The Ratio Index for Budgeted Learning, with Applications

Ashish Goel * Sanjeev Khanna † Brad Null ‡

October 3, 2008

Abstract

In the budgeted learning problem, we are allowed to experiment on a set of alternatives (given a fixed experimentation budget) with the goal of picking a single alternative with the largest possible expected payoff. Constant factor approximation algorithms for this problem were developed by Guha and Munagala by rounding a linear program that couples the various alternatives together. In this paper we present an index for this problem, which we call the ratio index, which also guarantees a constant factor approximation. Index-based policies have the advantage that a single number (i.e. the index) can be computed for each alternative irrespective of all other alternatives, and the alternative with the highest index is experimented upon. This is analogous to the famous Gittins index for the discounted multi-armed bandit problem.

The ratio index has several interesting structural properties. First, we show that it can be computed in strongly polynomial time. Second, we show that with the appropriate discount factor, the Gittins index and our ratio index are constant factor approximations of each other, and hence the Gittins index also gives a constant factor approximation to the budgeted learning problem. Finally, we show that the ratio index can be used to create an index-based policy that achieves an $O(1)$-approximation for the finite horizon version of the multi-armed bandit problem. Moreover, the policy does not require any knowledge of the horizon (whereas we compare its performance against an optimal strategy that is aware of the horizon). This yields the following surprising result: there is an index-based policy that achieves an $O(1)$-approximation for the multi-armed bandit problem, oblivious to the underlying discount factor.

1 Introduction

The classical multi-armed bandit problem provides an elegant model to study the tradeoff between collecting rewards in the present based on the current state of knowledge (exploitation) versus deferring rewards to the future in favor of gaining more knowledge (exploration[1]) [2]. Specifically, in this model a user has a choice of bandit-arms to play, and at each time step it must decide which arm to play. The expected reward from playing a bandit-arm depends on the state of the bandit-arm where the state represents a "prior" belief on the bandit-arm. Each time a bandit-arm is played, this prior gets updated according to some transition matrix defined on the state space. For instance, a typical assumption on the bandit-arms is that they have (α, β)-priors: the success probability of an (α, β)-bandit-arm is $\alpha/(\alpha+\beta)$; in case of a success a reward of 1 is obtained and α gets incremented, whereas in case of a failure no reward is obtained and β gets incremented. The user wishes to maximize the total expected discounted reward over time. This simple setting effectively models many applications. A canonical example is exploring the effectiveness of different treatments in clinical trials while maximizing the benefit received by patients.

The discount factor in a multi-armed bandit problem may be viewed as modulating the horizon over which the strategy explores to identify the bandit-arm with maximum expected reward, before switching to exploitation. This facet of the multi-armed bandit problem is explicitly captured by the *budgeted learning problem*, recently studied by Guha and Munagala [19]. The input to the budgeted learning problem is the same as for the multi-armed bandit problem, except the discount factor is replaced by a horizon h. The goal is to identify the bandit-arm with maximum expected reward using

*Departments of Management Science and Engineering and (by courtesy) Computer Science, Stanford University. Research supported by NSF ITR grant 0428868, NSF CAREER award 0339262, and gifts from Google, Microsoft, and Cisco. Email: ashishg@stanford.edu .

†Department of Computer and Information Science, University of Pennsylvania, Philadelphia PA. Email: sanjeev@cis.upenn.edu. Supported in part by a Guggenheim Fellowship, an IBM faculty award, and NSF Award CCF-0635084.

‡Department of Management Science and Engineering, Stanford University. Research supported by NSF ITR grant 0428868. Email: null@stanford.edu .

[1] We will use the terms experimentation and exploration interchangeably in this paper, depending on the context.

at most h steps of exploration. The work of [19] gives a constant factor approximation for the budgeted learning problem via a linear programming based approach that determines the allocation of exploration and exploitation budgets across the various arms. The budgeted learning problem is the main object of study in this paper.

The multi-armed bandit problem admits an elegant solution: compute a score for each bandit-arm using only the current state of the bandit-arm and the discount factor, *independent* of all other bandit-arms in the system, and then play the bandit-arm with the highest score. This score is known as the *Gittins index*, and many proofs are known to show this is an optimal strategy (e.g., see [9]). The optimality of this "index-based" strategy implies that this problem exhibits a "separability" property whereby the optimal decision at each step is obtained by computations performed *separately* for each bandit-arm. This structural insight translates into efficient decision making algorithms. In fact, for commonly used prior update rules and discount rates, extensive collections of pre-computed Gittins indices exist, enabling in principle, a simple lookup-based approach for optimal decision-making. There are multiple definitions of what it means for a problem to have an "index". We will use the term index in its strongest form, i.e., where the index of an arm depends *only on the state of that arm*. This is also sometimes called a decomposable index (eg. [3, 4]).

The inherent appeal and efficiency of index-based policies is the unifying theme underlying our work. We show that many interesting and non-trivial variations of the multi-armed bandit problem, including the budgeted learning problem and the finite horizon problem, can all be well-approximated by index-based policies. Moreover, our approach gives decision strategies that are *oblivious* to parameters such as the underlying horizon or the discount factor while being constant-factor competitive to optimal strategies that are fully aware of these parameters.

1.1 Our results We will study this problem when the state space of each arm satisfies the "martingale property", i.e., if we play an arm multiple times, the sequence of expected rewards is a martingale. This is a natural assumption for multi-armed bandit and related problems, e.g. the commonly used (α, β) priors satisfy this property.

An Index for Budgeted Learning Problems: Our first result is that the budgeted learning problem admits an approximate index, which we call the *ratio index*. Informally speaking, given a single bandit-arm and an exploration budget of h steps, the ratio index for that arm is the maximum expected exploitation reward per unit of the exploration and exploitation budget utilized. The ratio index suggests the following natural algorithm: at each step, play the arm with the highest ratio index. We show that this simple greedy algorithm gives a constant factor approximation to the budgeted learning problem. An $O(1)$-approximation algorithm for this problem is already known [19]. However, the algorithm of [19] is based on solving a *coupled* LP over all the arms, whereas the ratio index can be computed for each arm in isolation, much like the Gittins index. The ratio index has many other interesting properties. For example:

(1) We show that the Gittins index with discount factor $(1 - 1/h)$ and the ratio index over horizon h are within a constant factor of each other. This gives the following surprising result:

THEOREM 1.1. *Given an exploration budget h, playing at each step the arm with the highest Gittins index, with discount factor $1 - 1/h$, yields a constant factor approximation to the budgeted learning problem.*

The proof relies on comparing the "decision-trees" of the ratio index and Gittins index strategies. Even in retrospect, it is not clear to us how such a result could be derived using an LP-based formulation such as the one used by Guha and Munagala [19]. Interestingly, the policy described in theorem 1.1 is known to often work well in practice [24]. Nonetheless, before the work of Guha and Munagala [19], we do not know of any provable guarantees for polynomial time algorithms in this setting. And until now, we do not know of any formal guarantees that relate the exponential discounting approach (which yields the Gittins index) and the budgeted learning approach.

(2) The ratio index can be computed in time which is strongly polynomial in the size of the state space (independent of h) of each arm if the state space is acyclic, and strongly polynomial in the size of the state space and h if the state space is general. Our proof of this fact involves recursively analyzing the basic feasible solutions of an underlying LP for computing optimum single arm strategies and using the structure of the basic feasible solutions to prove that these strategies have a simple form.

Finite Horizon and Discount-Oblivious Multi-Armed Bandits: We next study an important and natural variation of the budgeted learning problem, called the *finite horizon* multi-armed bandit problem.

We are given a finite horizon h, and the goal is to maximize the expected reward collected during the horizon. Thus, in contrast to the budgeted learning problem, the horizon h is being used for both exploration and exploitation, and no payoffs are obtained after time h. We show the following result using the ratio index:

THEOREM 1.2. *There is an index-based policy that gives a constant factor approximation to the finite horizon multi-armed bandit problem.*

Finally, we study the role of the discount factor in the design of an optimal strategy for the exploration-exploitation tradeoff. Small variations in discount factors can alter the choice of bandit-arm played at any step, highlighting the sensitivity of the Gittins index to the discount rate. We study the "Discount-oblivious" multi-armed bandit problem where the underlying discount factor is not known, and in fact, may even vary from one time step to the next. A finite horizon problem can be viewed as a special case of this general setting where the discount factor is 1 for the first h steps and is 0 for all subsequent steps. There is a useful relationship between the finite horizon and discount oblivious versions of the multi-armed problem: a strategy is κ-approximate for the discount-oblivious multi-armed bandit problem iff it is κ-approximate (*simultaneously*) for all finite horizons. Using this connection, and building on Theorem 1.2, we show the following result:

THEOREM 1.3. *There is an index-based policy that gives a constant factor approximation for the multi-armed bandit problem with respect to all possible discount factors simultaneously.*

Our proof of both of these results is based on the following easy consequence of the ratio index approach to the budgeted learning problem. For any constant β, the expected profit of the optimal h/β-horizon strategy is an $\Omega(1)$-fraction of the expected profit of an optimal h-horizon strategy. Using this result, we design an algorithm that alternates between budgeted exploration and exploitation, using geometrically increasing horizons; each increasing horizon competing against a lower discount rate on future rewards. It is worth noting that this result can also be shown using the LP-based proof of Guha and Munagala. However, the following corollary is a consequence of our index-based approach and the relation between ratio and Gittins indices.

COROLLARY 1.1. *The strategy that alternates between exploring the arm with the highest Gittins index, and exploiting the arm with the highest reward, in phases of geometrically increasing length (and discount factor $1 - 1/t$ during a phase of length t) provides a constant factor approximation to the multi-armed bandit problem* simultaneously *for all finite horizons and for all discount factors.*

1.2 Related Work and Organization There are many sources for the canonical work on Gittins indices, particularly with reference to (α, β) bandits and Bernoulli bandit processes [10, 11, 12, 9]. Glazebrook and others have studied approximation algorithms for other extensions to multi-armed bandit problems [13, 14]. Their approach builds upon the concept of achievable regions and general conservation laws and a related linear programming approach built by Tsoucas, Bertsimas, Nino-Mora, and others [1, 3, 26]. Relaxed linear programming based approaches to extensions of the multi-armed bandit problem have also been developed, e.g. for restless bandits [28, 4]. Our work on the ratio index builds on the insights obtained from the LP relaxation based approach of Guha and Munagala [19] as well as related work in model-driven optimization [15, 18] and stochastic packing [7, 8, 16, 20]. Additionally, related LP formulations have been developed for multi-stage stochastic optimization [6, 25].

In the theoretical computer science community, multi-armed bandits have primarily been studied in an adversarial setting, with the goal being to minimize the regret (see [5] for a nice overview). A typical guarantee in these settings is that the *total regret* after T steps grows as $\tilde{O}(\sqrt{TN})$ where N is the number of alternatives, assuming the partial information model (i.e. only the reward for the alternative that is actually played is revealed), which corresponds well to our setting. These results assume no prior beliefs, unlike our decision theoretic framework. However, the regret based bounds in the adversarial setting are meaningless unless $T > N$. The decision theoretic framework which has a rich history (starting perhaps with Wald's work in 1947 [27]) is more suited to the situation where the number of exploration steps is drastically limited, as is often the case. A typical setting, for example, is one where an advertiser that can advertise on 100,000 possible phrases and is willing to pay for 100 clicks to decide which keyword attracts visitors that convert into paid customers. So a traditional regret based bound may not be very meaningful in this setting.

In section 2 we define the budgeted learning problem and the ratio index, and prove that the ratio index is a constant factor approximation to the budgeted learning problem. Section 3 establishes that the Gittins and ratio indices are constant factor approximations of each other. We also show here that playing the arm with the largest Gittins index (with a suitable discount factor), gives a constant factor approximation to the bud-

geted learning problem. Section 4 presents index-based policies for finite horizon and discount oblivious versions of the multi-armed bandit problem. In section 5, we present a strongly polynomial algorithm to compute the ratio index as well as several useful insights into its structural properties.

2 The Budgeted Learning Problem and the Ratio Index

2.1 The Budgeted Learning Problem. *We are given n arms. Arm i has state space T_i, with initial state ρ_i. Experimenting on an arm i in state $u \in T_i$ results in the arm entering state $v \in T_i$ with known probability P_{uv}. The payoff of state u is given as $\zeta(u)$. Given an experimentation budget h, we are interested in finding the optimal policy, π^*, so that $\mathbf{E}_{\pi^*}[\max_{i\in\{1,\ldots,n\}} \zeta(v_i)]$ is maximum among all policies, where v_i is the state of arm i after the policy has been executed (the number of experiments cannot exceed h).*

We will use T to denote $\cup_i T_i$. For convenience, we will assume that the T_i are disjoint and that $P_{uv} = 0$ if u and v are in the state spaces of different bandit-arms; this can be easily enforced by duplicating any shared states. The initial states represent a prior belief on the payoff from the bandit-arms. We will assume that the expected payoff is a martingale, i.e., $\zeta(u) = \sum_{v\in T} P_{uv}\zeta(v)$; the martingale assumption is crucial to our results. We will also assume without loss of generality that the state space of any arm is acyclic and truncated at depth h.

The martingale property has some useful and easy consequences which we will use repeatedly:

1. For an arbitrary policy let $p(t)$ denote its expected payoff if it is terminated after t experiment steps. Then, $p(t)$ is non-decreasing in t. In other words, extra experiments can never hurt.

2. Given a *single arm*, no policy can have a higher expected payoff than the one which does no exploration and simply chooses the initial state as the winner; in other words, extra experiments can never help given just one arm.

The proof of the following theorem is deferred to the full version of this paper [17]. It is conceivable (though not obvious to us) that this theorem can also be obtained via the "indexability characterization" of [3]. In any case, the proof is quite elementary and provides useful intuition.

THEOREM 2.1. *There is no exact index for the budgeted learning problem.*

2.2 The ratio index We will now define the ratio index, which is an approximate index for this problem. At any given time, the *current state* of the system is denoted by $\mathbf{S} = \{u_1, u_2, \ldots, u_n, \delta\}$, which captures the current states of all the arms, and the budget left (i.e. the number of experimentation steps that are still remaining), δ. The initial state of the system has all the arms in their initial states, and $\delta = h$. Since we use the term state for both the system and an arm, we will disambiguate where necessary by referring to these as "system-state" and "arm-state" respectively. A policy π is a function which takes as input a system-state $\mathbf{S}$ and either returns an arm i for experimentation (i.e. explores the arm-state u_i), or chooses an arm i as a winner and terminates (i.e. exploits the arm-state u_i), or simply terminates (abandons). If $\delta = 0$ then the only options are to abandon or exploit. The martingale property (see the comment at the end of section 2.1) implies that there always exists an optimal policy which explores some arm iff $\delta > 0$ and exploits some arm iff $\delta = 0$. We now introduce two vectors x^π and z^π. The probability that arm-state u is the final exploited state by policy π is given by x^π_u. The probability that arm-state u is explored by policy π is given by z^π_u. We define the *cost* of policy π as

$$\mathcal{C}(\pi) = \frac{\sum_{u\in T} z^\pi_u}{h} + \sum_{u\in T} x^\pi_u.$$

Observe that $\mathcal{C}(\pi) \leq 2$, for any policy π. The profit of policy π is defined as

$$\mathcal{P}(\pi) = \sum_{u\in T} x^\pi_u \zeta(u).$$

Observe that our definition of policy is an adaptive one; the decisions made in step $j > 1$ depend on the entire system-state at time j and hence on the outcome of previous experimentation steps. Further, it is easy to see that randomized strategies can not do any better than deterministic strategies.

If we drop the requirement that a policy must either exploit or abandon when the remaining budget δ is 0, we obtain what we call a *pseudo-policy*. A single arm policy is one which makes all its decisions based only on the state of a single pre-determined arm i, ignoring all other arms. We are now ready to define the ratio index and prove that it leads quite naturally to an approximation of the budgeted learning problem.

Ratio Index. *The ratio index $r(u, h)$ of a bandit-arm (say arm i) in initial state u and with experimentation budget h, is defined as*

$$\max_\pi \frac{\mathcal{P}(\pi)}{\mathcal{C}(\pi)},$$

where the max *is over all single arm pseudo-policies* π *which have initial arm-state* u*, budget* h*, state space* T_i*, and cost* $\mathcal{C}(\pi) > 0$*. We refer to a policy which yields the ratio index as a ratio index policy for state* u*, denoted* $\pi_r(u, h)$*.*

Even though we allow pseudo-policies in the definition of the ratio index, any ratio index policy respects the budget constraint:

LEMMA 2.1. *Any ratio index policy for an initial state* u *has cost at most* 1*.*

Proof. Because of the martingale property, no single arm policy starting from arm-state u can obtain profit more than $\zeta(u)$. Hence, any single arm policy π that has cost more than 1 must have a smaller ratio (of expected profit to expected cost) than the single arm policy which exploits in state u. ∎

Greedy Algorithm. *Suppose the initial experimentation budget is* h*, and the current system-state is given by* $\mathbf{S} = \{u_1, u_2, \ldots, u_n, \delta\}$*. If* $\delta > 0$*, the greedy algorithm explores the arm* i *with the maximum ratio index,* $r(u_i, h)$*, with ties broken arbitrarily but consistently. If* $\delta = 0$ *the greedy algorithm exploits the arm* i *with maximum current expected reward* $\zeta(u_i)$*. We denote the greedy algorithm by* G*.*

Note: The greedy algorithm uses the same h at every step to compute the ratio index. Hence, given a table of the ratio index of every state in T (which can be precomputed efficiently as specified in the section 5), we can implement this algorithm using a simple min-heap and the complexity of each step would be just $O(\log n)$, which is much better than solving a coupled LP with $3nh$ variables.

2.3 Analysis of the greedy algorithm We now show that the greedy algorithm gives an $O(1)$-approximation to the budgeted learning problem.

LEMMA 2.2. *A ratio index policy for arm-state* u*,* $\pi_r(u, h)$*, does not abandon any arm-state* v *with* $r(v, h) > r(u, h)$ *and does not explore or exploit any arm-state* v *with* $r(v, h) < r(u, h)$*.*

The proof of the above lemma is deferred to the full version of this paper [17]. Now consider the following algorithm, which we call the *persistent* algorithm, denoted G'.

The persistent algorithm G′: Given a system-state $\mathbf{S}$, let i be the arm with the highest ratio index $r(u_i, h)$ where u_i denotes the current state of arm i. Play arm i in accordance with the policy $\pi_r(u_i, h)$ until the policy chooses to exploit or abandon. If $\pi_r(u_i, h)$ abandons, let $\mathbf{S}'$ be the resulting system-state. Repeat the process starting with $\mathbf{S}'$. If at any time, the system-state is such that $\delta = 0$, immediately exploit the arm that has the highest current ratio index.

Observe that as for the greedy algorithm, the ratio index used by the persistent algorithm G' is computed using a fixed budget h; the number of remaining exploration steps δ is used only to terminate G'.

LEMMA 2.3. *The expected profit of the greedy algorithm* G *is at least as much as the expected profit of the persistent algorithm* G'*.*

Proof. Couple the greedy and the persistent algorithms such that if they both explore an arm in a given state, that arm transitions to the same state for both strategies. Let $I = \langle i_1, i_2, \ldots, i_\kappa \rangle$ denote the sequence of arms explored by G' before exploitation; here $\kappa \leq h$ since G' can exploit early. Let $J = \langle j_1, j_2, \ldots, j_h \rangle$ denote the set of arms explored by G. By Lemma 2.2, we can conclude that I is a prefix of J. By the martingale property, early termination can never result in increased profit; the lemma follows. ∎

Thus it suffices to analyze G'. Given two single arm pseudo-policies $\tilde{\pi}$ and π for an arm i, we say that $\tilde{\pi} \succeq \pi$, if for all arm-states $u \in T_i$ for which π explores (exploits) arm i, $\tilde{\pi}$ also explores (exploits) the arm i. Notice that $\tilde{\pi}$ might choose to continue exploration/exploitation when π abandons an arm-state. Informally, $\tilde{\pi} \succeq \pi$ means that policy $\tilde{\pi}$ can be played after policy π has been played to completion. We will now state a useful technical lemma; the proof is in the full version [17].

LEMMA 2.4. *Given two arbitrary single arm pseudo-policies* π*,* π' *for arm* i *in initial arm-state* u*, there exists another single arm pseudo-policy* $\tilde{\pi}$ *starting in the same initial arm-state* u *such that, (1)* $\tilde{\pi} \succeq \pi$*, (2)* $\mathcal{C}(\tilde{\pi}) - \mathcal{C}(\pi) \leq \mathcal{C}(\pi')$*, and (3)* $\mathcal{P}(\tilde{\pi}) \geq \mathcal{P}(\pi')$*.*

The above property is akin to submodularity. We now state our main theorem, which says that the greedy algorithm G gives a constant factor approximation to the optimal policy. Let $B_\pi(h, \mathbf{S})$ denote the expected profit obtained by strategy π for the budgeted learning problem run with budget h and initial system-state $\mathbf{S}$, and let $B^*(h, \mathbf{S})$ denote the expected profit obtained by an optimum strategy with the same parameters. We will omit the system-state when it is the same for all the strategies involved.

THEOREM 2.2. $B_G(h) \geq 0.22 B^*(h)$

Proof. From Lemma 2.3, it suffices to analyze the persistent algorithm G' rather than the greedy algorithm G.

We divide the persistent algorithm into stages, starting from stage 1. Let i_1 be the arm with the highest ratio index at the beginning of stage 1 (and hence the arm that will be played by G' at the first step). Since the arms evolve probabilistically, the first stage (as well as subsequent stages) will result in a distribution over system-states. Let $\mathbf{S}_j$ denote the system-state at the start of stage j, and let $\mathcal{D}_j$ denote the distribution of $\mathbf{S}_j$. Let u_j be the arm-state with the highest ratio index among the arm-states which have a non-zero probability, say γ_j, in $\mathcal{D}_j$, and let i_j be the corresponding arm. The j-th stage of G' is to simply move to the next stage if the arm i_j is not in state u_j (which happens with probability $1-\gamma_j$); we call this stage "empty" in this case. If the arm i_j is in state u_j, then the j-th stage of G' is to mimic an optimum ratio index policy for state u_j. If the exploration budget gets exhausted during this mimicking process, then the j-th stage exploits arm i_j right away and the policy terminates; the cost of the extra exploitation is not charged to this stage of the policy. By the martingale property, this early termination can only increase the expected profit of the j-th stage. If the j-th stage exploits an arm, then the persistent algorithm terminates as well.

Let π_j denote the policy corresponding to the j-th stage. Let p_j and c_j be the cumulative expected profit and expected cost of the first j stages. Use $\Delta_p(j)$ and $\Delta_c(j)$ to denote the expected profit and the expected cost of the j-th stage, conditioned on this stage being played (i.e. the j-th stage being non-empty and the persistent algorithm not terminating before reaching the j-th stage). The following statement is a corollary of Lemma 2.4: the proof is a digression from the current theorem and is deferred to the full version [17].

COROLLARY 2.1. *At the beginning of stage j, there exists a single arm pseudo-policy with profit to cost ratio at least $(\mathcal{P}(\pi^*)-p_{j-1})/2$.*

Since the persistent algorithm follows an optimum ratio index policy we are guaranteed that $\Delta_p(j)/\Delta_c(j) \geq (\mathcal{P}(\pi^*)-p_{j-1})/2$. By Markov's inequality, the probability that the budget has not been exhausted before stage j starts is at least $1-c_{j-1}$. Also, recall that γ_j is the probability that the j-th stage is non-empty. The expected unconditioned profit of the j-th stage is at least $\gamma_j(1-c_{j-1})\Delta_p(j)$. The expected unconditioned cost of the j-th stage is at most $\gamma_j\Delta_c(j)$. Hence, we get

$$\frac{p_j-p_{j-1}}{c_j-c_{j-1}} \geq \frac{\mathcal{P}(\pi^*)-p_{j-1}}{2}\cdot(1-c_{j-1}).$$

Thus, the profit obtained by the persistent algorithm is more than the one attained by the following differential process, where p is the cumulative profit and c is the cumulative cost, and $p^* = \mathcal{P}(\pi^*)$ (view the process as increasing the expected cost from 0 to 1):

$$\frac{dp}{dc} = \frac{p^*-p}{2}(1-c).$$

Integrating from $c=0$ to $c=1$, we get that the expected profit is at least $(1-e^{-0.25})p^* \geq 0.22p^*$. ■

Thus, we have shown the existence of a simple index which yields almost as good an approximation ratio as the LP-based approach of Guha and Munagala. The results above assume that each exploration step has the same cost, but can easily be extended to the weighted exploration cost case. We can also modify the proof slightly to obtain the following corollary[2]:

COROLLARY 2.2. $B^*(h/2) \geq 0.17B^*(h)$.

Combining Corollary 2.2 and Theorem 2.2, we obtain the following corollary:

COROLLARY 2.3. $B_G(h/2) = \Omega(B^*(h))$.

3 Relating the Gittins and Ratio indices

We will use S_δ to denote a standard (stationary) bandit-arm with a fixed reward of δ. We will use $\mathcal{A}$ to denote a given bandit-arm in some initial state u. A *Gittins index strategy* $\mathcal{S}$ takes as input an arm $\mathcal{A}$ with an unknown reward distribution (but a known initial state) and a standard bandit-arm S_δ for some $\delta \geq 0$, and gives a strategy for maximizing the discounted reward for a multi-armed bandit with $\mathcal{A}$ and S_δ as its two bandit-arms. Thus each node in the decision tree of $\mathcal{S}$ is labeled as playing either the given arm $\mathcal{A}$ or the standard bandit S_δ. We can assume w.l.o.g. that once the strategy $\mathcal{S}$ plays the standard bandit at a node in the tree, it plays it forever from then onwards. The *Gittins index* of an arm $\mathcal{A}$ is defined to be the least δ such that the Gittins index strategy with input arms $\mathcal{A}$ and S_δ is indifferent between playing either one of them at time 0. We will assume u to be the initial state of $\mathcal{A}$ in the remainder of this section, and drop its explicit mention. Let $r(h)$ denote the ratio index for $\mathcal{A}$ when the horizon is limited to h. Let $\rho(\theta)$ denote the Gittins index for $\mathcal{A}$ when the discount factor is uniform for some $0<\theta<1$. The following lemmas show that the Gittins and the ratio indices are constant factor approximations of each other. The proofs involve transforming the Gittins index strategy to the ratio index strategy (and vice verse) and are in the full version [17].

LEMMA 3.1. *For any $h \geq 2$, $\rho(\theta) \geq r(h)\left(1-\frac{1}{h}\right)^h$ where $\theta = (1-\frac{1}{h})$. Thus as $h \to \infty$, $\rho(\theta) \geq r(h)/e$.*

[2]This corollary can also be obtained using the LP-based framework of Guha and Munagala [19].

LEMMA 3.2. *For any $h \geq 2$, $\rho(\theta) \leq (2+4e)r(h)$ where $\theta = (1 - \frac{1}{h})$.*

LEMMA 3.3. *Let $\rho_i(t)$ denote the Gittins index of arm i at time t, where the discount factor θ is $1-1/h$. Playing the arm with the highest value of $\rho_i(t)$ for $t = 1, 2, \ldots, h$ and then picking the arm with the highest expected payoff at time t results in a constant factor approximation to the budgeted learning problem.*

While the constants in our proofs are large, the algorithms are simple and intuitive. For instance, Schnieder and Moore [24] and Madani, Lizotte and Greiner [22] have studied the policy defined in lemma 3.3 and other similar policies, and found that they often work well in practice.

4 Finite Horizon and Discount Oblivious Multi-armed Bandits

In the traditional multi-armed bandit problem, we are given a fixed discount factor $\theta \in (0,1)$ and allowed to play one arm at each time. If the reward at time t is $r(t)$ then the total discounted reward is $\sum_{t \geq 0} \theta^t r(t)$. Always playing the arm with the currently highest Gittins index maximizes the expected total discounted reward; however the Gittins index of an arm depends crucially on the parameter θ. In this section, we discuss both *finite horizon* and *discount oblivious* versions of the multi-armed bandit problem.

In the finite horizon multi-armed bandit problem, we are given a fixed number of steps, h, as in the budgeted learning problem. However, unlike the budgeted learning problem, the objective of the finite horizon problem is to maximize the total (undiscounted) expected reward obtained *during* the first h steps. This models many important problems such as optimally placing bets with a fixed number of chips, and optimally assigning impressions to advertisers [23].

In the discount oblivious multi-armed bandit problem, we want to find a strategy that provides a constant factor approximation to the optimum reward for all $\theta \in (0,1)$ *simultaneously*. It is not clear up front that such a strategy exists. In fact, we will allow the discounts to be even more general. Let $\Lambda = \langle \Lambda_0, \Lambda_1, \Lambda_2, ... \rangle$ be an infinite sequence of discount factors that satisfies the property $1 = \Lambda_0 \geq \Lambda_1 \geq \Lambda_2 \geq ...$ and where $\Lambda_t \to 0$ as $t \to \infty$. We will call such a sequence a discount factor sequence. Let the system-state $\mathbf{S}$ denote the vector of all the arm-states. We will use $D_\pi(\Lambda, \mathbf{S})$ to denote the total expected discounted reward of any strategy π for discount factor sequence Λ starting from system-state $\mathbf{S}$. If strategy π obtains reward $r(t)$ at time t when started in initial system-state $\mathbf{S}$, then $D_\pi(\Lambda, \mathbf{S}) = \sum_{t=0}^{\infty} \Lambda_t r(t)$. Setting $\Lambda_t = \theta^t$ leads to the standard multi-armed bandit problem. Setting $\Lambda_t = 1$ for $t < h$ and $\Lambda_t = 0$ otherwise leads to a fixed horizon problem where we only get the reward from the first h time steps.

We will use $F_\pi(h, \mathbf{S})$ to denote the total (undiscounted) expected reward over a window of h steps of any strategy π, starting from $\mathbf{S}$. We will use $D^*(\Lambda, \mathbf{S})$, $F^*(h, \mathbf{S})$ to denote the optimum values for the two problems. We will omit the parameter $\mathbf{S}$ when it is the same for all strategies under discussion. All proofs are deferred to the full version [17].

4.1 An approximate index for the finite horizon problem

Recall (from section 2) that $B_G(h)$ and $B^*(h)$ denote the expected profit of the greedy algorithm (which always explores the arm with the largest ratio index) and the optimum strategy respectively, for the budgeted multi-armed bandit problem. We first relate the budgeted learning and finite horizon problems:

LEMMA 4.1. *For any positive integer h, we have $\lceil \frac{h}{2} \rceil \cdot B^* \left(\lfloor \frac{h}{2} \rfloor \right) \leq F^*(h) \leq h \cdot B^*(h)$.*

We will now define two index-based strategies for the finite horizon problem, assuming horizon h and initial system-state $\mathbf{S}$:.

1. For the first $\lfloor h/2 \rfloor$ steps, play the arm with the highest ratio index, where the ratio index is computed assuming a budget of $\lfloor h/2 \rfloor$. For the remaining $\lceil h/2 \rceil$ steps, play the arm with the highest expected reward. We will denote this strategy as RATIOSWITCH$(h, \mathbf{S})$ since it switches from using the ratio index (in the first half) to using the expected profit as an index in the second half.

2. Similarly, define GITTINSSWITCH$(h, \mathbf{S})$ as the strategy which plays the arm with the highest Gittins index (assuming a discount factor $1-1/\lfloor h/2 \rfloor$) during the first $\lfloor h/2 \rfloor$ steps, and then switches to using the arm with the highest expected reward.

Observe that both strategies use an index at each step, and the choice of index does not depend on the state of the system; it only depends on the time step. As before, we will omit the system-state parameter $\mathbf{S}$ when is it is the same for all strategies under discussion.

THEOREM 4.1. $F_{\text{RATIOSWITCH}(h)}(h) = \Omega(F^*(h))$.

Combining lemma 3.3 with the proof of theorem 4.1, we obtain:

THEOREM 4.2. $F_{\text{GITTINSSWITCH}(h)}(h) = \Omega(F^*(h))$.

To the best of our knowledge, this is the first index for the finite horizon problem with provable approximation

guarantees. It would be interesting to obtain a smooth version of GittinsSwitch(h) which does not need to make the discrete jump from a discount factor of $1-2/h$ in the first half to playing the arm with the highest expected reward (i.e. to a discount factor of 0) in the second half.

4.2 An approximate index for the discount oblivious problem We will first establish a connection between the discount oblivious and finite horizon problems and then use this connection to obtain a simple index-based approximation algorithm for the discount oblivious problem.

Lemma 4.2. *For any κ, a strategy gives a κ-approximation simultaneously for all discount factor sequences Λ iff it gives a κ-approximation simultaneously to the fixed horizon problems with all horizons $h \geq 0$.*

Let RatioScale be the following discount oblivious strategy: play in sequence the strategies RatioSwitch$(1, \mathbf{S_0})$, RatioSwitch$(2, \mathbf{S_1})$, RatioSwitch$(4, \mathbf{S_3})$, RatioSwitch$(8, \mathbf{S_7}), \ldots$, where each RatioSwitch$(2^k)$ is started from the state of the system after time $2^k - 1$, denoted $\mathbf{S_{2^k-1}}$; this is the state in which the arms are left by the previous RatioSwitch strategy. $\mathbf{S_0}$ is the initial state of the system.

Like RatioSwitch, RatioScale is also an index-based strategy; the index used at any time step t depends only on t. Analogously, GittinsScale plays the sequence GittinsSwitch$(1, \mathbf{S_0})$, GittinsSwitch$(2, \mathbf{S_1})$, GittinsSwitch$(4, \mathbf{S_3})$, GittinsSwitch$(8, \mathbf{S_7}), \ldots$.

Since the state of the system at the start of RatioSwitch(2^i) depends on the outcomes of the previous steps, the following technical lemma, which is an easy consequence of the Martingale property, will be useful. This lemma states that performing an arbitrary sequence of extra explorations at the beginning cannot hurt the optimum solution for the budgeted learning problem. Observe that the state $\mathbf{T}$ is itself a random variable in this lemma; the expectation is over all values of $\mathbf{T}$.

Lemma 4.3. *Let π_1 be any arbitrary finite sequence of explorations starting from system-state $\mathbf{S}$. Let $\mathbf{T}$ be the system-state at the end of π_1. Let π_2 be an optimum h step strategy for the budgeted learning problem starting from the system-state $\mathbf{T}$. Then $\mathbf{E}[B_{\pi_2}(h, \mathbf{T})] \geq B^*(h, \mathbf{S})$.*

Using the above lemma, we can show the following:

Lemma 4.4. *For any positive integer $h \geq 1$, the expected reward of the discount oblivious strategy* RatioScale *in the first h steps is $\Omega(F^*(h, \mathbf{S_0}))$.*

Lemma 4.5. *For any positive integer $h \geq 1$, the expected reward of the discount oblivious strategy* GittinsScale *in the first h steps is $\Omega(F^*(h, \mathbf{S_0}))$.*

Invoking lemma 4.2 now gives us:

Theorem 4.3. *Strategies* RatioScale *and* GittinsScale *both give a constant factor approximation to the multi-armed bandit problem simultaneously for all discount factor sequences Λ.*

5 Computing the Ratio Index

We will now sketch how the ratio index can be computed. In the process, we will also get several useful insights into its structural properties. Given a single bandit-arm i, an initial state ρ for i, an exploration budget of h, and a state space T_i truncated to depth h, we view T_i as a layered DAG of depth h, which is to say that for any arm-state, u, in layer j, if $P_{uv} > 0$, then v must be in layer $j+1$. As explained in section 5.1, this is without loss of generality. We let Σ be the number of nodes in the layered DAG. Additionally, for any state u in T_i, we use T_i^u to denote the sub-DAG of T_i with root u; thus $T_i^\rho = T_i$.

For the purposes of this section, we require the use of *randomized single arm policies*. Whereas a *deterministic* single arm policy (corresponding to arm-state v) will always either explore v, exploit v, or abandon with probability 1, a randomized policy, π, selects $e_v, p_v : e_v, p_v \geq 0, e_v + p_v \leq 1$ where e_v represents the probability π explores in this state, p_v represents the probability π exploits in this state, and $1 - e_v - p_v$ represents the probability π abandons in this state. The vectors x^π and z^π are defined for randomized policies as for deterministic policies, as are the profit $\mathcal{P}(\pi)$ and cost $\mathcal{C}(\pi)$ of the policy. Our approach below will calculate the ratio index $r(u, h)$ for all $u \in T_i$ as well as the entire *profit curve* $\mathcal{P}_u(\cdot)$ for all u where $\mathcal{P}_u(\mathcal{C}_u) = \max_\pi \mathcal{P}(\pi)$ where the max is over all randomized single arm policies π with initial state u and $\mathcal{C}(\pi) \leq \mathcal{C}_u$. We show that there exists a deterministic policy that induces the maximum $\mathcal{P}_u(\mathcal{C}_u)/\mathcal{C}_u$ over all $\mathcal{C}_u > 0$ (and in fact our algorithm will find such a policy). Thus, the value $\max \mathcal{P}_u(\mathcal{C}_u)/\mathcal{C}_u$ is the ratio index for u given h, i.e., $r(u, h)$. Our algorithm relies heavily upon the following theorem on the structure of the profit curve.

Theorem 5.1. *The profit curve, $\mathcal{P}_u(\cdot)$, for any given state u is concave and piecewise linear with at most $2\Sigma_u$ segments where Σ_u represents the number of states in T_i^u.*

The proof of this theorem involves several steps and is deferred to the full version [17]. Towards proving the theorem, we show that as the budget increases along the profit curve for u, a *monotonicity property* holds that for every state $v \in T_i^u$, both p_v and $e_v + p_v$ are non-decreasing.

LEMMA 5.1. *For any $C^{(1)}, C^{(2)}$, with $C^{(2)} > C^{(1)}$, there exist optimal solutions $\langle e^{(1)}, p^{(1)} \rangle$ and $\langle e^{(2)}, p^{(2)} \rangle$ to $LP_u(C^{(1)})$ and $LP_u(C^{(2)})$ respectively such that $p_v^{(1)} \leq p_v^{(2)}$ and $e_v^{(1)} + p_v^{(1)} \leq e_v^{(2)} + p_v^{(2)}$ for all v in T_i^u.*

We further characterize the intersection of line segments of the profit curve as "corner" solutions and show that at these points $p_v \in \{0, 1\}$ and $e_v \in \{0, 1\}$ for *all* states in T_i^u. Thus, these points of the curve are induced by deterministic policies. Thus, the policy which induces the "corner" solution at the end of the first segment of the profit curve is a deterministic ratio index policy.

5.1 Algorithm for Computing the Profit Curve

The algorithm for computing the profit curve (and hence the ratio index) involves recursively calculating the profit curve for a state u given the profit curves for all of its successor states. We begin by constructing an *exploration profit curve* for u, $\mathcal{X}_u(\cdot)$, which denotes the optimal profit for any given cost conditioned on the fact that we are exploring at u (i.e. $e_u = 1$). We then take the concave envelope over this curve combined with the abandonment policy and the exploitation policy.

Superficially, it might seem that the number of segments of the profit curves could increase exponentially as we perform this process up the DAG. However, Theorem 5.1 guarantees that the number of segments remains bounded and the entire curve for u can be computed in time $O(d\Sigma_u \log \Sigma_u)$ given the successor curves, where d represents the maximum number of immediate descendants for any node. Thus, this algorithm is strongly polynomial (in Σ) for computing the entire profit curve of a state in the layered DAG, and hence, the ratio index. If the underlying state space of the bandit-arm is an unlayered DAG, we can make it layered by multiplying the number of states by at most Σ, so the algorithm is still strongly polynomial in Σ. If the underlying state space is not a DAG, we can convert it into a layered DAG by multiplying the number of states by at most h. Details of the algorithm and the analysis are in the full version [17].

Acknowledgements

The authors would like to thank Rajat Bhattacharjee and Sudipto Guha for helpful discussions, as well as anonymous referees for pointing us to the citations [24, 22, 27].

References

[1] P. Bhattacharya, L. Georgiadis, and P. Tsoucas. Extended Polymatroids, Properties, and Optimization. *Integer Programming and Combinatorial Optimization, IPCO2*, ed. E. Bala, G Cornnejols, and R. Kannan, Carnegie-Mellon University 298-315, 1992.

[2] R. Bellman. A Problem in the Sequential Design of Experiments. *Sankhia*, 16:221-229, 1956.

[3] D. Bertsimas and J. Nino-Mora. Conservation Laws, Extended Polymatroids and Multi-armed Bandit Problems. *Mathematics of Operations Research*, 21:257-306, 1996.

[4] D. Bertsimas and J. Nino-Mora. Restless Bandits, Linear Programming Relaxations, and a Primal-Dual Index Heuristic. *Operations Research*, 48:80-90, 2000.

[5] A. Blum and Y. Mansour. Learning, Regret Minimization, and Equilibria. *Algorithmic Game Theory*, ed. N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani. 79-102, 2007.

[6] M. Charikar, C. Chekuri, and M. Pal. Sampling Bounds for Stochastic Optimization. *APPROX-RANDOM* 257-269, 2005.

[7] B. C. Dean, M. X. Goemans, and J. Vondrak. Approximation the Stochastic Knapsack Problem: The Benefit of Adaptivity. *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, 208-217, 2004.

[8] B. C. Dean, M. X. Goemans, and J. Vondrak. Adaptivity and Approximation for Stochastic Packing Problems. *Porc. 16th ACM-SIAM Symp. on Discrete Algorithms*, 395-404, 2005.

[9] E. Frostig and G. Weiss. Four Proofs of Gittins' Multiarmed Bandit Theorem. *Applied Probability Trust*, November 10, 1999.

[10] J.C. Gittins and D. M. Jones. A Dynamic Allocation Index for the Sequential Design of Experiments. *Progress in Statistics: European Meeting of Statisticians, Budapest, 1972*, ed. J. Ganu, K. Sarkadi, and I. Vince. 241-266, 1974.

[11] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *J Royal Statistical Societe Series B*, 14:148-167, 1979.

[12] J. C. Gittins. *Multiarmed Bandits Allocation Indices*, Wiley, New York, 1989.

[13] K. D. Glazebrook and R. Garbe. Almost Optimal Policies for Stochastic Systems which Almost Satisfy Conservation Laws. *Annals of Operations Research*, 92:19-43, 1999.

[14] K. D. Glazebrook and D. J. Wilkinson. Index-Based Policies for Discounted Multi-armed Bandits on Parallel Machines. *The Annals of Applied Probability*, 10:3:877-896, 2000.

[15] A. Goel, S. Guha, and K. Munagala. Asking the Right Questions: Model-driven Optimization Using Probes. *Proc. ACM Symp. on Principles of Database Systems*, 2006.

[16] A. Goel and P. Indyk. Stochastic Load Balancing and Related Problems. *Proc. Symp. on Foundations of Computer Science*, 1999.

[17] A. Goel, S. Khanna, and B. Null. The Ratio Index for Budgeted Learning, with Applications. *Math arXiv:0810.0558v1,* `http://arxiv.org/abs/0810.0558` .

[18] S. Guha and K. Munagala. Model Driven Optimization Using Adaptive Probes. *Proc. ACM-SIAM Symp. on Discrete Algorithms*, 2007.

[19] S. Guha and K. Munagala. Approximation Algorithms for Budgeted Learning Problems. *STOC*, 2007.

[20] J. Kleinberg, Y. Rabini, and E. Tardos. Allocation Bandwidth for Bursty Connections. *SIAM J. Comput* 30(1), 2000.

[21] D. Luenberger. *Linear and Nonlinear Programming*, Reading, MA, 1984.

[22] O. Madani, D. Lizotte, and R. Greiner. Active Model Selection. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 357-365, 2004.

[23] P. Rusmevichientong and D. Williamson. An Adaptive Algorithm for Selecting Profitable Keywords for Search-based Advertising Services. *Proceedings of the 7th ACM conference on Electronic commerce,* 260-269, 2006.

[24] J. Schneider and A. Moore. Active Learning in Discrete Input Spaces. *Proceedings of the 34th Interface Symposium*, 2002.

[25] D. Shmoys and C. Swamy. Stochastic Optimization is (Almost) as Easy as Discrete Optimization. *Proc. 45th Symp. on Foundations of Computer Science* 228-237, 2004.

[26] P. Tsoucas. The Region of Achievable Performance in a Model of Klimov. Research Report RC16543, IBM T.J. Watson Research Center, Yorktown Heights, New York, 1991.

[27] A. Wald. *Sequential Analysis*, J. Wiley & Sons, New York, 212p, 1947.

[28] P. Whittle. Restless Bandits: Activity Allocation in a Changing World. *A Celebration of Applied Probability. J. Applied Probability*, 25A:287-298, 1988.