

# The Profit Curve for Budgeted Learning: Properties and Computation

Brad Null \*  
Stanford University

Ashish Goel †  
Stanford University

June 30, 2008

## Abstract

In the budgeted learning problem we 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. The ratio index, introduced by Goel et al. [5], leads to an index-based constant factor approximation algorithm for this problem. 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 value has been highlighted by the famous Gittins index for the discounted multi-armed bandit problem.

In this paper, we define the profit curve for budgeted learning, which is critical to calculating the ratio index. In the process, we derive several structural properties for the profit curve with significant consequences. Among these properties, we show that the profit curve is concave and piecewise linear, which, among other things, makes calculation of the ratio index using the profit curve elementary. We also show that the policies which induce the profit curve possess a certain monotonicity property critical to the proof that a greedy algorithm for budgeted learning with respect to the ratio index is a constant factor approximation algorithm. This monotonicity property also leads directly to a proof that the profit curve and the ratio index can be computed in strongly polynomial time. Thus, the greedy algorithm can be executed in strongly polynomial time, as can an index-based approximation algorithm for a discount oblivious version of the multi-armed bandit problem. Further, the techniques we develop may also be useful for deriving strongly polynomial algorithms for other Markov Decision Problems. We conclude with a detailed algorithm for computing the profit curve.

---

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

†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 an Alfred P. Sloan fellowship. Email: ashishg@stanford.edu .

# 1 Introduction

The classical multi-armed bandit problem provides an invaluable 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<sup>1</sup>). 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. The user wishes to maximize the total expected discounted reward over time.

This problem was first considered by Bellman (1956) [1] and became especially popular after the presentation of the Gittins index in the 1970’s, which provides an elegant and often easily implementable optimal index-based approach to this problem ([2], [3], [4]).

One typical assumption on the bandit-arms is that they have  $(\alpha, \beta)$ -priors: the success probability of an  $(\alpha, \beta)$ -bandit-arm is  $\alpha/(\alpha + \beta)$ ; in case of a success a reward of 1 is obtained and  $\alpha$  gets incremented, whereas in case of a failure no reward is obtained and  $\beta$  gets incremented. 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*. 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 at most  $h$  steps of exploration. Although the classical multi-armed bandit problem has received significantly more attention (likely due in large part to the attractiveness of the Gittins index and the NP-hardness of the budgeted learning problem [6]); in many applications the budgeted learning problem is often the more appropriate model. For example, Madani et al ([9], [7]) model the selection of treatments in clinical trials under a fixed \$2 million budget as a budgeted learning problem.

---

<sup>1</sup>We will use the terms experimentation and exploration interchangeably in this paper, depending on the context.

Further, it has been shown that a constant factor approximation algorithm for the budgeted learning problem can be modified to yield constant factor approximation algorithms for a wide range of interesting and non-trivial variations of the multi-armed bandit problem, including the finite horizon multi-armed bandit problem and the discount oblivious multi-armed bandit problem [5]. In the latter of these problems, the input is as in the classical multi-armed bandit problem, except that the discount rate need not be constant and additionally the user is oblivious to this discount rate term structure. For this problem, the approximation algorithm is not only constant factor competitive to any other oblivious algorithm, but constant-factor competitive to optimal strategies that are fully aware of this term structure as well. We also believe that this line of research will yield constant factor approximation algorithms for several other useful variants of the multi-armed bandit problem.

Goel et al. [5] introduced the ratio index for this problem and an associated greedy index-based policy, which leads to such a constant factor approximation for the budgeted learning problem.<sup>2</sup> 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, and it has been shown that the ratio index and the Gittins index are in fact constant factor approximations of each other [5].

## 1.1 Our contribution

In this paper, we present the profit curve, a function which is critical to calculating the ratio index. We develop two formulations by which a point on the profit curve can be computed (a dynamic program and a linear program), and upon analyzing these formulations, we derive several interesting structural properties for the profit curve. Among these properties, we show that the profit curve is concave and piecewise linear, which, among other things, makes calculation of the ratio index using the profit curve elementary. We also show that the profit curve possesses a certain monotonicity property which enables us to show that a greedy algorithm for budgeted learning with respect to the ratio index is a constant factor approximation algorithm. This monotonicity property also proves that both the profit curve and 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

---

<sup>2</sup>It should be noted that Guha and Munagala [6] also present a (non-index based) constant factor approximation for this problem.

if the state space is acyclic, and strongly polynomial in the size of the state space and  $h$  if the state space is general. Thus, not only can a greedy index-based constant factor approximation algorithm for the budgeted learning problem be implemented in strongly polynomial time, but so can an index-based constant factor approximation algorithm for the discount oblivious multi-armed bandit problem (see [5] for details of this algorithm).

Our proofs of several of these properties involve recursively analyzing the basic feasible solutions of the underlying linear program for computing points on the profit curve and using the structure of the basic feasible solutions to prove that these strategies have a simple form. These LPs can be viewed as modeling a Markov decision process. It is conceivable that the techniques we use may offer insight towards obtaining strongly polynomial algorithms for more general classes of Markov Decision Problems (see [10] for a description; the algorithm in [10] is not strongly polynomial in the discount factor so the problem is still open when discount factors are arbitrarily close to 1). We conclude with a detailed algorithm for computing the profit curve.

## 1.2 Organization

In section 2, we define the budgeted learning problem, the profit curve, and the ratio index. Section 3 then lays out the main properties of the profit curve and their implications, and section 4 presents detailed proofs of these properties. In section 5, we present a strongly polynomial algorithm to compute the profit curve and the ratio index.

# 2 The Budgeted Learning Problem, the Profit Curve 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  $\rho_i$ . 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,  $\pi^*$ , so that  $\mathbf{E}_{\pi^*}[\max_{i \in \{1, \dots, n\}} \zeta(v_i)]$  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.

At any given time, the *current state* of the system is denoted by  $\mathbf{S} = \{u_1, u_2, \dots, u_n, \delta\}$ , which captures the current states of all the arms, and the budget left (i.e. the number of experimentation steps that are still remaining),  $\delta$ . 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*  $\pi$  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  $\delta = 0$ . 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. 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. Though, it is easy to see that randomized strategies can not do any better than deterministic strategies, for our purposes, we require the use of *randomized single arm policies*. Whereas a *deterministic* single arm policy (corresponding to arm-state  $u$ ) will always either explore  $u$ , exploit  $u$ , or

abandon with probability 1, a randomized policy,  $\pi$ , selects  $e_u, p_u : e_u, p_u \geq 0, e_u + p_u \leq 1$  where  $e_u$  represents the probability  $\pi$  explores in this state,  $p_u$  represents the probability  $\pi$  exploits in this state, and  $1 - e_u - p_u$  represents the probability  $\pi$  abandons in this state.

## 2.2 The profit curve and the ratio index

We now introduce two vectors  $x^\pi$  and  $z^\pi$  corresponding to a randomized single arm policy  $\pi$  (on any given arm  $i$ ). The probability that arm-state  $u (\in T_i)$  is the final exploited state by policy  $\pi$  is given by  $x_u^\pi$ . The probability that arm-state  $u$  is explored by policy  $\pi$  is given by  $z_u^\pi$ . We define the *cost* of policy  $\pi$  as

$$\mathcal{C}(\pi) = \frac{\sum_{u \in T_i} z_u^\pi}{h} + \sum_{u \in T_i} x_u^\pi.$$

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

$$\mathcal{P}(\pi) = \sum_{u \in T_i} x_u^\pi \zeta(u).$$

We are now ready to define the *profit curve* and *ratio index* for arm  $i$  at state  $u$ .

**Profit Curve.** The profit curve  $\mathcal{P}_u(\cdot)$  of a bandit-arm (say arm  $i$ ) in initial state  $u$  and with experimentation budget  $h$ , is defined as a function over all  $\mathcal{C}_u \geq 0$  where

$$\mathcal{P}_u(\mathcal{C}_u) = \max_{\pi} \mathcal{P}(\pi)$$

where the max is over all randomized single arm policies  $\pi$  which have initial arm-state  $u$ , budget  $h$ , state space  $T_i$ , and cost  $\mathcal{C}(\pi) \leq \mathcal{C}_u$ .

**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 randomized single arm policies  $\pi$  which have initial arm-state  $u$ , budget  $h$ , state

space  $T_i$ , and cost  $\mathcal{C}(\pi) > 0$ .

It is easy to see then that the ratio index will be equivalent to the maximum slope on the profit curve. We refer to a policy which yields the ratio index as a ratio index policy for state  $u$ , denoted  $\pi_r(u, h)$ .

### 3 Main Properties of the Profit Curve and their Implications

In section 4 we derive several properties of the profit curve and ratio index. As this section can become quite technical, we present the more notable implications here.

*The profit curve is concave and piecewise linear, with each “corner” solution representing a deterministic policy.* In Lemmas 4.1 and 4.4 we show that the profit curve for a state  $u$  is concave and piecewise linear (with  $\mathcal{P}_u(0) = 0$ ), thus implying that the ratio index is simply the slope of the first line segment of the profit curve. Further, lemma 4.4 characterizes 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  $T_i$ . 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.**

*The ratio index policy for an arm-state will not abandon any descendant arm-state with higher ratio index, nor will it explore or exploit any descendant arm-state with lower ratio index.* The central proof of [5] shows that the following persistent algorithm ( $\mathbf{G}'$ ) is constant factor (.22) optimal for the budgeted learning problem.

**The persistent algorithm  $\mathbf{G}'$ :** 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 currently highest ratio index.

Corollary 4.6 shows that 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  $r(v, h) < r(u, h)$ . Using this

structural result on the ratio index, we can deduce that the series of experiments in  $\mathbf{G}'$  is a prefix of that in the Greedy algorithm ( $\mathbf{G}$ ) below (see [5] for proof details).

**The Greedy Algorithm  $\mathbf{G}$ :** Suppose the initial experimentation budget is  $h$ , and the current system-state is given by  $\mathbf{S} = \{u_1, u_2, \dots, u_n, \delta\}$ . 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  $\zeta(u_i)$ .

Thus, since additional experimentation cannot hurt (by the Martingale property), **the greedy algorithm is constant factor (.22) optimal for the budgeted learning problem.**

*As the budget increases along the profit curve for  $u$ , for every state  $v \in T_i$ , both  $p_v$  and  $e_v + p_v$  are non-decreasing.* We show in lemma 4.7 that as the budget increases along the profit curve for  $u$ , a *monotonicity property* holds that for every state  $v \in T_i$ , both  $p_v$  and  $e_v + p_v$  are non-decreasing. This result directly implies theorem 4.8, which shows that **the profit curve for state  $u$  can have at most  $2\Sigma_u$  segments.** This implies (as we will see in section 5) that **the profit curve can be computed in strongly polynomial time.** As mentioned in section 1.1 this further implies that the greedy algorithm above is strongly polynomial, as is an index-based approximation algorithm for the discount oblivious multi-armed bandit problem.

## 4 Derivation of the Properties of the Profit Curve

In what follows, we will only be concerned with a single bandit-arm  $i$ , an initial state  $\rho$  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 later, this is without loss of generality. We let  $\Sigma$  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  $T_i^\rho = T_i$ .

We begin by considering two methods of calculating  $\mathcal{P}_u(\mathcal{C}_u)$  that will be used in our discussions. The first is a recursive equation ( $\text{RE}_u(\mathcal{C}_u)$ ) that can be used to calculate  $\mathcal{P}_u(\mathcal{C}_u)$  for a given state  $u$  and budget  $\mathcal{C}_u$  provided that we have the entire profit curves of all successor states. This equation is

$$\mathcal{P}_u(\mathcal{C}_u) = \max_{p_u, e_u, E^u} p_u \zeta(u) + e_u \sum_{v \in D(u)} P_{uv} \mathcal{P}_v(E_v^u)$$

the constraints are:

$$p_u + e_u \leq 1$$

$$p_u + e_u(1/h + \sum_{v \in D(u)} P_{uv} E_v^u) \leq \mathcal{C}_u$$

$$p_u, e_u, E^u \geq 0$$

$D(u)$  is the set of immediate descendants of  $u$ . The decision variables  $p_u$  and  $e_u$  represent the probability of exploiting and exploring in  $u$  respectively (as in the definition of a randomized policy). The vector  $E^u$  represents the budgets we would allocate to each of the immediate descendants of  $u$  should we visit them. Recall that  $P_{uv}$  is the probability of transitioning to state  $v$  given we are experimenting in state  $u$ . We assume  $E_v^u = 0$  if  $v \notin D(u)$ .

Alternatively, the following LP ( $LP_u(\mathcal{C}_u)$ ) (similar to the one in [6]) for a given state  $u$  of bandit-arm  $i$  reveals a policy  $\langle w, x, z \rangle$ , which induces  $\mathcal{P}_u(\mathcal{C}_u)$  for a given  $\mathcal{C}_u$ .

$$\begin{aligned} & \max_{w, x, z} \sum_{v \in T_i^u} x_v \zeta(v) \\ \text{s.t.} & \frac{\sum_{v \in T_i^u} z_v}{h} + \sum_{v \in T_i^u} x_v \leq \mathcal{C}_u \\ & \sum_{y: v \in D(y)} z_y P_{yv} = w_v \quad \forall v \in T_i^u \setminus \{u\} \\ & w_u = 1 \\ & x_v + z_v \leq w_v \quad \forall v \in T_i^u \\ & x_v, z_v \geq 0 \quad \forall v \in T_i^u \end{aligned}$$

For any state  $v \in T_i^u$ ,  $w_v$  represents the probability the bandit-arm enters  $v$ ,  $x_v$  represents the unconditional probability of exploiting in  $v$ , and  $z_v$  represents the unconditional probability of experimenting in  $v$ . Given the stochastic nature of the  $P$  matrix and the fact that all of the  $z$  values are less than or equal to their corresponding  $w$  values, we can see that each element of  $w$  will be bounded between 0 and 1. Thus,  $x$  and  $z$  will also be automatically bounded above by 1. Thus, we do not need further constraints bounding these

variables.

#### 4.1 Some basic properties of the profit curve

Using both the recursive equation and the LP, we can show the following.

**Lemma 4.1**  $\mathcal{P}_u(\cdot)$  is a concave, nondecreasing function and if  $\mathcal{C}_u \geq 1$ ,  $\mathcal{P}_u(\mathcal{C}_u) = \mathcal{P}_u(1) = \zeta(u)$ .

**Proof:** Given any  $\mathcal{C}_u$  and associated policy  $\{p_u, e_u, E^u\}$  for  $\text{RE}_u(\mathcal{C}_u)$  that realizes  $\mathcal{P}_u(\mathcal{C}_u)$ , observe first that this policy is also feasible for any cost greater than or equal to  $\mathcal{C}_u$ . Thus,  $\mathcal{P}_u(\mathcal{C}_u)$  is a nondecreasing function.

We prove the remainder of the lemma by induction. Looking at  $\mathcal{P}_u(\mathcal{C}_u)$  for a state  $u$  that is at depth  $h$  it is easy to see that,  $\mathcal{P}_u(\mathcal{C}_u) = \mathcal{C}_u * \zeta(u)$  if  $\mathcal{C}_u \leq 1$  and  $\zeta(u)$  if  $\mathcal{C}_u > 1$ . Thus, at depth  $h$  if  $\mathcal{C}_u \geq 1$ ,  $\mathcal{P}_u(\mathcal{C}_u) = \mathcal{P}_u(1) = \zeta(u)$  and  $\mathcal{P}_u(\cdot)$  is concave.

Now assume these properties hold for all states at depth  $i + 1$  and look at a state  $u$  at depth  $i$ . Our profit is obviously non-decreasing in each of the decision variables  $\{p_u, e_u, E^u\}$ . Set  $E_v^u = 1 \forall v \in D(u)$ . From the induction hypothesis,  $\sum_{v \in D(u)} P_{uv} \mathcal{P}_v(1) = \sum_{v \in D(u)} P_{uv} \zeta(v)$  and by the martingale property,  $\sum_{v \in D(u)} P_{uv} \zeta(v) = \zeta(u)$ , so our objective becomes  $\max p_u \zeta(u) + e_u \zeta(u)$  or equivalently  $\max(p_u + e_u) \zeta(u)$ . Since  $(p_u + e_u) \leq 1$ , this can be no larger than  $\zeta(u)$ . But clearly  $\zeta(u)$  can be achieved with a cost of 1 by setting  $p_u = 1$  and  $e_u = 0$ . Henceforth,  $\mathcal{P}_u(\mathcal{C}_u) = \zeta(u) \forall \mathcal{C}_u \geq 1$ .

With respect to concavity, for any two points on the profit curve of  $u$ ,  $\mathcal{P}_u(C^{(1)})$  and  $\mathcal{P}_u(C^{(2)})$ , corresponding to  $C^{(1)} < C^{(2)}$  let  $\langle w^{(1)}, x^{(1)}, z^{(1)} \rangle$  and  $\langle w^{(2)}, x^{(2)}, z^{(2)} \rangle$  be associated policies respectively as defined on  $LP_u(\mathcal{C}_u)$ . Consider the policy  $\langle \frac{w^{(1)}+w^{(2)}}{2}, \frac{x^{(1)}+x^{(2)}}{2}, \frac{z^{(1)}+z^{(2)}}{2} \rangle$ . This policy is feasible for the problem of finding the profit associated with budget  $\frac{C^{(1)}+C^{(2)}}{2}$ , so  $\mathcal{P}_u(\frac{C^{(1)}+C^{(2)}}{2}) \geq \sum_{v \in T_i^u} \frac{x_v^{(1)}+x_v^{(2)}}{2} \zeta(v) = \frac{\mathcal{P}_u(C^{(1)})+\mathcal{P}_u(C^{(2)})}{2}$ . Thus the profit curve for  $u$  is concave. ■

With respect to  $LP_u(\mathcal{C}_u)$ , we define the vectors  $e$  and  $p$  where  $e_v = z_v/w_v$  and  $p_v = x_v/w_v$  if  $w_v > 0$ . Otherwise, we require only that  $e_v, p_v \geq 0$  and  $e_v + p_v \leq 1$ . Thus  $e_v$  and  $p_v$  are the conditional probability of exploring and exploiting, respectively, given that we are in state  $v$ . Thus, the two vectors  $\langle e, p \rangle$  define a randomized policy that induces a point on  $\mathcal{P}_u(\cdot)$ . Alternatively, we could define the same policy with the

three vectors  $\langle w, x, z \rangle$ . In what follows, we will freely interchange between the two notations. Note that the one thing we must be careful to observe is that for any policy and state where  $w_v = 0$ , there are infinitely many equivalent assignments of  $e_v$  and  $p_v$ .

We will now appeal to linear programming theory with respect to  $LP_u(\mathcal{C}_u)$  to derive several important properties of the profit curve. We begin with the following:

**Lemma 4.2** *For any  $\mathcal{C}_u \in [0, 1]$  there exists an optimal policy with respect to  $LP_u(\mathcal{C}_u)$  such that  $p_v \in \{0, 1\}$  and  $e_v \in \{0, 1\}$  for all but at most one state in  $T_i^u$ .*

**Proof:** Let's consider a basic feasible optimal solution to  $LP_u(\mathcal{C}_u)$ ,  $\langle w^*, x^*, z^* \rangle$ . We know such a solution exists since the LP is bounded and a basic feasible solution to the LP exists (for instance the solution that corresponds to setting  $x_v = z_v = 0 \forall v$ ). Let us create  $\hat{T}_i^u$  by removing all states  $v$  from  $T_i^u$  for which  $x_v^* = z_v^* = 0$ . (Note: Since we may be removing some children of states remaining in  $\hat{T}_i^u$ , the martingale property may no longer hold with respect to  $\hat{T}_i^u$ .) Let us create  $\hat{L}P_u(\mathcal{C}_u)$  by replacing  $T_i^u$  in  $LP_u(\mathcal{C}_u)$  with  $\hat{T}_i^u$ .  $\hat{L}P_u(\mathcal{C}_u)$  will have the same optimal objective value as  $LP_u(\mathcal{C}_u)$  and an optimal solution  $\langle \hat{w}^*, \hat{x}^*, \hat{z}^* \rangle$  such that  $\hat{w}_v^* = w_v^*$ ,  $\hat{x}_v^* = x_v^*$  and  $\hat{z}_v^* = z_v^*$  for every state  $v \in \hat{T}_i^u$ .

Let us define the number of states in  $\hat{T}_i^u$  as  $\hat{\Sigma}_u$ .  $\hat{L}P_u(\mathcal{C}_u)$  has  $3\hat{\Sigma}_u$  variables,  $2\hat{\Sigma}_u$  non-negativity constraints, and  $2\hat{\Sigma}_u + 1$  other constraints. Thus  $\hat{L}P_u(\mathcal{C}_u)$  has a basic feasible optimal solution, which will have at least  $\hat{\Sigma}_u - 1$  variables equal to zero. (For discussion of LP theory and the role of basic feasible solutions, see [8] or a similar resource.)

If exactly  $\hat{\Sigma}_u - 1$  variables are equal to zero, all constraints of the type  $x_v + z_v \leq w_v$  must be tight. By virtue of how we created  $\hat{T}_i^u$ , we know that each of these zero variables must correspond to a distinct state. Thus, for these  $\hat{\Sigma}_u - 1$  states either  $\hat{x}_v^* = \hat{w}_v^* > 0$  or  $\hat{z}_v^* = \hat{w}_v^* > 0$ , and for the only remaining state, (call it  $y$ ),  $\hat{x}_y^* > 0$  and  $\hat{z}_y^* > 0$ .

Alternatively there could be  $\hat{\Sigma}_u$  variables equal to zero (but no more since for all states  $\hat{x}_v^* + \hat{z}_v^* > 0$ ), in which case, for at least  $\hat{\Sigma}_u - 1$  states either  $\hat{x}_v^* = \hat{w}_v^* > 0$  or  $\hat{z}_v^* = \hat{w}_v^* > 0$ .

Looking at this in terms of  $\hat{p}$  and  $\hat{e}$ , in at least  $\hat{\Sigma}_u - 1$  states either  $\hat{p}_v^* = 1$  or  $\hat{e}_v^* = 1$ . For all states  $v$  in  $T_i^u/\hat{T}_i^u$  we could arbitrarily assign  $p_v$  and  $e_v$ , so this property holds with respect to  $LP_u(\mathcal{C}_u)$  as well. ■

From the analysis above, we can see that for a given  $\mathcal{C}_u$ ,  $LP_u(\mathcal{C}_u)$  has an optimal basic feasible solution  $\langle w^*, x^*, z^* \rangle$  that takes one of the following three forms:

1. For every state  $v$  where  $x_v^* + z_v^* > 0$ ,  $x_v^* + z_v^* = w_v^*$ , and in exactly one state  $y$ ,  $x_y^* > 0$  and  $z_y^* > 0$ .  
In this case exactly  $3\hat{\Sigma}_u$  constraints of  $L\hat{P}_u(\mathcal{C}_u)$  are binding (including the cost constraint).
2. For every state  $v$  where  $x_v^* + z_v^* > 0$ , either  $x_v^* = 0$  or  $z_v^* = 0$ , and in exactly one state  $y$ ,  $x_y^* + z_y^* < w_y^*$ .  
In this case exactly  $3\hat{\Sigma}_u$  constraints of  $L\hat{P}_u(\mathcal{C}_u)$  are binding (including the cost constraint).
3. For every state  $v$  where  $x_v^* + z_v^* > 0$ , either  $x_v^* = 0$  or  $z_v^* = 0$ , and  $x_v^* + z_v^* = w_v^*$ . In this case either  $3\hat{\Sigma}_u + 1$  constraints of  $L\hat{P}_u(\mathcal{C}_u)$  are binding, or else the cost constraint is not binding which implies the slope of the profit curve at this point is zero.

In the first two types of policies, we will call the state  $y$  for which it does not hold that  $p_y^*, e_y^* \in \{0, 1\}$  the *transitional state*. Note that policy type 3 does not have a transitional state, and thus corresponds to a deterministic policy.

Further leveraging our understanding of the basic feasible solutions of  $LP_u()$  and linear programming theory, we establish the following pair of lemmas.

**Lemma 4.3** *Let  $\langle e^*, p^* \rangle$  be a basic feasible optimal policy for  $LP_u(\mathcal{C}_u)$ . If  $\langle e^*, p^* \rangle$  has a transitional state, then there exists an  $\epsilon > 0$  such that there exists an optimal policy  $\langle e^{*(1)}, p^{*(1)} \rangle$  for  $LP_u(\mathcal{C}_u^{(1)})$  where  $\mathcal{C}_u^{(1)} \in [\mathcal{C}_u - \epsilon, \mathcal{C}_u + \epsilon]$  that has the same transitional state  $y$  and for all states  $v \neq y$   $e_v^{*(1)} = e_v^*$  and  $p_v^{*(1)} = p_v^*$ .*

**Proof:** This result follows directly from linear programming theory. Given that we have a transitional state, we have a non-degenerate solution to  $L\hat{P}_u()$  (with exactly  $3\hat{\Sigma}_u$  constraints at equality). Thus, small changes in the budget will result in a solution that has the same set of binding constraints. These results carry over to  $LP_u()$ . (We again refer the reader to [8] or another suitable optimization text for more discussion of linear programming theory and sensitivity analysis.)

With respect to our problem, this implies that for every non-transitional state  $v$ , if  $x_v^* = 0$ ,  $x_v^{*(1)} = 0$ , if  $z_v^* = 0$ ,  $z_v^{*(1)} = 0$ , if  $x_v^* = w_v^*$ ,  $x_v^{*(1)} = w_v^{*(1)}$ , and if  $z_v^* = w_v^*$ ,  $z_v^{*(1)} = w_v^{*(1)}$ . Or equivalently,

$$e_v^{*(1)} = e_v^*, p_v^{*(1)} = p_v^*.$$

■

This leads naturally to the following result:

**Lemma 4.4** *The profit curve is piecewise linear. Further, each “corner” solution or point connecting two segments of the curve can be achieved by a deterministic policy.*

**Proof:** Again by linear programming theory, beginning at a non-degenerate solution to  $LP_u(\cdot)$  and making incremental changes to the budget will change the optimal objective value at a constant rate (the shadow price of the budget constraint) until a new constraint becomes binding. With respect to  $LP_u(\mathcal{C}_u)$  this implies that if the optimal solution to this LP has a transitional state, then there exist some  $C^{(1)} < \mathcal{C}_u$  and  $C^{(2)} > \mathcal{C}_u$  such that the segment of the profit curve from  $\mathcal{P}_u(C^{(1)})$  to  $\mathcal{P}_u(C^{(2)})$  is linear.

Furthermore, since we know that each end of this line segment must have an additional binding constraint, the optimal policy associated with these points on the curve must be of type 3 above (i.e. a deterministic policy).

■

This result, combined with the fact that profit curves are concave, implies that for any state,  $u$ , the ratio index can always be calculated with infinitesimal cost, i.e.  $r(u, h) = \mathcal{P}'_{+u}(0)$ , where  $\mathcal{P}'_{+u}(\mathcal{C}_u)$  denotes the right-sided derivative of the profit curve evaluated at  $\mathcal{C}_u$ . Further, the policy at the end of the first line segment of  $\mathcal{P}_u(\cdot)$  is a deterministic ratio index policy.

## 4.2 Monotonicity properties of the profit curve

We are now ready to begin establishing the monotonic properties of the profit curve. We begin by establishing the following four properties. In the following, recall that  $\mathcal{P}'_{+u}(\mathcal{C}_u)$  represents the right-sided derivative of  $\mathcal{P}_u(\cdot)$  evaluated at  $\mathcal{C}_u$ . Correspondingly,  $\mathcal{P}'_{-u}(\mathcal{C}_u)$  represents the left-sided derivative of  $\mathcal{P}_u(\cdot)$  evaluated at  $\mathcal{C}_u$ .

**Lemma 4.5** *There exists an optimal solution  $\langle e^*, p^* \rangle$  to  $LP_u(C)$ , such that for any state  $v$  where  $w_v^* > 0$ :*

1. *If  $p_v^* > 0$ ,  $\mathcal{P}'_{-v}(1) \geq \mathcal{P}'_{-u}(C)$*

2. If  $e_v^* > 0$ ,  $\mathcal{P}'_{+u}(C) \geq \mathcal{P}'_{-v}(1)$
3. If  $e_v^* > 0$ ,  $r(v, h) \geq \mathcal{P}'_{-u}(C)$
4. If  $1 - e_v^* - p_v^* > 0$ ,  $\mathcal{P}'_{+u}(C) \geq r(v, h)$

**Proof:** First, consider the set of all optimal solutions to  $LP_u(C)$ . Arbitrarily observe one optimal solution  $\langle \bar{e}, \bar{p} \rangle$ . Use  $\mathcal{S}$  to denote the set of all optimal solutions  $\langle e, p \rangle$  to  $LP_u(C)$  such that  $w_v = \bar{w}_v$ . Select  $\langle e^*, p^* \rangle \in \mathcal{S}$  such that  $p_v^* = \max_{\mathcal{S}} p_v$ ,  $p_v^* + e_v^* = \max_{\mathcal{S}} p_v + e_v$ . We now prove the first of these four properties. The remaining proofs are similar and can be found in appendix A.

1. If  $p_v^* > 0$ ,  $\mathcal{P}'_{-v}(1) \geq \mathcal{P}'_{-u}(C)$ : Given  $v$  such that  $w_v^* > 0$  and  $p_v^* > 0$  assume  $\mathcal{P}'_{-v}(1) < \mathcal{P}'_{-u}(C)$ . Denote by  $E_v^*$  the amount of budget devoted to  $v$  and its descendants. Thus, we know that total profit garnered from  $v$  and its descendants is  $\mathcal{P}_v(E_v^*)$ . Let  $\langle \tilde{e}, \tilde{p} \rangle_v$  be the policy that induces  $\mathcal{P}_v(E_v^* - \epsilon)$  on  $T_i^v$ . Let  $\langle \tilde{e}, \tilde{p} \rangle$  be a policy that follows  $\langle \tilde{e}, \tilde{p} \rangle_v$  in the subtree of  $v$  and follows  $\langle e^*, p^* \rangle$  otherwise. The policy  $\langle \tilde{e}, \tilde{p} \rangle$  is a feasible solution to the problem of finding the point on the profit curve of  $u$  with cost  $C - w_v^* \epsilon$ . This policy has profit  $\mathcal{P}_u(C) - w_v^* \epsilon \mathcal{P}'_{-v}(E_v^*)$ . Further, as  $p_v^* > 0$ , we know that either  $p_v^* = 1$  (in which case  $E_v^* = 1$ ) or else  $v$  is the transitional state for  $LP_v(E_v^*)$ . In the latter case, one end of the line segment of  $\mathcal{P}_v(\cdot)$  which contains  $\mathcal{P}_v(E_v^*)$  must have  $p_v^* = 1$ . This obviously corresponds to  $\mathcal{P}_v(1)$ . Thus,  $\mathcal{P}_v(E_v^*)$  must be on the last segment of the profit curve of  $v$ , so  $\mathcal{P}'_{-v}(E_v^*) = \mathcal{P}'_{-v}(1)$ . Thus,

$$\begin{aligned} \mathcal{P}_u(C - w_v^* \epsilon) &\geq \mathcal{P}_u(C) - w_v^* \epsilon \mathcal{P}'_{-v}(1) \\ &> \mathcal{P}_u(C) - w_v^* \epsilon \mathcal{P}'_{-u}(C) \end{aligned}$$

Thus,  $\mathcal{P}_u(C) - \mathcal{P}_u(C - w_v^* \epsilon) < w_v^* \epsilon \mathcal{P}'_{-u}(C)$ . By concavity, this cannot be true, so it must be true that  $\mathcal{P}'_{-v}(1) \geq \mathcal{P}'_{-u}(C)$ . ■

This result directly implies the following.

**Corollary 4.6** *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  $r(v, h) < r(u, h)$ .*

**Proof:** With respect to lemma 4.5, select  $C$  such that  $0 < C < \mathcal{C}(\pi_r(u, h))$ . Thus,  $\mathcal{P}'_{-u}(C) = \mathcal{P}'_{+u}(C) = r(u, h)$ . since  $\mathcal{P}'_{-v}(1) \leq r(v, h)$ , the first, third and fourth properties proved in lemma 4.5 respectively imply that for the optimal randomized policy corresponding to this point on the profit curve: (1) if  $r(u, h) > r(v, h)$ , then  $p_v^* = 0$ ; (3) if  $r(u, h) > r(v, h)$ , then  $e_v^* = 0$ ; and (4) if  $r(v, h) > r(u, h)$ , then  $1 - e_v^* - p_v^* = 0$ . As  $v$  cannot be a transitional state under any of these conditions, then by the argument in lemma 4.4 these properties must hold for the ratio index policy as well. ■

As mentioned in section 3, this result proves that the persistent algorithm  $\mathbf{G}'$  is a prefix of the greedy algorithm  $\mathbf{G}$ , and thus the greedy algorithm is a constant factor approximation algorithm for the budgeted learning problem as well. We are now ready to prove the following monotonicity result, which follows from the concavity of the profit curve.

**Lemma 4.7** *For any state  $u$  and  $C^{(1)}$  and  $C^{(2)}$ , where  $0 \leq C^{(1)} < C^{(2)} \leq 1$ , 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  $T_i^u$ .*

**Proof:** Given the four properties from lemma 4.5, we can see that for any state  $v$ , as the slope of  $\mathcal{P}_u(C)$  decreases with increasing  $C$ , there will be up to three regions where we would first abandon at  $v$ , then we would explore at  $v$ , then we would exploit. The only remaining technical issue that remains is if the slope of the profit curve of  $u$  exactly equals  $\mathcal{P}'_{-v}(1)$  or  $r(v)$ . In these cases, we are of necessity on a line segment of  $\mathcal{P}_u(\cdot)$  where  $v$  is the transitional state. In these cases, at the two endpoints of the segment we must have one of the three following pairs of values for  $v$ :

1.  $p_v = e_v = 0; p_v = 1$
2.  $p_v = e_v = 0; e_v = 1$
3.  $e_v = 1; p_v = 1$

For the first two of these cases, it is obvious that the state on the right corresponds to higher cost. Thus, our monotonicity property holds. For the third case, we know by the martingale property that setting  $p_v = 1$

yields maximum possible profit at the node. Thus, the monotonicity property must hold in this case as well. ■

We index the  $B_u$  “corner” solutions on  $\mathcal{P}_u(\mathcal{C}_u)$  as  $s_i, \dots, s_{B_u}$  corresponding to the budget associated with their rightmost end point. The above lemma further implies that if for every state  $v \in T_i^u$  and every “corner” solution,  $s_i$  on  $\mathcal{P}_u(\mathcal{C}_u)$  we apply a label  $L_{s_i}(v) \in \{A, E, P\}$  where  $A$  corresponds to “abandoning” at  $v$  ( $p_v = e_v = 0$ ),  $E$  corresponds to “exploring” ( $e_v = 1$ ), and  $P$  corresponds to “exploiting” ( $p_v = 1$ ), then for any  $v$  and any  $i, j$  such that  $i \leq j$ , if  $L_{s_i}(v) = P$ , then  $L_{s_j}(v) = P$ , and if  $L_{s_i}(v) = E$ , then  $L_{s_j}(v) \neq A$ . Thus, we can find a set of solutions such that as we increase the budget, once a state is labeled “P” it will always remain a “P” and once it becomes an “E” it will never become an “A”. Thus, each state can only change labels at most twice. As every successive “corner” solution must involve changing the label of at least one state, the profit curve for a state can have at most  $2\Sigma_u$  segments (where  $\Sigma_u$  represents the number of states in  $T_i^u$ ). This establishes our main theorem.

**Theorem 4.8** *The profit curve  $\mathcal{P}_u(\mathcal{C}_u)$  for any state  $u$  can have at most  $2\Sigma_u$  segments, where  $\Sigma_u$  is the number of states in  $T_i^u$ .*

In section 5 we present an algorithm that computes the profit curve of  $u$  given the profit curves of all successor states which is polynomial in the number of segments of successor profit curves. Thus, as we have established that the number of segments of the profit curve of any state  $v$  is bounded by  $\Sigma_v$ , then  $\mathcal{P}_u()$  can be computed in strongly polynomial time, as can  $\mathcal{P}_\rho()$ .

## 5 A Strongly Polynomial 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  $u$ ,  $\mathcal{X}_u()$ , 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. Figure 1 shows a typical example of the relationship between these two curves.

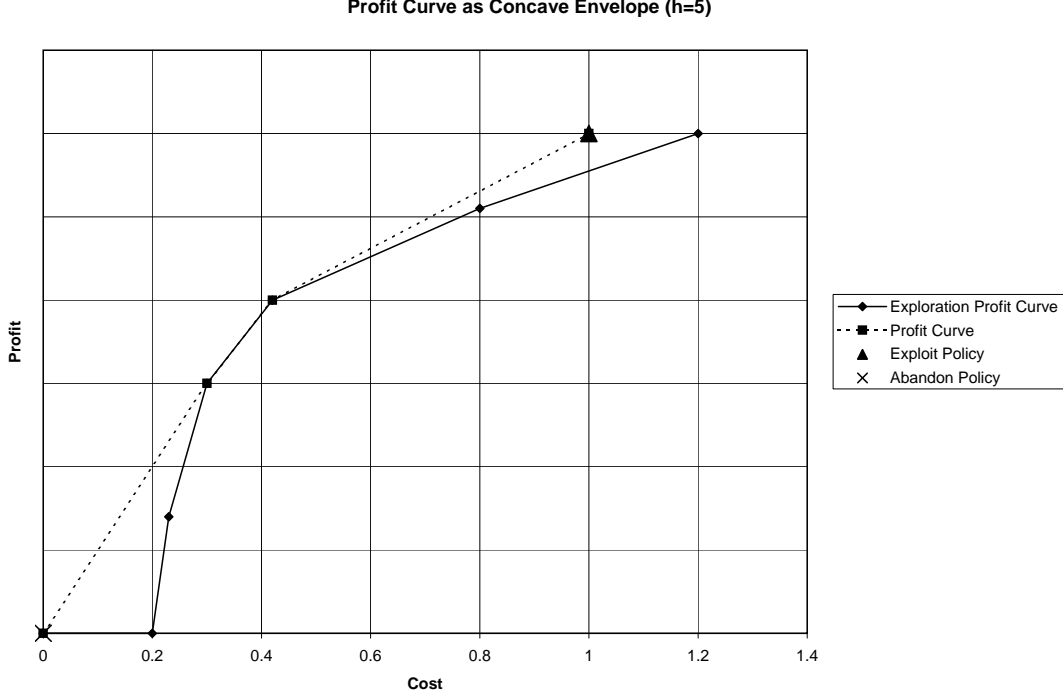


Figure 1: The relation between the profit and the exploration-profit curves

Under these conditions (i.e.  $e_u = 1$ ), we can modify the recursive equation we introduced earlier to calculate the profit curve for  $u$  ( $\mathcal{P}_u(\cdot)$ ) into a simpler equation to calculate the exploration profit curve

$$\mathcal{X}_u(\mathcal{C}_u) = \max_{E^u} \sum_{v \in D(u)} P_{uv} \mathcal{P}_v(E_v^u)$$

such that:

$$\begin{aligned} 1/h + \sum_{v \in D(u)} P_{uv} E_v^u &\leq \mathcal{C}_u \\ E^u &\geq 0 \end{aligned}$$

Recall that  $E_v^u$  is the budget allocated to  $v$  should we visit it immediately after  $u$ , and  $E^u$  is the vector of these budgets for each of the immediate descendants of  $u$ .

For each  $v \in D(u)$ , let  $B_v$  represent the number of “corner” solutions on  $\mathcal{P}_v(\cdot)$  and denote the cost of the  $i$ th such “corner” solution as  $s_i^v$  (where  $s_0^v = 0$  and  $s_{B_v}^v = 1$ ). We can then create the following modified recursive equation to calculate the exploration profit curve.

$$\max_{\epsilon^u} \sum_{v \in D(u)} P_{uv} \sum_{i=1}^{B_v} [\mathcal{P}_v(s_i^v) - \mathcal{P}_v(s_{i-1}^v)] \epsilon_{v,i}^u$$

s.t.

$$1/h + \sum_{v \in D(u)} P_{uv} \sum_{i=1}^{B_v} (s_i^v - s_{i-1}^v) \epsilon_{v,i}^u \leq \mathcal{C}_u$$

$$0 \leq \epsilon_{v,i}^u \leq 1$$

$$\forall v \in D(u), i \in \{1, \dots, B_v\}$$

This is a linear program where each decision variable,  $\epsilon_{v,i}^u$ , selects some fraction of the  $i$ th segment of the profit curve of  $v$ .  $\epsilon^u$  represents the collection of all such variables. As the profit curve is concave (by lemma 4.1),  $\frac{\mathcal{P}_v(s_i^v) - \mathcal{P}_v(s_{i-1}^v)}{s_i^v - s_{i-1}^v} \leq \frac{\mathcal{P}_v(s_j^v) - \mathcal{P}_v(s_{j-1}^v)}{s_j^v - s_{j-1}^v} \forall i \geq j$ . Thus, there exists an optimal solution which only assigns  $\epsilon_{v,i}^u > 0$  if  $\epsilon_{v,i-1}^u = 1$  for any  $i > 1$ .

Further, through inspection we can see that the optimal solution for any  $\mathcal{C}_u$  is to select the segments in order of decreasing slope until all budget is exhausted. By ordering segments thus, we can easily construct  $\mathcal{X}_u(\cdot)$ . The algorithm below orders the segments of the elements of  $D(u)$  and builds  $\mathcal{X}_u(\cdot)$ , storing the costs ( $c_i$ ) and budgets allocated to all descendants ( $E_v^u = \sum_i \epsilon_{v,i}^u$ ) for each ‘‘corner’’ solution of  $\mathcal{X}_u(\cdot)$ .

**Algorithm:** COMPUTEEXPLORATIONPROFITCURVE

1) For all  $v \in D(u), i \in \{1, \dots, B_v\}$

/\*Compute profit, cost, and slope for each line segment of  $\mathcal{P}_v(\cdot)$ \*/

Set  $\pi_i^v = \mathcal{P}_v(s_i^v) - \mathcal{P}_v(s_{i-1}^v)$

Set  $c_i^v = s_i^v - s_{i-1}^v$

Set  $M_i^v = \pi_i^v / c_i^v$

2) Sort the  $\sum_{v \in D(u)} B_v$  elements of the form  $M_i^v$  from largest to smallest

Let  $d(j)$  index the node for the  $j$ th largest element in the list

Let  $t(j)$  indicate which segment of  $V_j$  this is

/\* $M_{t(j)}^{d(j)}$  = slope of the  $j$ th line segment in the ordered list.\*/

3) Set  $\hat{S}_0 = \frac{1}{h}$  /\*fixed cost\*/,  $\hat{X}_0 = 0$  /\*initial profit\*/,  $\hat{E}_{v,0} = 0 \forall v \in D(u)$

/\*initial budgets for descendants\*/

4) For  $i = 1$  to  $\sum_{v \in D(u)} B_v$

/\*Add next segment to the curve\*/

/\*Compute current total profit ( $\hat{X}_i$ ) and cost ( $\hat{S}_i$ )\*/

Let  $\hat{X}_i = \hat{X}_{i-1} + P_{u,d(i)} \pi_{t(i)}^{d(i)}$

```

Let  $\hat{S}_i = \hat{S}_{i-1} + P_{u,d(i)} c_{t(i)}^{d(i)}$ 
/*Compute budgets allocated to descendants ( $\hat{\mathcal{E}}_{v,i}$ ) at the end*/
/*of each segment (needed to represent the policy)*/
Let  $\hat{\mathcal{E}}_{d(i),i} = \hat{\mathcal{E}}_{d(i),i-1} + c_{t(i)}^{d(i)}$ 
Let  $\hat{\mathcal{E}}_{v,i} = \hat{\mathcal{E}}_{v,i-1} \quad \forall v \neq d(i)$ 
5) Set  $n = 1, k_1 = 1$ 
6) For  $i = 2$  to  $\sum_{v \in D(u)} B_v$ 
/*Find changes in slope on the exploration profit curve*/
If  $\frac{\hat{X}_i - \hat{X}_{i-1}}{\hat{S}_i - \hat{S}_{i-1}} \neq \frac{\hat{X}_{i-1} - \hat{X}_{i-2}}{\hat{S}_{i-1} - \hat{S}_{i-2}}$ 
     $n++$ 
     $k_n = i$ 
7) For  $m = 1$  to  $n$ 
/*Merge together segments of the exploration profit*/
/*curve with the same slope*/
 $S_m = \hat{S}_{k_m}, \mathcal{X}_u(c_m) = \hat{X}_{k_m}, \mathcal{E}_{v,m}^u = \hat{\mathcal{E}}_{v,k_m}$ 
8)  $B_u^{\mathcal{X}} = n$ 

```

Algorithm COMPUTEEXPLORATIONPROFITCURVE represents the exploration profit curve for state  $u$  by returning the number of line segments of the curve (not including the zero slope segment from cost of 0 to  $S_0 = 1/h$ ),  $B_u^{\mathcal{X}}$ , as well as the cost ( $S_i$ ), profit ( $\mathcal{X}_u(S_i)$ ), and vector of budgets to allocate to all immediate descendants ( $\mathcal{E}_{\cdot,i}^u$ ) corresponding to the endpoint of each segment.

Given that the profit curves of all descendants are concave, the sorting of line segments in step 2) equates to simply interleaving the segments of the different states and can be performed in  $O(d \sum_u \log \Sigma_u)$  time using a simple min-heap, where  $d$  is the maximum number of immediate descendants for a node. After sorting these segments, steps 3) and 4) then determine the cost and profit associated with adding each segment to the exploration profit curve. Finally, as there may and likely will be duplicates in the sorted list of slopes, steps 5) through 7) merge all segments of the curve with the same slope.

Given the exploration profit curve, it is much easier to calculate the profit curve for a state. From lemma 4.1, we know that  $\mathcal{P}_u(1) = \zeta(u)$ . Further, this must be the only ‘‘corner’’ solution corresponding

to exploiting at  $u$  ( $p_u = 1$ ). All other "corner" solutions must thus correspond to exploring at  $u$  ( $e_u = 1$ ) and thus must correspond to points on  $\mathcal{X}_u()$ . As we know  $\mathcal{P}_u()$  is concave, we can simply take the concave envelope of the points  $(0, \mathcal{P}_u(0) = 0)$ ,  $(S_i, \mathcal{X}_u(S_i)) \forall i \in \{1, \dots, B_u^{\mathcal{X}}\}$ , and  $(1, \mathcal{P}_u(1) = \zeta(u))$ . The algorithm below does this.

**Algorithm:** COMPUTEPROFITCURVE

```

1) Find  $j^* = \arg \max_{j \in \{1, \dots, B_u^{\mathcal{X}}\}} R_j$  where  $R_j = \frac{\mathcal{X}_u(S_j)}{S_j}$ 
2a) If  $R_{j^*} \leq \zeta(u)$  /*The ratio index policy exploits immediately*/
    Set  $r(u) = \zeta(u), s_1^u = 1, \mathcal{P}_u(s_1^u) = \zeta(u), E_v^u(1) = 0 \ \forall v \in D(u), B_u = 1$ 
2b) Else /*The ratio index policy explores*/
    Set  $r(u) = R_{j^*}, s_1^u = S_{j^*}, \mathcal{P}_u(s_1^u) = \mathcal{X}_u(S_{j^*}), E^u(1) = \mathcal{E}_{:,j^*}^u$ 
    Set  $i = 1, j = j^* + 1$ .
    While  $\frac{\mathcal{X}_u(S_j) - \mathcal{X}_u(S_{j-1})}{S_j - S_{j-1}} > \frac{\zeta(u) - \mathcal{X}_u(S_{j-1})}{1 - S_{j-1}}$ 
        /*Greater marginal return to explore than exploit*/
         $s_{i+1}^u = S_j$ 
         $\mathcal{P}_u(s_{i+1}^u) = \mathcal{X}_u(S_j)$ 
         $E^u(i+1) = \mathcal{E}_{:,j}^u$ 
         $i = i + 1, j = j + 1$ 
    Set  $s_{i+1}^u = 1, \mathcal{P}(s_{i+1}^u) = \zeta(u), B_u = i + 1$ .

```

Algorithm COMPUTEPROFITCURVE runs in  $O(d\Sigma_u)$  time. Step 1) computes the value  $R_j$  for each segment of  $\mathcal{X}_u()$ , which represents to slope of the line segment from the origin to the point  $(c_i, \mathcal{X}_u(c_i))$ . In the event that  $R_{j^*} \leq \zeta(u)$ , the ratio index policy is to exploit immediately and we are done determining the profit curve for  $u$  (step 2a). Otherwise, once we have found the ratio index policy, we continue to look at higher budget exploration policies to determine subsequent segments of the profit curve (step 2b). The quantity  $E^u(i)$  represents the budgets allocated to each of the immediate descendants at the end of the  $i$ th segment of the profit curve. These values are only required to represent the actual policy, not calculate the ratio index or profit curve of any state. The quantity  $\frac{\zeta(u) - \mathcal{X}_u(c_{j-1})}{1 - c_{j-1}}$  is the marginal ratio of transitioning from  $s_i^u$  to exploitation at  $u$ . Once the slopes of the segments of  $\mathcal{X}_u()$  are no larger than this, it is optimal to transition to exploitation at  $u$ .

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 4.8 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, the total time to compute the profit curves for all states in the layered DAG is  $O(d\Sigma^2 \log \Sigma)$ , and this algorithm is strongly polynomial (in  $\Sigma$ ) 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  $\Sigma$ , so the algorithm is still strongly polynomial in  $\Sigma$ . 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  $h$ .

## **Acknowledgements**

The authors would like to thank Rajat Bhattacharjee and Sanjeev Khanna for helpful discussions.

## References

- [1] R. Bellman. A problem in the Sequential Design of Experiments. *Sankhya*, 16:221-229, 1956.
- [2] 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.
- [3] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *J Royal Statistical Societe Series B*, 14:148-167, 1979.
- [4] J. C. Gittins. *Multiarmed Bandits Allocation Indices*, Wiley, New York, 1989.
- [5] A. Goel, S. Khanna, and B. Null. The Ratio Index for Budgeted Learning, with Applications. submitted, 2008. <http://www.stanford.edu/~null/ratio.htm>
- [6] S. Guha and K. Munagala. Approximation Algorithms for Budgeted Learning Problems. *STOC*, 2007.
- [7] D. Lizotte, O. Madani, and R. Greiner. Budgeted learning of Naive Bayes classifiers. *UAI-2003*, 2003.
- [8] D. Luenberger. *Linear and Nonlinear Programming*, Reading, MA, 1984.
- [9] O. Madani, D. Lizotte, and R. Greiner. Active Model Selection. *ACM International Conference Proceeding Series; Proceedings of the 20th conference on Uncertainty in artificial intelligence*. 70: 357 - 365, 2004.
- [10] Y. Ye. A New Complexity Result on Solving the Markov Decision Problem. *Mathematics of Operations Research*, 30(3):733-749, 2005.

## A Proof of final three parts of Lemma 4.5

Below are the proofs for the final three properties of lemma 4.5.

**Proof:**

1. If  $e_v^* > 0$ ,  $\mathcal{P}'_{+u}(C) \geq \mathcal{P}'_{-v}(1)$ : Given  $v$  such that  $w_v^* > 0$  and  $e_v^* > 0$  assume  $\mathcal{P}'_{-v}(1) > \mathcal{P}'_{+u}(C)$ . Denote by  $E_v^*$  the amount of budget devoted to  $v$  and its descendants. Observe that it must be the case that  $E_v^* < 1$ . Otherwise there exists some other optimal solution  $\langle e^{*(1)}, p^{*(1)} \rangle$  to  $LP_u(C)$  with  $p_v^{*(1)} = 1, e_v^{*(1)} = 0$ . Thus, we know that total profit garnered from  $v$  and its descendants is  $\mathcal{P}_v(E_v^*)$ . Let  $\langle \tilde{e}, \tilde{p} \rangle_v$  be the policy that induces  $\mathcal{P}_v(E_v^* + \epsilon)$  on  $T_i^v$ . Let  $\langle \tilde{e}, \tilde{p} \rangle$  be a policy that follows  $\langle \tilde{e}, \tilde{p} \rangle_v$  in the subtree of  $v$  and follows  $\langle e^*, p^* \rangle$  otherwise. The policy  $\langle \tilde{e}, \tilde{p} \rangle$  is a feasible solution to the problem of finding the point on the profit curve of  $u$  with cost  $C + w_v^* \epsilon$ . This policy has profit  $\mathcal{P}_u(C) + w_v^* \epsilon \mathcal{P}'_{+v}(E_v^*)$ . Thus,

$$\mathcal{P}_u(C + w_v^* \epsilon) \geq \mathcal{P}_u(C) + w_v^* \epsilon \mathcal{P}'_{+v}(E_v^*)$$

Further, as  $E_v^* < 1$ ,  $\mathcal{P}'_{+v}(E_v^*) \geq \mathcal{P}'_{-v}(1)$ . Thus,

$$\mathcal{P}_u(C + w_v^* \epsilon) \geq \mathcal{P}_u(C) + w_v^* \epsilon \mathcal{P}'_{-v}(1)$$

$$\Rightarrow \mathcal{P}_u(C + w_v^* \epsilon) - \mathcal{P}_u(C) > w_v^* \epsilon \mathcal{P}'_u(C)$$

By concavity, this cannot be true, so it must be true that  $\mathcal{P}'_{+u}(C) \geq \mathcal{P}'_{-v}(1)$ .

3. If  $e_v^* > 0, r(v, h) \geq \mathcal{P}'_{-u}(C)$ : Given  $v$  such that  $w_v^* > 0$  and  $e_v^* > 0$  assume  $r(v, h) < \mathcal{P}'_{-u}(C)$ . Denote by  $E_v^*$  the amount of budget devoted to  $v$  and its descendants. Thus, we know that total profit garnered from  $v$  and its descendants is  $\mathcal{P}_v(E_v^*)$ . Let  $\langle \tilde{e}, \tilde{p} \rangle_v$  be the policy that induces  $\mathcal{P}_v(E_v^* - \epsilon)$  on  $T_i^v$ . Let  $\langle \tilde{e}, \tilde{p} \rangle$  be a policy that follows  $\langle \tilde{e}, \tilde{p} \rangle_v$  in the subtree of  $v$  and follows  $\langle e^*, p^* \rangle$  otherwise. The policy  $\langle \tilde{e}, \tilde{p} \rangle$  is a feasible solution to the problem of finding the point on the profit curve of  $u$  with cost  $C - w_v^* \epsilon$ . This policy has profit  $\mathcal{P}_u(C) - w_v^* \epsilon \mathcal{P}'_{-v}(E_v^*)$ . Further, as  $E_v^* > 0$  (since  $e_v^* > 0$ ), we know that  $\mathcal{P}'_{-v}(E_v^*) \leq r(v, h)$ . Thus,

$$\mathcal{P}_u(C - w_v^* \epsilon) \geq \mathcal{P}_u(C) - w_v^* \epsilon r(v, h)$$

$$> \mathcal{P}_u(C) - w_v^* \epsilon \mathcal{P}'_{-u}(C)$$

Thus,  $\mathcal{P}_u(C) - \mathcal{P}_u(C - w_v^* \epsilon) < w_v^* \epsilon \mathcal{P}'_u(C)$ . By concavity, this cannot be true, so it must be true that  $\mathcal{P}'_{-v}(1) \geq \mathcal{P}'_{-u}(C)$ .

4. If  $1 - e_v^* - p_v^* > 0, \mathcal{P}'_{+u}(C) \geq r(v, h)$ : Given  $v$  such that  $w_v^* > 0$  and  $1 - e_v^* - p_v^* > 0$  assume  $r(v, h) > \mathcal{P}'_{+u}(C)$ . Denote by  $E_v^*$  the amount of budget devoted to  $v$  and its descendants. Thus, we know that total profit garnered from  $v$  and its descendants is  $\mathcal{P}_v(E_v^*)$ . Let  $\langle \tilde{e}, \tilde{p} \rangle_v$  be the policy that induces  $\mathcal{P}_v(E_v^* + \epsilon)$  on  $T_i^v$ . Let  $\langle \tilde{e}, \tilde{p} \rangle$  be a policy that follows  $\langle \tilde{e}, \tilde{p} \rangle_v$  in the subtree of  $v$  and follows  $\langle e^*, p^* \rangle$  otherwise. The policy  $\langle \tilde{e}, \tilde{p} \rangle$  is a feasible solution to the problem of finding the point on the profit curve of  $u$  with cost  $C + w_v^* \epsilon$ . This policy has profit  $\mathcal{P}_u(C) + w_v^* \epsilon \mathcal{P}'_{+v}(E_v^*)$ . Thus,

$$\mathcal{P}_u(C + w_v^* \epsilon) \geq \mathcal{P}_u(C) + w_v^* \epsilon \mathcal{P}'_{+v}(E_v^*)$$

Further, as  $1 - e_v^* - p_v^* > 0$ ,  $\mathcal{P}'_{+v}(E_v^*) = r(v, h)$ . (Since neither  $e_v^* = 1$  nor  $p_v^* = 1$ , at least one must be zero and the other is either zero or the transitional state. Thus, at one end of this segment of  $\mathcal{P}_v(\cdot)$  must be the abandonment policy, so we are on the first segment of the profit curve of  $v$ .) Thus,

$$\mathcal{P}_u(C + w_v^* \epsilon) \geq \mathcal{P}_u(C) + w_v^* \epsilon r(v, h)$$

$$\Rightarrow \mathcal{P}_u(C + w_v^* \epsilon) - \mathcal{P}_u(C) > w_v^* \epsilon \mathcal{P}'_u(C)$$

By concavity, this cannot be true, so it must be true that  $\mathcal{P}'_{+u}(C) \geq r(v, h)$ .

■