The Annals of Applied Probability
1992, Vol. 2, No. 4, 1024–1033

ON THE GITTINS INDEX FOR MULTIARMED BANDITS

BY RICHARD WEBER

University of Cambridge

This paper considers the multiarmed bandit problem and presents a new proof of the optimality of the Gittins index policy. The proof is intuitive and does not require an interchange argument. The insight it affords is used to give a streamlined summary of previous research and to prove a new result: The optimal value function is a submodular set function of the available projects.

1. The multiarmed bandit problem. The multiarmed bandit problem is concerned with the question of how to dynamically allocate a single resource amongst several alternative projects. It models problems that arise in many contexts, for example, the scheduling of jobs on a single machine and the design of sequential clinical trials.

Consider a gambler who is presented with the opportunity to play any of n one-armed bandit machines. He wishes to allocate his successive plays amongst these machines to maximize his expected total-discounted reward. He does this one play at a time, on the basis of prior information and observations to date. More formally, consider a family of n alternative bandit processes (FABP), defined as a discounted Markov decision process in which actions $(a_1, \ldots, a_n)$ are available at decision epochs $0, 1, \ldots$. The states of the bandits are $x_1(t), \ldots, x_n(t)$. Taking action a_j corresponds to "playing bandit j." The result is that a random reward $R_j(x_j(t))$ is obtained, the state of bandit j changes in a known Markov fashion and the states of all other bandits are unchanged. Assume rewards are nonnegative and uniformly bounded. Let $j(t)$ denote the bandit that is played at epoch t when the evolution of the process is determined by policy π. The aim is to find a policy having the greatest value of expected total-discounted reward,

$$V_\pi(x) = E_\pi\left[\sum_{t=0}^{\infty} \beta^t R_{j(t)}\big(x_{j(t)}(t)\big) \,\Big|\, x(0) = x\right], \tag{1}$$

where $0 < \beta < 1$.

It is well known that the solution to the problem can be characterised by functions G_j, having the property that playing bandit j is optimal at t if and only if

$$G_j\big(x_j(t)\big) = \max_{1 \le i \le n} G_i\big(x_i(t)\big).$$

Received February 1992.
AMS 1980 *subject classifications*. Primary 60G40, 90B35; secondary 62L05, 90C40.
Key words and phrases. Multiarmed bandit problem, stochastic scheduling, Markov decision processes, optimal stopping, sequential methods.

These functions are called *Gittins indices*. The remarkable thing is that G_j depends only on information concerning bandit j. This greatly reduces the dimensionality of the problem and its solution. For example, if we were scheduling jobs on a single machine we would simply compute an index for each job and schedule the job of greatest index. The following theorem was proved by Gittins and Jones (1974).

THEOREM 1. *The optimal policy is to play at each epoch a bandit of greatest Gittins index.*

Section 2 of this paper presents a simple proof of Theorem 1. It is intended to be intuitive and to require almost no notation. Some special cases and generalisations are discussed in Sections 3–8. In Section 4 we relate our proof to that of Gittins and Jones and also to a proof due to Whittle (1980).

2. Optimality of Gittins index policy. Suppose bandit j is the only bandit and the gambler may either play it or not. As a device for the proof, we shall imagine that each time the gambler plays bandit j he pays a fixed charge. This charge, which we shall call the *prevailing charge*, is the same each time he plays bandit j. The gambler is to play the bandit for some number of epochs, observing the state as it evolves, and stop when it is unfavorable to continue playing (because the prevailing charge is too great). If the prevailing charge is sufficiently small, it will be profitable to play the bandit at least once more (in the sense of expected total-discounted profit). If the charge is too great then any further play of bandit j will be loss-making. As a function of the state of bandit j, say x_j, we define the *fair charge* $\gamma_j(x_j)$ as the level of prevailing charge for which optimal play would be neither profitable nor loss-making. Thus,

$$(2)\quad \gamma_j(x_j) = \sup\left\{\gamma\colon \sup_{\pi} E_{\pi}\left[\sum_{t=0}^{\tau-1}\beta^t\big(R_j\big(x_j(t)\big)-\gamma\big)\big|x_j(0)=x_j\right]\geq 0\right\},$$

where the policy π determines a stopping time $\tau \geq 1$.

Now imagine that the prevailing charge for bandit j is reduced to the level of the fair charge whenever the gambler would otherwise stop playing, that is, whenever the value of the fair charge in the present state falls below the value of prevailing charge. Equation (2) expresses the fact that if the gambler starts to play when the prevailing charge is equal to the fair charge and adopts a policy of stopping at the first time it is not optimal to continue, then his expected total-discounted profit is zero. Given that the value of the prevailing charge is reduced at precisely the time he would otherwise have stopped, the gambler may continue to play the bandit as a fair game for further epochs. If the prevailing charge is always reduced in this manner, then the gambler need never stop; if he plays bandit j forever he experiences a fair game (in terms of expected total-discounted profit).

PROOF OF THEOREM 1. Thus far we have considered only a single bandit. Now consider a gambler who at each epoch must play one of n alternative bandits. Suppose that initially the prevailing charge for each bandit is set equal to its fair charge and then the prevailing charges are reduced periodically in the manner described above. Then the gambler never pays more than the fair charge to play any bandit and he can ensure his expected total-discounted profit is nonnegative (e.g., by playing the same bandit at all times). However, it is also clear that his expected total-discounted profit cannot be positive, since it would have to be positive for at least one bandit taken alone and this is disallowed because the prevailing charges were set equal to the fair charges. Notice that if at any time the fair charge for bandit j has risen above its prevailing charge, then playing bandit j is strictly profitable and the effect of discounting will be to strictly lessen the gambler's potential expected total-discounted profit from bandit j if he does not immediately continue playing bandit j. Since whenever the prevailing charge was last reset, its value was chosen so that the gambler will only just break even if he plays bandit j optimally, any policy that fails to continue immediately to play bandit j in such a circumstance must incur an expected total-discounted loss. Conversely, a policy is optimal provided it never fails to immediately continue play of a bandit in such a circumstance. These ideas are summarised in two remarks.

REMARK 1. Each bandit of itself presents the gambler with an opportunity for a fair game, but only if he plays it optimally. If he plays bandit j suboptimally, then his expected total-discounted reward from bandit j will be less than the expected total-discounted charge that he pays to bandit j.

REMARK 2. The gambler plays optimally provided that whenever he starts playing a given bandit he continues to do so without interruption as long as that bandit's fair charge remains greater than its prevailing charge.

The following key points are the heart of this proof. Notice that the sequence of prevailing charges for each bandit is a nonincreasing function of the number of plays the bandit has received. The function is random; its values only become known as the state of the bandit evolves. However, by definition, it is a sequence that is independent of the policy adopted. It is enough to know that for each bandit the sequence of prevailing charges is nonincreasing to see that if the gambler adopts the policy of always playing a bandit of greatest prevailing charge, then he incurs the charges in a nonincreasing sequence. This interleaving of the charges into a single nonincreasing sequence is unique (in terms of charges, not the bandits). Thus the policy of playing the bandit of greatest prevailing charge (or equivalently, of greatest fair charge) maximizes the expected total-discounted charge paid by the gambler. Now by the first remark, this quantity is an upper bound for the expected total-discounted reward obtained by the gambler under any policy. By the

second remark, this bound is achieved by the proposed policy, since it ensures that the gambler continues to play a bandit without interruption for any series of epochs that its fair charge exceeds its prevailing charge. Thus we have a fair game in which the gambler's expected total-discounted profit is 0. Simply define the Gittins index of bandit j as its fair charge divided by $(1 - \beta)$,

$$G_j(x_j) = \gamma_j(x_j)/(1 - \beta). \tag{3}$$

The divisor $(1 - \beta)$ is introduced for consistency with other expositions. Since we have seen that it is optimal to always play a bandit with the greatest fair charge, this proves the theorem. □

We consider a few connections, special cases and generalisations.

3. A formula for the index. From (2) we have

$$\gamma_j(x_j) = \sup_{\pi} \frac{E_\pi\left[\sum_{t=0}^{\tau-1}\beta^t R_j(x_j(t))|x_j(0) = x_j\right]}{E_\pi\left[\sum_{t=0}^{\tau-1}\beta^t|x_j(0) = x_j\right]}. \tag{4}$$

Thus $\gamma_j(x_j)$ can be interpreted as the maximal value of the ratio of expected discounted reward to expected discounted time under policies that choose a stopping time $\tau \geq 1$.

For the original model in Section 1, with no prevailing charges, (2) and (3) imply that $G_j(x_j)$ can be interpreted as the value of a lump sum retirement payment such that the gambler would be indifferent between receipt of the lump sum or retirement and receipt of the lump sum after some optimal number of further plays.

4. Connection to other proofs. The characterisation of the Gittins index as the solution to a stopping problem in (4) is not new; it was at the heart of Gittins and Jones' original proof. Their proof used an interchange argument, that in a more general setting has been well explained by Varaiya, Walrand and Buyukkoc (1985). The proof is based on considering a suboptimal policy, that at epoch 0 plays a bandit j that is of less than greatest index. Such a suboptimal policy can be improved by exchanging that play of bandit j with the first play, made at some later random time, of the bandit that actually has the greatest Gittins index at time 0. This leads to a fairly complicated reshuffling of the order in which the bandits are played, but the effect is tractable. A stopping problem is key to the analysis. The argument can be repeated for epochs $1, \ldots, t - 1$, at which point the improved policy is identical to the Gittins index policy for the first t epochs. Because of discounting, the rewards obtained from epoch t onwards result in a vanishingly small amount of suboptimality as t is increased toward infinity.

Whittle's proof is different, but also based on a stopping problem. He introduces the idea of a retirement option and imagines that the gambler may at any time decide to stop playing and collect a retirement reward M. He

expresses the value function for this modified problem in terms of the value functions for n similarly modified single-armed bandit problems.

To understand the connection between Whittle's formulation and the proof in this paper, we can imagine there is a separate retirement option for each bandit. However, the retirement payment is received as a pension of $(1 - \beta)M_j$ per period. Suppose that during each period the gambler may "unretire" with respect to exactly one bandit, but in doing so he forgoes for that period the pension associated with the bandit. Suppose the pension for bandit j is reduced to a "fair value" whenever the gambler would otherwise have decided that the value of the pension was sufficiently attractive to retire from bandit j permanently. As before, it is clear that the Gittins index policy maximizes the expected total-discounted pension that the gambler forgoes, since it interleaves nonincreasing sequences into a nonincreasing order, and this provides an upper bound on the expected total-discounted reward he obtains under any policy. But under the Gittins index policy the gambler plays optimally and forgoes pension payments that have the same expected total-discounted value as the reward he obtains.

5. Suboptimality bounds. It can be interesting to compare the Gittins index policy with others that might be employed. Our formulation in terms of prevailing and fair charges makes it particularly easy to derive some bounds. Recall, by Remark 2 in the proof of Theorem 1, that the gambler's play is optimal provided that whenever he starts play on a bandit, he continues to do so without interruption as long as its fair charge is greater than its prevailing charge. This notion is summarised in the following definition.

DEFINITION. A policy is said to be *index consistent* if once play of a given bandit commences, play of that bandit continues without interruption while its Gittins index remains greater than its initial value.

Remark 1 in the proof of Theorem 1 showed that for any policy π, the expected total-discounted reward in (1) is bounded above by the expected total of the discounted charges that are incurred. The prevailing charge for bandit j at time t is the least fair charge so far, namely, $\min_{0 \le s \le t}\{\gamma_j(x_j(s))\}$. So

$$V_\pi(x) \le E_\pi\left[\sum_{t=0}^{\infty} \beta^t \min_{0 \le s \le t} \left\{\gamma_{j(t)}\left(x_{j(t)}(s)\right)\right\} \middle| x(0) = x\right]. \tag{5}$$

Consider a fixed realisation of the sequence of states through which each bandit evolves. It is clear that if one follows a policy that differs from the Gittins policy, then the value of the greatest prevailing charge, $\max_i \min_{0 \le s \le t} \gamma_i\{x_i(s)\}$, is at least as great as it would have been under the Gittins policy. Combining this observation with (5) gives the following bound, previously given by Glazebrook (1990).

THEOREM 2. *Suppose π^* is the optimal (Gittins) policy. The suboptimality of a policy π has the bound*

$$V_{\pi^*}(x) - V_\pi(x) \le (1-\beta)E_\pi\left[\sum_{t=0}^{\infty}\beta^t\left(\max_i \min_{0\le s\le t} G_i(x_i(s)) - \frac{R_{j(t)}(x_{j(t)}(t))}{(1-\beta)}\right)\middle| x(0)=x\right]. \tag{6}$$

Glazebrook (1982) has given another bound. Its bound is usually weaker than (6), but has the advantage of being expressed entirely in terms of Gittins indices.

THEOREM 3.

$$V_{\pi^*}(x) - V_\pi(x) \le E_\pi\left[\sum_{t=0}^{\infty}\beta^t\left(\max_i G_i(x_i(t)) - G_{j(t)}(x_{j(t)}(t))\right)\middle| x(0)=x\right]. \tag{7}$$

PROOF. Let π be an arbitrary policy and define $\pi^*(t)$ and $\pi(t)$ as policies that are identical to π for times less than t. Thereafter, $\pi^*(t)$ is identical to the Gittins policy; $\pi(t)$ plays $j(t)$ at time t, continues playing $j(t)$ until its fair charge drops below $\gamma_{j(t)}(x_{j(t)})$ and is identical to the Gittins index policy thereafter. Since $\pi(t)$ and $\pi^*(t)$ are the same for times less than t, they incur the same prevailing charges until t and will have reached the same state at t. A little thought reveals that for each time $s \ge t$, the prevailing charge incurred under $\pi^*(t)$ minus the charge incurred under $\pi(t)$ is never more than the value of this difference at time t, namely, $\max_i \gamma_i(x_i(t)) - \gamma_{j(t)}(x_{j(t)}(t))$. This is because the first time that all prevailing charges are less than $\gamma_{j(t)}(x_{j(t)}(t))$ is the same under $\pi^*(t)$ and $\pi(t)$ and at that time they will have reached the same state. The only charges that might possibly be taken in different orders prior to this time are those with values in the interval $[\gamma_{j(t)}(x_{j(t)}(t)), \max_i \gamma_i(x_i(t))]$. From time t onward, both policies are index consistent, so the expected total-discounted charge equals the expected total-discounted reward. Thus averaging over realisations,

$$V_{\pi^*(t)}(x) - V_{\pi(t)}(x) \le \beta^t E_\pi\left[\frac{\max_i \gamma_i(x_i(t)) - \gamma_{j(t)}(x_{j(t)}(t))}{1-\beta}\middle| x(0)=x\right]. \tag{8}$$

Note that since $\pi(t)$ and $\pi^*(t+1)$ are identical through time t and $\pi^*(t+1)$ is optimal thereafter, $V_{\pi(t)}(x) \le V_{\pi^*(t+1)}(x)$. Also, $\pi^*(0)$ is the Gittins policy. Using these facts and summing (8) from $t = 0$ to infinity gives (7). □

There is a further interesting inequality that can be derived using the ideas in Section 2. Suppose S denotes the set of all bandits, $\{1, \ldots, n\}$. Then for any $I \subseteq S$, let $P(I)$ denote the restriction of the problem to bandits in I and now let $V(I)$ denote the maximal expected total-discounted reward [suppressing the

dependence on $x(0)$, which we assume to be fixed]. The following theorem says that V is a submodular set function.

THEOREM 4. *For any* $I, J \subseteq S$,

$$V(I \cap J) + V(I \cup J) \leq V(I) + V(J). \tag{9}$$

PROOF. Consider a fixed realisation of the states through which bandit j evolves. Let γ_{jk} denote the prevailing charge of bandit j on this realisation after it has been played k times, $\gamma_{j0} \geq \gamma_{j1} \geq \cdots$. Suppose $U_t(I)$ is the sum of undiscounted charges paid while playing $P(I)$ optimally during the first t plays. It is computed by interleaving into nonincreasing order the sequences $\{\gamma_{jk}\}_{k=1}^{\infty}$, $j \in I$, and then summing the first t elements of the resulting sequence. It is not hard to see that for all $t \geq 0$,

$$U_t(I \cap J) + U_t(I \cup J) \leq U_t(I) + U_t(J). \tag{10}$$

The theorem follows by multiplying (10) by β^t, summing on t from 0 to infinity and taking an expected value over realisations. □

If all bandits are statistically the same and initially in the same state, then a consequence of (9) is that $V(S)$ is a concave increasing function of n. (Simply take $I = \{1, \ldots, n-1\}$ and $J = \{2, \ldots, n\}$.) A special case of (9) has been proved by Tsitsiklis (1986). He supposes that there is one bandit, say bandit 1, that always pays $(1-\beta)M$ per play; so choosing to play bandit 1 corresponds to taking a retirement option. By considering $I = \{1, i\}$ and $J = S \setminus \{i\}$, we have the bound $V(S) \leq V(\{i, 1) + V(S \setminus \{i\}) - M$.

6. The finite horizon scenario. In general, the Gittins index policy does not maximize the expected total-discounted reward obtained by a finite time t, nor does any other index policy (except in the trivial case $t = 1$, when a one-step look-ahead policy is optimal). However, an important exception is the *deteriorating case*, in which with probability 1 the fair charge is nonincreasing. In this case the prevailing charge is at all times equal to the fair charge, so we have $\gamma_j(x_j) = E[R_j(x_j)]$, and the Gittins policy is the one-step look-ahead policy. The deteriorating case requires that $R_j(x_j(t))$ be nonincreasing with probability 1.

For the deteriorating case it is clear that the proof still holds if the objective function is to minimize the expected total-cost up to time t. Moreover, if rewards are monotone likelihood ratio ordered and nonincreasing in that ordering, then the Gittins index policy stochastically maximizes for all t the total reward obtained by time t. An example is where R_j is 0 or 1 with probabilities $1 - r_j$ and r_j, and r_j is nonincreasing with probability 1 following each play of bandit j.

7. The nonpreemptive scenario. In previous sections we have assumed that the gambler plays preemptively: In other words, at each epoch he has a free choice amongst bandits. One can impose a constraint on preemption, by

saying that once the gambler starts to play a particular bandit he must do so until its state reaches a given set in which preemption is allowed. Essentially, the same arguments apply, but the calculation of the appropriate prevailing charge takes into account the fact that having once embarked upon play of bandit j, the gambler may only discontinue play when its state reaches a certain set. The formula for the index takes the same form as (2), but the stopping time τ is restricted to times at which the bandit's state enters that set.

The *climactic case* is one in which the optimal policy is naturally nonpreemptive. It refers to the circumstances in which with probability 1 the fair charge is nondecreasing, up to some random time that the bandit enters an absorbing state 0, in which the fair charge is 0. Whittle calls state 0 the *lapsed state* of a project. It is a state in which no further rewards can be obtained. Then τ is interpreted as the time at which the bandit first enters the lapsed state. The optimal policy is to choose the bandit of greatest index and play it until it lapses. The bandits are played nonpreemptively until all have reached their lapsed state. An example is the scheduling jobs of stochastic processing times on a single machine. Suppose job j has an increasing hazard rate and makes only the one payment R_j only upon completion (i.e., entry to the lapsed state). The index is then

$$G_j = \frac{R_j E\left[\beta^{X_j}\right]}{1 - E\left[\beta^{X_{j+1}}\right]}.$$

In this case it is meaningful to think of an undiscounted problem in which a holding cost R_j is made for each epoch that job j is not yet complete. Then $(1-\beta)G_j \to R_j/E[X_j]$ as $\beta \to 1$, and it can be shown that the policy based on these indices minimizes the expected holding cost incurred until all jobs are complete.

8. Branching bandits. It is possible to extend the results to a model in which additional bandits become available at times after the start. Let bandits be of L types. Suppose that once having begun to play a bandit of type j the gambler is commited to T_j plays of that bandit, during which rewards are obtained and at the end of which $N_{j1}, \ldots, N_{jL}$ bandits of types $1, \ldots, L$, are obtained. Thus bandits arise as a branching process. Here T_j and the N_{jk} are random variables, not necessarily independent, but they are independent from bandit to bandit and identically distributed for bandits of the same types.

Again, we imagine that j is the only bandit present initially and define the fair charge, γ_j, as the greatest cost per play that the gambler would be willing to pay to play bandit j for the random number of plays T_j, then make further numbers of plays of some or all of the bandits that branch from it or its descendants. We define the prevailing charge of any bandit that is not present initially as the minimum of its own fair charge and the prevailing charge that applied to its parent at the time of branching. Under this definition, the sequence of prevailing charges for a given bandit and its descendants is again

nonincreasing and independent of the policy employed. The sequences of charges are uniquely interleaved into a single nonincreasing sequence if the gambler always plays the bandit of greatest index. Thus playing the bandit of greatest index maximizes the expected total discounted sum of charges. Because they are fair charges, their expected total discounted sum is an upper bound for the expected total discounted reward that the gambler can obtain. He achieves this upper bound if he plays optimally, which means playing optimally within each branch of the process.

Thus we have seen that there is a Gittins index that can be computed for each type of bandit and it can be interpreted as the bandit's fair charge when it is the only bandit available at the start. Of course the value of this index will depend on the statistics of the other types of bandit that can arise amongst its descendants. But in fact, the dependency is only on bandits of greater Gittins index, since to play optimally within a branch that begins with a bandit of type j means not playing any descendant having a lesser index value and playing those that arise and are of greater index value according to the priority of Gittins indices.

The idea of a branching bandit process is based on the work of Weiss (1988). The process that arises through branching of each of the initial bandits has also been called a *superprocess*. We have seen that one ought to choose between superprocesses according to their Gittins indices and play optimally within each. In particular, this model can be used to analyse the optimal nonpreemptive scheduling of a $M/GI/1$ queue. See Weiss (1988) and Whittle (1983) for more details.

9. Conclusion. We have provided a short and intuitive proof of the optimality of the Gittins index policy and summarised in a unified manner some of its more immediate properties and generalisations. The subject is a rich one and has been explored by others elsewhere. The reader is recommended to Gittins (1989) for an authoritative treatment and exposition of various areas of application. I am grateful to Gideon Weiss for reading a first draft of this paper and suggesting the interpretation of Whittle's formulation in terms of a pension and to Kevin Glazebrook for drawing my attention to the paper of Tsitsiklis.

REFERENCES

GITTINS, J. C. (1989). *Bandit Processes and Dynamic Allocation Indices*. Wiley, New York.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (J. Gani, ed.) 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.

TSITSIKLIS, J. N. (1986). A lemma on the multiarmed bandit problem. *IEEE Trans. Automat. Control* **31** 576–577.

VARAIYA, P., WALRAND, J. and BUYUKKOC, C. (1985). Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Automat. Control* **30** 426–439.
WEISS, G. (1988). Branching bandit processes. *Probab. Engr. Inform. Sci.* **2** 269–278.
WHITTLE, P. (1980). Multiarmed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42** 143–149.
WHITTLE, P. (1983). *Optimization Over Time* **1**. Wiley, Chichester.

JUDGE INSTITUTE OF MANAGEMENT STUDIES
ENGINEERING DEPARTMENT
UNIVERSITY OF CAMBRIDGE
MILL LANE
CAMBRIDGE CB2 1RX
ENGLAND