A NOTE ON BANDITS WITH A TWIST

AKSHAY-KUMAR KATTA* AND JAY SETHURAMAN †

Abstract. A variant of the multi-armed bandit problem was recently introduced by Dimitriu, Tetali and Winkler. For this model (and a mild generalization) we propose faster algorithms to compute the Gittins index. The indexability of such models follows from earlier work of Nash on generalized bandits.

Key words. Multiarmed bandit problem, generalized bandit problem, stochastic scheduling, priority rule, Gittins index, game

AMS subject classifications. 60J10, 66C99, 60G40, 90B35, 90C40

1. Introduction. The multi-armed bandit problem is a well studied optimization problem concerned with dynamically allocating a single resource amongst several competing projects. In the basic version of this problem, there are N independent projects, each of which can be in one of many possible states. At each $t = 1, 2, \ldots$, we must operate exactly one of the projects; as a result, we earn a (possibly random) reward that may depend on the state of the operated project, which undergoes a Markovian state transition. The states of all the other projects remain frozen. Future earnings are discounted by a factor β, and our objective is to decide the order in which we must operate the various projects to maximize the expected total discounted reward earned. Gittins and Jones [7] showed that to each project i, we can attach an index which depends only on the state of project i, and is independent of the states of all the other projects, and that operating a project with the largest index at any point in time is optimal. (Such problems are said to be *indexable*.) Since their original proof, many alternative and insightful proofs have appeared, see [14, 10, 12, 5, 9, 11, 1, 3]. In addition, several natural extensions and variations of the basic multi-armed bandit model have been considered, see [8, 14, 15, 13, 2]. Especially relevant to this work is the generalized bandit model of Nash [8], which considers a class of bandit problems with a more general reward structure. In such a model, the reward obtained from a transition in one project depends in a multiplicatively separable way on the states of all the other projects. Nash [8] proved that this more general class of bandit problems is indexable.

In this paper we consider a variant of the multi-armed bandit problem that has been recently introduced in [4]. Here, as before, we are required to operate exactly one of the projects, except that we are forced to stop upon reaching certain "target" states. The formulation in [4] is in terms of costs (instead of rewards), and is used to model situations in which there are multiple ways to accomplish a certain task, and the goal is to find the "best" way; termination is assumed to be inevitable, and our objective is to operate the projects so as to minimize the total expected cost incurred until termination. By letting the "multiplicative factors" in the generalized bandit model to be zero for the target states and one for the non-target states, we see that the (discounted version of the) model considered here can be viewed as a special-case of the generalized bandit problem.

2. Model & Related Work. There are n bandit processes; the i^{th} process is a Markov chain with a finite state space $\mathcal{S}_i$, and a *sink* $t_i \in \mathcal{S}_i$. For convenience, we

*IEOR Department, Columbia University, New York, NY, ark2001@columbia.edu
†IEOR Department, Columbia University, New York, NY, jay@ieor.columbia.edu

assume that the state spaces of the different bandit processes are disjoint. Time is discrete and is indexed by t. If the i^{th} bandit is at some state $x \in \mathcal{S}_i$ and is operated at time t, then the bandit moves to state $y \in \mathcal{S}_i$ with probability p_{xy}, and a (possibly random) cost C_{xy} is incurred; if $y = t_i$, we stop, otherwise we must choose a bandit to operate at time $t + 1$. Our objective is to operate the bandits over time so as to minimize the expected total cost incurred before termination, which we assume is inevitable (so the expected total cost is finite). For simplicity, we assume that the C_{xy} are deterministic, noting that much of what follows holds true for random C_{xy} by simply replacing the random variables by their expected values.

We shall call the special case in which $C_{xy} > 0$ for all non-sink states x as the *positive-cost* model, distinguishing it from the *general* model in which no assumptions are made about C_{xy}. The indexability of the positive-cost model and the general cost model can be inferred from the classical results of Gittins [6] and Nash [8] respectively, by letting $\beta \to 1$. In [4], the authors prove the indexability of the positive-cost model by adapting Weber's elegant intuitive proof to this setting; in addition, they provide two algorithms to compute the Gittins index, both with complexity $O(n^5)$, where n is the number of non-sink states. Our main observation is that standard techniques result in an $O(n^3)$ algorithm to compute the Gittins index for the general model (and hence for the positive-cost model as well); this matches the complexity of the most efficient algorithm to compute the Gittins index in the usual multi-armed bandit problem [11, 12].

3. Computing the Gittins Index. Since the model considered here is indexable, we focus on a single bandit and show how the Gittins index can be computed for each of its states. Without loss of generality, we assume that the bandit has a single sink, which can be accessed from every other state; let F_x denote the probability of going from state x to the sink in one step. Also, let $C_x \equiv \sum_y C_{xy} p_{xy}$ be the expected cost of operating the bandit when it is in state x. For convenience, we also assume that $F_x > 0$ for every non-sink state x. Later we show how this assumption can be relaxed.

An alternative characterization of the Gittins index is the key to computing it efficiently, so we discuss this briefly. Consider a "game" in which, at each step, one is faced with two choices: continuing to operate the bandit, which costs (on average) C_x if the bandit is in state x, or quitting by paying a fee of M dollars. It is well-known that the Gittins index, ν_x, of a state x is the unique value of M at which one is indifferent between operating the bandit in state x and quitting.

Suppose state x has the smallest Gittins index, and suppose the bandit is currently in state x. Let the fee in the game described earlier be ν_x. By definition, it is optimal to operate the bandit once, and quit by paying ν_x if the resulting state is not a sink; thus $\nu_x = C_x/F_x$. Unfortunately, we do not know the state with the smallest Gittins index, so we test all possibilities. From the alternative characterization of the Gittins index mentioned earlier, it is clear that x is a state with the smallest Gittins index if and only if

$$x = \arg\min_{y \in \mathcal{S}} \frac{C_y}{F_y}.$$

Having identified a state with the smallest Gittins index, we can now "reduce" the bandit by eliminating x in the following manner (see [11]). Consider any non-sink state $y \neq x$ with $p_{yx} > 0$. In computing the Gittins index of y, we may assume that whenever we make a transition to x, we continue to operate the bandit until we leave

x to reach some state z (which may possibly be y itself or even the sink); this sequence of plays may be regarded as a single play with a "cost"

$$\hat{c}_{yz} \;=\; C_{yx} + C_{xx}\left\{\frac{1}{1-p_{xx}} - 1\right\} + C_{xz},$$

and a transition probability

$$\hat{p}_{yz} \;=\; p_{yx}\, p_{xz}/(1-p_{xx}).$$

We note that $\hat{c}_{yz}$ is the expected cost incurred during this composite play, which can be broken down into three components: the first transition from y to x, costing C_{yx}; the successive self-transitions at x, whose expected number is $1/(1-p_{xx})-1$, each costing C_{xx}; and the last transition from x to z, costing C_{xz}. The conditional probability of an (x,z) transition, given that a transition from x to another state occurs is $p_{xz}/(1-p_{xx})$, which justifies the expression for $\hat{p}_{yz}$. If the (y,z) arc does not already exist, we introduce one, and let $C_{yz} = \hat{c}_{yz}$, $p_{yz} \;=\; \hat{p}_{yz}$; if the (y,z) arc already exists, the cost for a y to z transition is updated as

$$C_{yz} \leftarrow \frac{p_{yz}C_{yz} + \hat{p}_{yz}\hat{c}_{yz}}{p_{yz} + \hat{p}_{yz}},$$

and the transition probability from y to z now becomes

$$p_{yz} \leftarrow p_{yz} + \hat{p}_{yz}.$$

For a bandit with n states, a state with the smallest Gittins index can be determined in $O(n)$ time; the reduction algorithm needs to examine $O(n^2)$ pairs, each of which requires $O(1)$ time; so the complexity per iteration is $O(n^2)$ when there are n states. The (reduced) bandit now has one less state; we proceed as before by identifying a state with the minimum Gittins index, eliminating this state to further reduce the bandit, etc. After $(n-1)$ applications of the reduction algorithm we will have determined the Gittins index for all the non-sink states; thus the overall complexity of computing the Gittins index for an n-state bandit is easily seen to be $O(n^3)$.

We now show how the assumption $F_x > 0$ for all non-sink states x can be relaxed. Let x be a non-sink state with $F_x = 0$. If $C_x \leq 0$, then we will always operate the bandit in state x, so $\nu_x = -\infty$. (Such states must be reduced first.) If $C_x > 0$, it is clear that x cannot be a state with the minimum index; in fact, it is easy to see that some state adjacent to x must have a lower index (see [4]). In this case, the index of state x will be determined by the algorithm at a later point.

Finally, we note that the algorithm proposed here can be extended to more general versions of the problem such as the semi-Markov version (time is not slotted), and the discounted version. We leave the obvious modifications to the reader.

4. Acknowledgements. A version of the problem described here was the subject of the first author's final project in a graduate course on dynamic programming taught by the second author. We thank John Tsitsiklis for sharing his thoughts on this problem and Kevin Glazebrook for telling us about the relevance of Nash's work on generalized bandits; in addition, we thank them both for their comments on an earlier version of this paper. This research was supported by an NSF grant DMI-0093981 and by an IBM partnership award.

REFERENCES

[1] D. Bertsimas and J. Nino-Mora (1996) Conservation laws, extended polymatroids and multi-armed bandit problems: A polyhedral approach to indexable systems, *Mathematics of Operations Research*, **21**(2):257–306.

[2] J. H. Crosbie and K. Glazebrook (2000) Index policies and a novel performance space structure for a class of generalized branching bandit problems, *Mathematics of Operations Research*, **25**(2):281–297.

[3] M. Dacre, K. Glazebrook, and J. Nino-Mora (1999) The achievable region approach to the optimal control of stochastic systems, *J. Roy. Statist. Soc. Ser. B*, **61**(4):747–791.

[4] I. Dumitriu, P. Tetali, and P. Winkler (2003) On Playing Golf with Two Balls, *SIAM J. Disc. Math.*, **16**(4):604–615.

[5] J. C. Gittins (1989) Multi-Armed Bandit Allocation Indices, Wiley, New York.

[6] J. C. Gittins (1979) Bandit processes and dynamic allocation indices, *J. Roy. Statist. Soc. Ser. B*, **41**, 148–177.

[7] J. C. Gittins and D. M. Jones (1974) A dynamic allocation index for the sequential design of experiments, Read at the 1972 European Meeting of Statisticians, Budapest, *Progress in Statistics*, (J. Gani et al.), 241–266, North-Holland.

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

[9] R. Weber (1992) On the Gittins Index for Multiarmed Bandits, *Annals of Applied Probability*, **2**(4):1024–1033.

[10] J. N. Tsitsiklis (1986) A lemma on the multi-armed bandit problem, *IEEE Trans. Automat. Control*, **AC-31**, 576–577.

[11] J. N. Tsitsiklis (1994) A Short Proof of the Gittins Index Theorem, *Annals of Applied Probability*, **4**(1):194–199.

[12] P. Varaiya, J. Walrand, and C. Buyukkoc (1985) Extensions of the multi-armed bandit problem: The discounted case, *IEEE Trans. Automat. Control*, **AC-30**, 426–439.

[13] G. Weiss (1988) Branching bandit processes, *Probab. Engng. Inform. Sci.*, **2**, 269–278.

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

[15] P. Whittle (1981) Arm acquiring bandits, *Ann. Probab.*, **9**, 284–292.