February 1993 LIDS-P-2171

A SHORT PROOF OF THE GITTINS INDEX THEOREM [1]

John N. Tsitsiklis[2]

Abstract

We provide a short and elementary proof of the Gittins index theorem for the multi-armed bandit problem, for the case where each bandit is modeled as a finite-state semi-Markov process. We also indicate how this proof can be extended to the branching bandits problem.

Submitted to: Annals of Applied Probability

[1] Research supported by the ARO under grant DAAL03-92-G-0115 and by Siemens A. G.

[2] Room 35–214, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. e-mail: jnt@athena.mit.edu

I. INTRODUCTION

There is a long history of alternative proofs of the Gittins index theorem for the multi–armed bandit problem. The original proof of Gittins and Jones [GiJ74] relied on an interchange argument. A different interchange argument was provided by Varaiya et al. [VWB85] and was simplified further in [Wal88]. The simplest interchange argument available seems to be the one by Weiss [Wei88] which in fact establishes an index theorem for the more general branching bandits model. A different proof, based on dynamic programming, was provided by Whittle [Whi80] and subsequently simplified by Tsitsiklis [Tsi86]. Weber [Web92] has outlined a new proof that avoids any calculations and rests on more qualitative reasoning. Finally, a proof based on a polyhedral characterization of a suitably defined "performance region" has been provided by Tsoucas [Tso91], for the case of the average–cost Klimov problem, and by Bertsimas and Nino–Mora [BeN93] for many other classes of multi–armed bandit problems.

This paper presents yet another proof of the same result. This proof has some common elements with the proof in [Wei88] but appears to be simpler in that it is based on a simple inductive argument and uses only trivial calculations. The induction is in terms of the cardinality of the state spaces of the bandits involved and, for this reason, the proof is valid only for the case of finite–state bandits.

The rest of the paper is organized as follows. Section 2 presents the model to be employed and the proof of our main result. Section 3 contains some discussion on how to handle the more general branching bandits problem.

II. THE MULTI–ARMED BANDIT MODEL

There are n bandit processes. The ith such process is a semi–Markov process with a finite state space $\mathcal{X}_i$. We assume, for simplicity, that the state spaces of the different bandits are disjoint and we let $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_n$. If the ith bandit is at some state $x \in \mathcal{X}_i$ and is selected to be "played", then a random reward $R(x)$ is received and the bandit remains active over a time period of random length $T(x)$. After $T(x)$ time units, the play is completed and the bandit moves to a random new state y. At that point, we are free to choose the same or another bandit to be played.

We assume that the joint probability distribution of the random vector $(T(x), R(x), y)$ is known and is the same for every play of bandit i for which bandit i is at the same state $x \in \mathcal{X}_i$. In addition, the random vectors corresponding to different plays of the same or of different bandits are assumed to be statistically independent.

A *policy* for the multi–armed bandit problem is defined as a mapping $\pi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \mapsto \{1, \ldots, n\}$ which at time zero and at any play completion time chooses a bandit to be played next, as a function of the current states of the n bandits. Given a particular policy, the time t_i at which the ith play starts and the reward R_i received at that time are well-defined random variables. Let $\beta > 0$ be a discount rate. We are interested in the problem of finding a policy that maximizes the expected discounted reward

$$E\Big[\sum_{i=1}^{\infty} R_i e^{-\beta t_i}\Big]$$

for every initial state.

We now comment on some consequences of the general fact that it is only $E[R(x)]$ and not the entire probability distribution of $R(x)$ that matters. Let us fix a particular policy. Let x_i be the state of the bandit that is played at the ith play and let $\mathcal{F}_i$ stand for all random variables realized during the first $i - 1$ plays. In particular, t_i and x_i are determined by the outcomes of the first $i - 1$ plays and are contained in $\mathcal{F}_i$. Conditioned on $\mathcal{F}_i$, the expected discounted reward resulting from the ith play is given by

$$e^{-\beta t_i} E[R(x_i) \mid x_i].$$

Let us now consider the following alternative reward structure: whenever a bandit at some state x is played, then rewards are received throughout the duration of that play at a constant rate $r(x)$, where

$$r(x) = \frac{E[R(x)]}{E[\int_0^{T(x)} e^{-\beta t}\,dt]}. \tag{2.1}$$

Under this new reward structure, the expected reward resulting from the ith play, conditioned on $\mathcal{F}_i$, is given by

$$E\Big[\int_{t_i}^{t_i+T(x_i)} e^{-\beta t} r(x_i)\,dt \mid \mathcal{F}_i\Big] = e^{-\beta t_i} r(x_i) E\Big[\int_0^{T(x_i)} e^{-\beta t}\,dt \mid x_i\Big] = e^{-\beta t_i} E[R(x_i) \mid x_i],$$

where the last equality follows from (2.1). It follows that under either reward structure, the infinite–horizon expected discounted reward of any policy is the same. We will be using this fact later in our proof.

We say that a policy is a *priority rule* if there is an ordering of the elements of $\mathcal{X} = \mathcal{X}_1 \cup \cdots \mathcal{X}_n$ such that at each decision point, the bandit whose state is ordered highest is chosen. Our basic result is the following.

Theorem 2.1: If each $\mathcal{X}_i$, $i = 1, \ldots, n$, is finite, then there exists a priority rule which is optimal.

Proof: Let N be the cardinality of the set $\mathcal{X}$. The proof proceeds by induction on N.

If $N = 1$, we have a single bandit and the only available policy is trivially a priority rule.

Let us now assume that the result is true for all multi–armed bandit problems for which $N = K$, where K is some positive integer. We consider a multi–armed bandit problem for which $N = K + 1$, and we will show that there exists an optimal policy which is a priority rule. This will complete the induction and the proof of the theorem.

Let us pick some state $s^* \in \mathcal{X}$ such that $r(s^*) = \max_{x \in \mathcal{X}} r(x)$. Let i^* be such that $s^* \in \mathcal{X}_{i^*}$. The following lemma states that s^* can be chosen as a top priority state.

Lemma 2.1: There exists an optimal policy that obeys the following rule: whenever bandit i^* is at state s^*, then bandit i^* is played.

Proof: Consider an optimal policy π. Suppose that at time 0, bandit i^* is at state s^*. If policy π chooses to play bandit i^*, then there is nothing to prove. Suppose now that π chooses some other bandit to play. Define the random variable τ as the first time at which bandit i^* is played under policy π. (We let $\tau = \infty$ if bandit i^* is never played.)

We now define a new policy, call it π', which plays bandit i^* once and from then on mimics the actions of policy π. However, when (and if) policy π plays bandit i^* for the first time, policy π skips that play of bandit i^*. Let $\overline{r}(t)$ be the reward rate, as a function of time, under policy π. Using the definition of s^*, we have $\overline{r}(t) \leq r(s^*)$ for all t.

The expected discounted reward $J(\pi)$ under policy π is given by

$$J(\pi) = E\Big[\int_0^\tau \overline{r}(t)e^{-\beta t}\,dt + e^{-\beta\tau}\int_0^{T(s^*)} r(s^*)e^{-\beta t}\,dt + \int_{\tau+T(s^*)}^\infty \overline{r}(t)e^{-\beta t}\,dt\Big].$$

Similarly, the expected discounted reward $J(\pi')$ under policy π' is given by

$$J(\pi') = E\Big[\int_0^{T(s^*)} r(s^*)e^{-\beta t}\,dt + e^{-\beta T(s^*)}\int_0^\tau \overline{r}(t)e^{-\beta t}\,dt + \int_{\tau+T(s^*)}^\infty \overline{r}(t)e^{-\beta t}\,dt.\Big]$$

We wish to show that $J(\pi') \geq J(\pi)$. Equivalently, that

$$E\Big[(1 - e^{-\beta\tau})\int_0^{T(s^*)} r(s^*)e^{-\beta t}\,dt\Big] \geq E\Big[(1 - e^{-\beta T(s^*)})\int_0^\tau \overline{r}(t)e^{-\beta t}\,dt\Big]. \tag{2.2}$$

We note that if $\overline{r}(t)$ were equal to $r(s^*)$ for all t, then the two sides of Eq. (2.2) would be equal. This observation and the fact $\overline{r}(t) \leq r(s^*)$ show that Eq. (2.2) is valid and, therefore, $J(\pi') \geq J(\pi)$. Since π was assumed optimal, π' is also optimal. But if it is optimal to give top priority to state s^* at time 0, then (by the optimality of stationary policies) it is also optimal to give top priority to state s^* at every decision time. **q.e.d.**

Lemma 2.1 states that there exists an optimal policy within the set of policies that give top priority to state s^*; call this set of policies $\Pi(s^*)$. We will now consider the problem of finding a policy which is optimal within the set $\Pi^*(s)$.

If s^* is the only possible state of bandit i^*, then the policy that always plays bandit i^* is evidently optimal and is a priority rule. We henceforth assume that $\mathcal{X}_{i^*}$ is not a singleton. Suppose that bandit i^* is in some state $x \neq s^*$ and that this bandit is played. If this play causes a transition to state s^*, bandit i^* will be played again and again until eventually a transition to some state different from s^* results. We can view this succession of plays as a single (composite) play which cannot be interrupted due to our restriction to $\Pi(s^*)$. This single play has a random duration $\hat{T}(x)$ equal to the total time elapsed until a transition to a state different than s^*. Furthermore, by the discussion preceding the statement of Theorem 2.1, the reward of every policy remains the same if the discounted reward $\int_0^{\hat{T}(x)} e^{-\beta t}\overline{r}(t)\,dt$ received during this composite play is replaced by a constant reward rate equal to

$$\hat{r}(x) = \frac{E[\int_0^{\hat{T}(x)} e^{-\beta t}\overline{r}(t)\,dt]}{E[\int_0^{\hat{T}(x)} e^{-\beta t}\,dt]}. \tag{2.3}$$

to be received throughout the duration of this composite play. We may thus replace bandit i^* by a new bandit in which state s^* is absent, $T(x)$ and $r(x)$ are replaced by $\hat{T}(x)$ and $\hat{r}(x)$, respectively, and the transition probabilities are suitably modified. We call this procedure "reducing bandit i^* by removing state s^*."

The above argument shows that the problem of finding an optimal policy within the class $\Pi(s^*)$ is a new multi–armed bandit problem for which the sum of the cardinalities of the state spaces of the different bandits is equal to K. The induction hypothesis shows that there exists a priority rule $\hat{\pi}$ which is optimal for the latter problem. It follows that there exists an piority rule which is optimal for the original problem: give top priority to state s^* and follow the priority rule $\hat{\pi}$ for the remaining states. **Q.E.D.**

To every state $x \in \mathcal{X}$, we associate a number $\gamma(x)$, which we will call an *index*, using the following procedure:

Index Algorithm:

a) Pick a state s^* such that $r(s^*) = \max_{x \in \mathcal{X}} r(x)$ and let $\gamma(s^*) = r(s^*)$. Let i^* be such that $s^* \in \mathcal{X}_{i^*}$.

b) If $\mathcal{X}_{i^*}$ is a singleton, remove bandit i^*. Else, reduce bandit i^* by removing state s^*. Go back to (a).

From the proof of Theorem 2.1, it is apparent that the statistics of the random variables $\hat{T}(x)$ and $\hat{r}(x)$, as well as the transition probabilities of the reduced bandit i^* are completely determined by the corresponding statistics and transition probabilities of the original bandit i^*. This shows that the indices of the various states of a particular bandit are completely determined by the statistics associated with that bandit. In other words, the index algorithm can be carried out separately for each different bandit, still yielding the same index values.

The Gittins index theorem establishes something more than Theorem 2.1. In particular, not only does it show that there exists a priority policy which is optimal, but also that an optimal priority ordering can be found by ordering the states according to the numerical values of a certain index which can be computed separately for each bandit. We can also get this stronger result as follows:

Theorem 2.2: Let the index of each state be determined according to our index algorithm. Then, any priority policy in which states with a higher index have higher priority, is optimal.

Proof: The proof of Theorem 2.1 shows that any priority policy that orders the states in the same order as they are picked by the index algorithm is optimal. Therefore, it only remains to show that x can be picked before y by the index algorithm if and only $\gamma(x) \geq \gamma(y)$. Given the recursive nature of the algorithm, it suffices to show that if state s^* is the first one picked by the index algorithm, and q^* is the next state to be picked, then $\gamma(s^*) \geq \gamma(q^*)$. Let i^* be such that $s^* \in \mathcal{X}_{i^*}$. If $q^* \in \mathcal{X}_{i^*}$, then, using Eq. (2.3), we have

$$\gamma(q^*) = \hat{r}(q^*) \leq \max_{x \in \mathcal{X}} r(x) = r(s^*) = \gamma(s^*).$$

If on the other hand $q^* \notin \mathcal{X}_{i^*}$, then $\gamma(q^*) = r(q^*) \leq r(s^*) = \gamma(s^*)$. **Q.E.D.**

For the case of discrete–time Markov bandits, our index algorithm is the same as the one in [VWB85]. This reference, as well as [Wal88], provides some more detail on how the needed calculations can be carried out. Our algorithm is also a special case of the algorithm in [Wei88].

III. DISCUSSION

The proof given here is very simple and it is quite surprising that it was not known earlier. Perhaps a reason is that for the proof to go through, we have to consider semi–Markov bandits rather than the usual discrete–time Markov bandits.

We finally remark that our proof is easily extended to cover the case of arm–acquiring bandits [Whi81] and the even more general case of branching bandits, thus recovering the results of [Wei88]. We assume that the reader is familiar with the framework of [Wei88] and we only point out a few minor modifications of the proof of Theorem 2.1 that are needed. Instead of assuming that the different bandits have disjoint state spaces, we now assume that all bandits share the same state space. We then use induction on the cardinality of this common state space. As in the proof of Theorem 2.1, we pick a top priority state s^* whose reward rate is maximal. We then "eliminate" state s^* and form a reduced bandit as follows: if a bandit at some state $x \neq s^*$ is played, then the play lasts until all type s^* descendants of that bandit have been eliminated; the reward rate during this composite play is also suitably defined, similarly with Eq. (2.3). The resulting index algorithm is identical to the algorithm in [Wei88]. .

REFERENCES

[BeN93] D. Bertsimas and J. Nino–Mora, "Conservation laws, extended polymatroids and the multi-armed bandit problem: a unified polyehdral approach", in preparation, 1993.

[GiJ74] J. C. Gittins, D. M. Jones, "A dynamic allocation index for the design of experiments," in J. Gani, K. Sarkadi and I. Vince (Eds.), *Progress in Statistics*, European Meeting of Statisticians, 1972, Vol. 1, North Holland, 1974, pp. 161–173.

[Tsi86] J. N. Tsitsiklis, "A lemma on the multi-armed bandit problem", *IEEE Transaction on Automatic Control*, 31, 1986, pp. 576–577.

[Tso91] P. Tsoucas, "The region of achievable performance in a model of Klimov", research report, I.B.M., 1991.

[VWB85] P. Varaiya, J. Walrand and C. Buyukkoc, "Extensions of the multi–armed bandit problem: the discounted case", *IEEE Transaction on Automatic Control*, Vol. AC-30, 5, 1985, pp. 426–439.

[Wal88] J. Walrand, *An Introduction to Queueing Networks*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[Web92] R. Weber, "On the Gittins index for multiarmed bandits", *The Annals of Applied Probability*, Vol. 2, 4, 1992, pp. 1024–1033.

[Wei88] G. Weiss, "Branching bandit processes", *Probability in the Engineering and Informational Sciences*, Vol. 2, 1988, pp. 269–278.

[Whi80] P. Whittle, "Multi–armed bandits and the Gittins index", J. Royal Statistical Society, B, 42, 2, 1980, pp. 143–149.

[Whi81] P. Whittle, "Arm acquiring bandits", *Annals of Probability*, Vol. 9, 1981, pp. 284–292.