

is a *Nash index* policy.

Such a policy uses the Nash index preference ordering to choose between MRPs on the basis of their indices. If two or more MRPs share the same index value, then a Nash index policy may choose any of them. In none of the results in the paper does it matter how such ties are broken. When we refer to some specific Nash index policy we shall assume that some tie-breaking rule is in force. The following result is a restatement of Nash's (1980) main result.

THEOREM 5 [Nash (1980)]. *Any Nash index policy is optimal for* $\{(MRP_i, Q_i), 1 \le i \le N\}$.

To take our ideas further, consider now the family $\{(\text{MDP}_i, Q_i), 1 \le i \le N\}$. We shall suppose that each MDP_i is provided with a policy π_i for choosing its actions. Equivalently, denote by $\{(\text{MDP}_i, \pi_i, Q_i), 1 \le i \le N\}$ the family described in special features 1–5 with the restriction that the action set $A_j(x_j) = \{\pi_j(x_j)\}$, $x_j \in \Omega_j$, $1 \le j \le N$; that is, actions within MDPs *must* be chosen according to policies π_j, $1 \le j \le N$. This construction reduces the family of competing MDPs to one of competing MRPs with influence because within MDPs choice has been eliminated. From Nash (1980) we deduce that optimal policies for $\{(\text{MDP}_i, \pi_i, Q_i), 1 \le i \le N\}$, choose *within* MDPs (necessarily) according to the π_j, $1 \le j \le N$, and *between* MDPs according to Nash indices $N_j(\pi_j, \cdot)$, $1 \le j \le N$. We shall use the notation $N(\pi_1, \pi_2, \ldots, \pi_N) \equiv N(\boldsymbol{\pi})$ to denote such a policy; that is,

$$\{N(\boldsymbol{\pi})\}(x) = a \in A_i(x_i) \Rightarrow a = \pi_i(x_i) \quad \text{and}$$

$$N_i(\pi_i, x_i) \succcurlyeq N_j(\pi_j, x_j), \qquad j \neq i, \qquad x \in \bigtimes_{j=1}^{N} \Omega_j.$$

The following is an immediate consequence of Theorem 5.

COROLLARY 6. *Any Nash index policy* $N(\boldsymbol{\pi})$ *is optimal for* $\{(MDP_i, \pi_i, Q_i), 1 \le i \le N\}$.

Corollary 6 asserts the optimality of $N(\boldsymbol{\pi})$ for the family $\{(\text{MDP}_i, Q_i), 1 \le i \le N\}$ for the *constrained* problem in which actions for MDP_i must be chosen according to policy π_i, $1 \le i \le N$. Our primary interest, though, concerns when there exists some choice of the π_i, $1 \le i \le N$, such that $N(\boldsymbol{\pi})$ is globally optimal. Theorem 7 states that the existence of a dominating policy for each constituent MDP with influence is sufficient to guarantee this.

THEOREM 7. *If each* (MDP_i, Q_i) *satisfies Condition* D *with optimal policy* $\hat{\pi}_i$ *being dominating, then* $N(\hat{\boldsymbol{\pi}}) \equiv N(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N)$ *is optimal for the family* $\{(MDP_i, Q_i),\ 1 \le i \le N\}$.

PROOF. Fix initial state $x \in \times_{j=1}^{N} \Omega_j$ and initial choice of action $a \in A(x)$. By standard dynamic programming arguments it is enough to show that

$$\mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\} \le \mathscr{R}\{x, N(\hat{\boldsymbol{\pi}})\}, \qquad a \in A(x), \qquad x \in \bigtimes_{j=1}^{N} \Omega_j. \tag{3.3}$$

In (3.3) the notation $\mathscr{R}(x, \pi)$ is used for the total discounted reward over an infinite horizon earned from policy π applied to the family $\{(\mathrm{MDP}_i, Q_i),\ 1 \le \mathrm{i} \le N\}$ when $X(0) = x$. Further $aN(\hat{\boldsymbol{\pi}})$ is used for the policy which chooses action a at $t = 0$ and thereafter operates $N(\hat{\boldsymbol{\pi}})$. Note that there is a trivial way of extending the system state to ensure that $aN(\hat{\boldsymbol{\pi}})$ is deterministic, stationary and Markov. By an argument due to Gittins [(1989), page 60] it is enough to establish that

$$\mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\} \le \mathscr{R}\{x, N(\hat{\boldsymbol{\pi}})\}, \qquad a \in A_i(x_i),$$

for those $x \in \times_{j=1}^{N} \Omega_j$ satisfying

$$N_i(x_i) \succcurlyeq N_j(x_j), \qquad j \ne i. \tag{3.4}$$

Now fix x and suppose w.l.o.g. that $N_1(x_1) \succcurlyeq N_j(x_j)$, $j \ne 1$. Fix $a \in A_1(x_1)$. We shall further suppose that (in an obvious notation) $N_1(a\hat{\pi}_1, x_1) \succcurlyeq N_j(x_j)$, $j \ne 1$. This latter condition will be removed subsequently.

We now proceed to develop formulae for $\mathscr{R}\{x, N(\hat{\boldsymbol{\pi}})\}$ and $\mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\}$. To this end, consider the application of policy $N(\hat{\pi}_2, \ldots, \hat{\pi}_N) \equiv N^1(\hat{\boldsymbol{\pi}})$ to $\{(\mathrm{MDP}_i, Q_i),\ 2 \le i \le N\}$. Denote by $R^1\{x^1, N^1(\hat{\boldsymbol{\pi}}), t, 0\}$ the total expected reward up to time t from initial state $x^1 = (x_2, \ldots, x_N) \in \times_{j=2}^{N} \Omega_j$ and by $X^1(t)$ the state of $\{(\mathrm{MDP}_i, Q_i),\ 2 \le i \le N\}$ under policy $N^1(\hat{\boldsymbol{\pi}})$ at t. We further develop the sequence $\{t_0, t_1, t_2, \ldots\}$ of random times as follows:

$$\begin{aligned} t_0 &= 0, \\ \{N^1(\hat{\boldsymbol{\pi}})\}(x^1) &= \hat{\pi}_i(x_i) \Rightarrow t_1 = \inf\big[t;\ t > 0 \text{ and } N_i\{X_i(t)\} \prec N_i(x_i)\big], \\ \{N^1(\hat{\boldsymbol{\pi}})\}\{X^1(t_1)\} &= \hat{\pi}_j\{X_j(t_1)\} \Rightarrow t_2 \\ &= \inf\big[t;\ t > t_1 \text{ and } N_j\{X_j(t)\} \prec N_j\{X_j(t_1)\}\big], \end{aligned}$$

and so on. For general $n \in \mathbb{Z}^+$,

$$\begin{aligned} \{N^1(\hat{\boldsymbol{\pi}})\}\{X^1(t_{n-1})\} &= \hat{\pi}_k\{X_k(t_{n-1})\} \Rightarrow t_n \\ &= \inf\big[t;\ t > t_{n-1} \text{ and } N_k\{X_k(t)\} \prec N_k\{X_k(t_{n-1})\}\big]. \end{aligned}$$

Corresponding to the sequence $\{t_0, t_1, t_2, \ldots\}$ we have the sequence $\{M_0, M_1, M_2, \ldots\}$ of Nash indices for the MDP chosen by policy $N^1(\hat{\boldsymbol{\pi}})$ at times $t_0, t_1, t_2, \ldots$, respectively. Plainly, by Lemma 4(vi), for all $n \in \mathbb{N}$,

$$\{N^1(\hat{\boldsymbol{\pi}})\}\{X^1(t_n)\} = \hat{\pi}_k\{X_k(t_n)\} \Rightarrow M_n = N_k\{X_k(t_n)\} = N_k\{\hat{\pi}_k, X_k(t_n)\}.$$

We also use the notation

$$R_1(x_1, \pi_1, \tau_1, M) = E_{\pi_1, \tau_1}\Bigg(\sum_{t=0}^{\tau_1 - 1} \alpha^t R_1\big[X_1(t), \pi_1\{X_1(t)\}\big] + \alpha^{\tau_1} Q_1\{X_1(\tau_1)\} M \mid X_1(0) = x_1 \Bigg).$$

Last, we use $\hat{\tau}_1(\hat{\pi}_1, M)$ for the stopping time on MDP_1 under policy $\hat{\pi}_1$, defined by

$$\begin{aligned} \hat{\tau}_1(\hat{\pi}_1, M) &= \inf\big[t; t \geq 0 \text{ and } N_1\{\hat{\pi}_1, X_1(t)\} \prec M\big] \\ &= \inf\big[t; t \geq 0 \text{ and } N_1\{X_1(t)\} \prec M\big]. \end{aligned}$$

The stopping time $\hat{\tau}_1(a\hat{\pi}_1, M)$ is defined similarly. If $M \preccurlyeq N_j(x_j)$ for some $j \neq 1$, then the assumption $N_1(a\hat{\pi}_1, x_1) \succcurlyeq N_j(x_j)$, $j \neq 1$, implies that for such an M, $\hat{\tau}_1(a\hat{\pi}_1, M)$ may be characterised as the stopping time on MDP under policy $a\hat{\pi}_1$ satisfying

$$\hat{\tau}_1(a\hat{\pi}_1, M) = \inf\big[t; t > 0 \text{ and } N_1\{X_1(t)\} \prec M\big].$$

With these preliminaries, a modified version of the accounting exercise used by Gittins [(1989), page 62] yields

$$\begin{aligned} \mathscr{R}\{x, N(\hat{\pi})\} = &\sum_{n=0}^{\infty} E_n\big[R_1\{x_1, \hat{\pi}_1, \hat{\tau}_1(\hat{\pi}_1, M_n), M_n\}\big] \\ &\times E_n\Bigg[\alpha^{t_n} \prod_{j \neq 1} Q_j\{X_j(t_n)\} - \alpha^{t_{n+1}} \prod_{j \neq 1} Q_j\{X_j(t_{n+1})\}\Bigg]. \end{aligned} \tag{3.5}$$

Note that in (3.5), E_n is an expectation operator conditional upon t_n and $X^1(t_n)$. We have an equivalent expression for $\mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\}$ obtained by replacing $\hat{\pi}_1$ by $a\hat{\pi}_1$ throughout (3.5). We now consider two cases for M_n.

CASE 1. $M_n < 0$ or $M_n = 0^-$. In this event it is clear from the fact that $\hat{\pi}_1$ is dominating for (MDP_1, Q_1) [Definition 6(ii)] together with Lemma 4 that, conditionally upon t_n,

$$\begin{aligned} R_1^4(x_1, M_n) &= R_1\{x_1, \hat{\pi}_1, \hat{\tau}_1(\hat{\pi}_1, M_n), M_n\} \\ &\leq R_1\{x_1, a\hat{\pi}_1, \hat{\tau}_1(a\hat{\pi}_1, M_n), M_n\}. \end{aligned} \tag{3.6}$$

In (3.6), R_1^4 refers to problem P^4, defined in terms of (MDP_1, Q_1). Further, if we regard $t_{n+1} - t_n$ as a stopping time on $\{(\mathrm{MDP}_i, Q_i), 2 \leq i \leq N\}$ with initial state $X^1(t_n)$ which achieves the Nash index M_n, then it is clear from Lemma 2(ii) that

$$t_{n+1} - t_n \in T_{N^1(\hat{\boldsymbol{\pi}})}\{X^1(t_n)\}. \tag{3.7}$$

It follows simply that

$$E_n\left[\alpha^{t_n}\prod_{j\neq 1}Q_j\{X_j(t_n)\} - \alpha^{t_{n+1}}\prod_{j\neq 1}Q_j\{X_j(t_{n+1})\}\right] \leq 0. \tag{3.8}$$

CASE 2. $M_n > 0$ or $M_n = 0^+$. For this case, the dominating nature of $\hat{\pi}_1$ leads via Definition 6(i) and Lemma 4 to the conclusion that, conditionally upon t_n,

$$\begin{aligned} R_1^3(x_1, M_n) &= R_1\{x_1, \hat{\pi}_1, \hat{\tau}_1(\hat{\pi}_1, M_n), M_n\} \\ &\geq R_1\{x_1, a\hat{\pi}_1, \hat{\tau}_1(a\hat{\pi}_1, M_n), M_n\}. \end{aligned} \tag{3.9}$$

In (3.9), R_1^3 refers to problem P^3, defined in terms of (MDP_1, Q_1). Also, because the Nash index M_n is positive, we conclude that

$$T_{N^1(\hat{\pi})}\{X^1(t_n)\} = \varnothing \tag{3.10}$$

and hence that

$$E_n\left[\alpha^{t_n}\prod_{j\neq 1}Q_j\{X_j(t_n)\} - \alpha^{t_{n+1}}\prod_{j\neq 1}Q_j\{X_j(t_{n+1})\}\right] \geq 0. \tag{3.11}$$

It is now an immediate consequence of (3.5) (together with its analogue for $a\hat{\pi}_1$) and (3.6), (3.8), (3.9) and (3.11) that

$$\mathscr{R}\{x, N(\hat{\boldsymbol{\pi}})\} \geq \mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\}, \tag{3.12}$$

as required.

We now remove the condition $N_1(a\hat{\pi}_1, x_1) \succcurlyeq N_j(x_j)$, $j \neq 1$, as follows: According to the policy $aN(\hat{\boldsymbol{\pi}})$, actions are chosen for the respective MDPs according to policies $a\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N$. Abbreviate the notation $N(a\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N)$ for the corresponding Nash index policy to $N(a\hat{\boldsymbol{\pi}})$. From Corollary 6,

$$\mathscr{R}\{x, N(a\hat{\boldsymbol{\pi}})\} \geq \mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\}. \tag{3.13}$$

Now consider the family $\{(\mathrm{MDP}_i, Q_i),\ 1 \leq i \leq N\}$ under policy $N(a\hat{\boldsymbol{\pi}})$. Let τ_a denote the first decision epoch at which $N(a\hat{\boldsymbol{\pi}})$ chooses action a for (MDP_1, Q_1). If no such time exists, $\tau_a = \infty$. In an obvious notation,

$$\begin{aligned} &\mathscr{R}\{x, N(a\,\hat{\boldsymbol{\pi}})\} \\ &\quad = \mathscr{R}\{x, N(a\hat{\boldsymbol{\pi}}), [0, \tau_a)\} + \mathscr{R}\{x, N(a\hat{\boldsymbol{\pi}}), [\tau_a, \infty)\} \\ &\quad = \mathscr{R}\{x, N(a\hat{\boldsymbol{\pi}}), [0, \tau_a)\} + \mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}}), [\pi_a, \infty)\}. \end{aligned} \tag{3.14}$$

Equation (3.14) signifies that from $\tau_a + 1$ onward, policy $N(\hat{\boldsymbol{\pi}})$ is followed. Now, by the characterisation of $N(a\hat{\boldsymbol{\pi}})$ and τ_a we must have

$$N_1\{a\hat{\pi}_1, X_1(\tau_a)\} \succcurlyeq N_j\{X_j(\tau_a)\}, \qquad j \neq 1, \text{a.s.}$$

Hence from the foregoing argument up to (3.12) we may conclude that

$$\mathscr{R}\{x, N(\hat{\boldsymbol{\pi}}), [\pi_a, \infty)\} \geq \mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}}), [\tau_a, \infty)\}, \tag{3.15}$$

namely, that we improve $N(a\hat{\boldsymbol{\pi}})$ by choosing actions according to $N(\hat{\boldsymbol{\pi}})$ from τ_a onward. However, such a policy always chooses actions for each (MDP_i, Q_i)

according to its dominating policy $\hat{\pi}_i$. By Corollary 6, such a policy can be no better than $N(\hat{\boldsymbol{\pi}})$. Hence from (3.13)–(3.15) we deduce that

$$\mathscr{R}\{x, N(\hat{\boldsymbol{\pi}})\} \geq \mathscr{R}\{x, N(a\hat{\boldsymbol{\pi}})\} \geq \mathscr{R}\{x, aN(\hat{\boldsymbol{\pi}})\},$$

as required. This concludes the proof. □

We have in Theorem 7 an appropriate generalisation of Whittle's (1980) result for families of competing MDPs in the absence of influence.

4. Families of stoppable bandit processes with influence. As an illustration of the material in Section 3, we discuss one of the simplest classes of families of competing MDPs with influence of practical interest. In doing so, we generalise results for families of stoppable bandit processes due to Glazebrook (1979, 1982) and discussed by Gittins (1989). Such processes (i.e., without influence) had been used by Bergman (1981) to model his buyer's problem. In these families, the objects of choice are decision processes with two actions available in each state, one of which (the "stopping" action) results in no change of state. To be explicit, by a *family of N stoppable bandit processes with influence* we mean a family of competing MDPs with influence $\{(\text{MDP}_i, Q_i),\ 1 \leq i \leq N\}$, as in features 1–5 of Section 3 with the following special features:

1′. $|A_j(x_j)| = 2$, $x_j \in \Omega_j$, $1 \leq j \leq N$. We write $A_j(x_j) = \{a_j, b_j\}$ and call action a_j "continue j" and action b_j "stop and invest in j."

2′. $P_j(x_j|x_j, b_j) = 1$, $x_j \in \Omega_j$, $1 \leq j \leq N$. Hence the application of the stop action b_j leaves the state of j unchanged.

3′. We write, for any $x_j \in \Omega_j$, $1 \leq j \leq N$,

$$R_j(x_j, a) = \begin{cases} R_j(x_j), & a = a_j, \\ \mu_j(x_j)(1-\alpha) > 0, & a = b_j. \end{cases} \tag{4.1}$$

As Gittins [(1989), page 63] points out, the preceding process could be a suitable model for an industrial research project. Each (MDP_j, Q_j) would model one of N possible routes to success in the project. The stopping action b_j corresponds to a decision to stop research and exploit the information gained in route j.

Note that should a stationary policy choose action b_j at time t in state $X(t)$ (and hence, in light of feature 2′, choose b_j at all subsequent decision epochs), the total reward earned during $[t, \infty)$ is

$$\alpha^t \left[\prod_{i \neq j} Q_i\{X_i(t)\}\right] \mu_j\{X_j(t)\}. \tag{4.2}$$

Interpret the quantity in (4.2) as a return from investing in j, influenced by the current states of the non-j options. Usage of "optimal" in Definition 8 is as in Definition 6.

DEFINITION 8. The *investment set* $S_j \subseteq \Omega_j$ for MDP_j is the set of states for which action b_j is uniquely optimal.

The following definition represents an appropriate development of Condition G described by Glazebrook (1982) to accommodate influence.

DEFINITION 9. The stoppable bandit process with influence (MDP_j, Q_j), described in features 1′–3′, satisfies Condition G′ if and only if

$$\mu_j(x_j) = Q_j(x_j)N_j(x_j), \qquad x_j \in S_j. \tag{4.3}$$

For stoppable bandit processes (i.e., without influence), the corresponding (simpler) Condition G has been commented upon and further analysed by Gittins (1989), Glazebrook (1979) and Glazebrook and Fay (1990). Our key result is the following theorem.

THEOREM 8. *For any stoppable bandit process with influence, Condition* G′ ⇔ *Condition* D. *There exists a dominating optimal policy* $\hat{\pi}_j$ *for* (MDP_j, Q_j) *satisfying Condition* G′ *such that*

$$\hat{\pi}_j(x_j) = \begin{cases} a_j, & x_j \notin S_j, \\ b_j, & x_j \in S_j. \end{cases}$$

PROOF. (i) *Condition G′ ⇒ Condition D.* Consider two cases in turn:

(a) $M \geq 0$. Problem $P^3(\cdot, M)$ for (MDP_j, Q_j) is solved via dynamic programming optimality equations:

$$\begin{aligned} R_j^3(x_j, M) = \max\Big(&MQ_j(x_j), \mu_j(x_j), R_j(x_j) \\ &+ \alpha E\big[R_j^3\{X_j(1), M | X_j(0) = x_j, a_j\}\big]\Big), \end{aligned} \tag{4.4}$$

in an obvious notation. From the characterisation of the Nash index (Definition 4), if $M \succcurlyeq N_j(x_j)$, then

$$R_j^3(x_j, M) = MQ_j(x_j) = \sup_{\tau} R_j(x_j, \pi_j, \tau, M)$$

for any policy π_j (and hence for $\hat{\pi}_j$).

Now consider the case $M \prec N_j(x_j) \Rightarrow M < N_j(x_j)$. Suppose that $\hat{\pi}_j(x_j) = b_j$ but that $\exists\, \pi_j, \tau$ such that

$$R_j(x_j, \pi_j, \tau, M) > \mu_j(x_j) = Q_j(x_j)N_j(x_j). \tag{4.5}$$

The equation in (4.5) follows from (4.3) and the characterisation of $\hat{\pi}_j$ in the statement of the theorem. However,

$$\begin{aligned} R_j(x_j, \pi_j, \tau, M) &= R_j(x_j, \pi_j, \tau, 0) + E_{\pi_j,\tau}\big[\alpha^{\tau} Q_j\{X_j(\tau)\} | X_j(0) = x_j\big]M \\ &\leq \Big(Q_j(x_j) - E_{\pi_j,\tau}\big[\alpha^{\tau} Q_j\{X_j(\tau)\} | X_j(0) = x_j\big]\Big)N_j(\pi_j, x_j) \\ &\quad + E_{\pi_j,\tau}\big[\alpha^{\tau} Q_j\{X_j(\tau)\} | X_j(0) = x_j\big]M \\ &\leq Q_j(x_j)\max\{N_j(\pi_j, x_j), M\} \leq Q_j(x_j)\max\{N_j(x_j), M\} \\ &= Q_j(x_j)N_j(x_j) = \mu_j(x_j). \end{aligned} \tag{4.6}$$

In the succession of equations and inequalities leading to (4.6), we utilise Lemma 1, Lemma 4 and Condition G′. However, in (4.5) and (4.6) we have now obtained a contradiction. Hence we conclude that under G′, if $M \geq 0$, then

$$M \prec N_j(x_j) \Rightarrow R_j^3(x_j, M) = \sup_{\tau} R_j(x_j, \hat{\pi}_j, \tau, M),$$

as required. This completes (a).

(b) $M \leq 0$. In this event, it is plainly never optimal to choose b_j in the dynamic program

$$R_j^4(x_j, M) = \min\Big(MQ_j(x_j), \mu_j(x_j), R_j(x_j) + \alpha E\big[R_j^4\{X_j(1), M | X_j(0) = x_j, a_j\}\big]\Big).$$

If $M \succcurlyeq N_j(x_j)$, an optimal choice is retirement, earning $MQ_j(x_j)$. If $N_j(x_j) \succ M$, an optimal policy for $P^4(x_j, M)$ for (MDP_j, Q_j) chooses a_j up to $\hat{\tau}(x_j, M)$ (see 2.16) and then retires. However, trivially from G′ and the characterisation of $\hat{\pi}_j$,

$$N_j(x_j) \succ M \Rightarrow N_j(x_j) < 0 \quad \text{or} \quad N_j(x_j) = 0^- \Rightarrow \hat{\pi}_j(x_j) = a_j.$$

Hence we conclude that under Condition G′, if $M < 0$, then

$$R_j^4(x_j, M) = \inf_{\tau} R_j(x_j, \hat{\pi}_j, \tau, M),$$

as required. This completes (b) and hence the proof of (i).

(ii) *Not Condition G′ ⇒ Not Condition D.* Not Condition G′ ⇒ $\exists\ x_j' \in S_j$ such that $Q_j(x_j')N_j(x_j') \succ \mu_j(x_j') > 0$. Consider two cases.

(a) $N_j(x_j') > 0$. Following Lemma 4(vi), $\exists\ \pi_j'$ with $\pi_j'(x_j') = a_j$ satisfying

$$Q_j(x_j')N_j(\pi_j', x_j') > \mu_j(x_j') > 0.$$

Following steps similar to those yielding (4.6), we can assert the existence of $M > 0$, satisfying $MQ_j(x_j') < \mu_j(x_j')$, such that for the stopping time τ' achieving $N_j(\pi_j', x_j')$,

$$\begin{aligned} R_j(x_j', \pi_j', \tau', M) &= R_j(x_j', \pi_j', \tau', 0) + E_{\pi_j', \tau'}\big[\alpha^{\tau'} Q_j\{X_j(\tau')\} | X_j(0) = x_j'\big] M \\ &= \Big(Q_j(x_j') - E_{\pi_j', \tau'}\big[\alpha^{\tau'} Q_j\{X_j(\tau')\} | X_j(0) = x_j'\big]\Big) N_j(\pi_j', x_j') \\ &\quad + E_{\pi_j', \tau'}\big[\alpha^{\tau'} Q_j\{X_j(\tau')\} | X_j(0) = x_j'\big] M > \mu_j(x_j'). \end{aligned} \tag{4.7}$$

Hence from (4.7), no optimal policy (i.e., choosing b_j in state x_j') can solve $P^3(M)$ for any such M. Therefore, no optimal policy dominates and Condition D cannot hold.

(b) $N_j(x_j') < 0$ or $N_j(x_j') = 0^-$. Following the discussion of (i)(b) it is clear that no optimal policy (i.e., choosing b_j in state x_j') can solve $P^4(M)$ for $M \prec N_j(x_j')$. Hence no optimal policy dominates and Condition D cannot hold. This concludes the proof of (ii) and of the theorem. □

The following is an immediate consequence of Theorems 7 and 8.

COROLLARY 9. *If each* (MDP_i, Q_i) *satisfies Condition* G′, *then* $N(\hat{\boldsymbol{\pi}})$ *is optimal for the family of stoppable bandit processes with influence* $\{(MDP_i, Q_i), 1 \le i \le N\}$, *where*

$$\hat{\pi}_i(x_i) = \begin{cases} a_i, & x_i \notin S_i, \\ b_i, & x_i \in S_i, 1 \le i \le N. \end{cases}$$

EXAMPLES. Consider now a family of N stoppable bandit processes with influence as a model of an industrial project in which (MDP_j, Q_j) models one of N possible routes to success. Each such route has two phases: an exploratory research phase and a development phase. This is modelled by writing

$$\Omega_j = \Omega_j^1 \cup \Omega_j^2, \qquad 1 \le j \le N,$$

where $X_j(0) \in \Omega_j^1$. So long as $X_j(t)$ remains within Ω_j^1, the research phase of route j is in progress. The development phase begins as soon as $X_j(t)$ enters Ω_j^2. Our assumptions about other aspects of this model reflect the ideas that influence is most likely to be an important issue during exploratory research and that investment returns are likely to be negligible until development starts, at which point they will begin to increase.

To reflect this, we allow Q_j, R_j and the transition law under a_j to be quite general during the research phase but require

$$\mu_j(x_j) = 0, \qquad x_j \in \Omega_j^1, \qquad 1 \le j \le N.$$

During the development phase, R_j and the transition law under a_j are again quite general, although note that regressing back to a state in Ω_j^1 is not allowed after entry to Ω_j^2. However, we shall assume that

$$P\left[Q_j\{X_j(t+1)\} = Q_j\{X_j(t)\} | X_j(t) = x_j, a_j\right] = 1, \qquad x \in \Omega_j^2, \qquad 1 \le j \le N,$$

$$P\left[\mu_j\{X_j(t+1)\} \ge \mu_j\{X_j(t)\} | X_j(t) = x_j, a_j\right] = 1, \qquad x \in \Omega_j^2, \qquad 1 \le j \le N.$$

It is not difficult to show that, under the foregoing assumptions, each (MDP_j, Q_j) satisfies Condition G′ and hence $N(\hat{\boldsymbol{\pi}})$ is optimal. The optimal policy $N(\hat{\boldsymbol{\pi}})$ will first give effort to routes in influential states (i.e., with negative indices) which have the potential to enhance the returns from other routes. Following this, it will then allocate effort on a reward rate basis, as measured by the Nash indices $N_j(\hat{\pi}_j, \cdot)$. As routes enter upon development, the investment options (b_j) will present an increasingly attractive reward rate alternative. The first route to enter its investment set under policy $N(\hat{\boldsymbol{\pi}})$ will be the one in which investment takes place.

REFERENCES

BERGMAN, S. W. (1981). Acceptance sampling: The buyer's problem. Ph.D. dissertation, Yale Univ.

BERGMAN, S. W. and GITTINS, J. C. (1985). *Statistical Methods for Planning Pharmaceutical Research*. Dekker, New York.

FAY, N. A. and GLAZEBROOK, K. D. (1987). On the scheduling of alternative stochastic jobs on a single machine. *Adv. in Appl. Probab.* **19** 955–973.

FAY, N. A. and GLAZEBROOK, K. D. (1989). A general model for the scheduling of alternative tasks on a single machine. *Probab. Engng. Inform. Sci.* **3** 199–221.

FAY, N. A. and WALRAND, J. C. (1991). On approximately optimal index strategies for generalised arm problems. *J. Appl. Probab.* **28** 602–612.

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 148–177.

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

GLAZEBROOK, K. D. (1979). Stoppable families of alternative bandit processes. *J. Appl. Probab.* **16** 843–854.

GLAZEBROOK, K. D. (1982). On a sufficient condition for superprocesses due to Whittle. *J. Appl. Probab.* **19** 99–110.

GLAZEBROOK, K. D. (1988). On a reduction principle in dynamic programming. *Adv. in Appl. Probab.* **20** 836–851.

GLAZEBROOK, K. D. (1991). Competing Markov decision processes. *Ann. Oper. Res.* **29** 537–564.

GLAZEBROOK, K. D. and FAY, N. A. (1990). Evaluating strategies for Markov decision processes in parallel. *Math. Oper. Res.* **15** 17–32.

GLAZEBROOK, K. D. and GREATRIX, S. (1993). On scheduling influential stochastic tasks on a single machine. *European J. Oper. Res.* To appear.

GLAZEBROOK, K. D. and OWEN, R. W. (1991). New results for generalised bandit processes. *Internat. J. Systems Sci.* **22** 479–494.

NASH, P. (1980). A generalised bandit problem. *J. Roy. Statist. Soc. Ser. B* **42** 165–169.

ROSS, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

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

VARAIYA, P., WALRAND, J. C. and BUYUKKOC, C. (1985). Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Automat. Control* **AC-30** 426–439.

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

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF NEWCASTLE UPON TYNE
NEWCASTLE UPON TYNE NE1 7RU
UNITED KINGDOM