

Approximation Algorithms for Budgeted Learning Problems

Sudipto Guha* Kamesh Munagala†

March 5, 2007

Abstract

We present the first approximation algorithms for a large class of budgeted learning problems. One classic example of the above is the budgeted multi-armed bandit problem. In this problem each arm of the bandit has an unknown reward distribution on which a prior is specified as input. The knowledge about the underlying distribution can be refined in the exploration phase by playing the arm and observing the rewards. However, there is a budget on the total number of plays allowed during exploration. After this exploration phase, the arm with the highest (posterior) expected reward is chosen for exploitation. The goal is to design the adaptive exploration phase subject to a budget constraint on the number of plays, in order to maximize the expected reward of the arm chosen for exploitation. While this problem is reasonably well understood in the infinite horizon setting or regret bounds, the budgeted version of the problem is NP-HARD. For this problem, and several generalizations, we provide approximate policies that achieve a reward within constant factor of the reward optimal policy. Our algorithms use a novel linear program rounding technique based on stochastic packing.

*Department of Computer and Information Sciences, University of Pennsylvania. Email: sudipto@cis.upenn.edu. Research supported in part by an Alfred P. Sloan Research Fellowship and by an NSF Award CCF-0430376.

†Department of Computer Science, Duke University. Email: kamesh@cs.duke.edu. Research supported in part by NSF CNS-0540347.

1 Introduction

In this paper, we study the problem of optimal design of experiments under a budget constraint. Consider first the following problem: we are given n coins and told that the probability p_i of obtaining head on tossing coin i is distributed according to a known prior distribution \mathcal{R}_i over $[0, 1]$. We can toss any coin any number of times – but we are limited to 100 tosses. If we wish to pick a coin i^* such that we maximize the probability of heads p_{i^*} , how should we proceed? What should be our strategy if the coin has 3 or more outcomes? Similar problems are ubiquitous in scenarios where data acquisition or running an experiment has a cost. For instance, in a clinical trial setting, we have several competing treatments and little data on the effectiveness of the individual treatments. If we could collect as much data as we would like, this is straightforward. The data however does not come for free, since it costs a lot of money to run a treatment on a patient; and for an uncommon enough disease, there are not enough patients. However we may have some degree of belief or partial information expressed through prior distributions about the treatments. If we seek to find the most effective treatment, the data that is collected should depend on the outcome of the treatments on previous patients. Given a budget for running the clinical trials to gather the training data, the question is how best to dynamically allocate the budget among gathering the data for competing treatments so that we find the most effective treatment. It is obvious that this problem extends to any classification scenario where we are interested in finding the most successful classifier from a few trials where we can verify the classifier by human experts.

The above problems are examples of the budgeted multi-armed bandit problem which has recently gained currency as active model selection or optimal design of experiments problems [25, 8, 28]. Abstractly, we are given a bandit with n arms (or treatments). The reward when arm i is played (*i.e.*, the treatment i is tried on some patient) follows some fixed distribution R_i . This distribution is over K non-negative values $\{a_1, a_2, \dots, a_K\}$. There is a cost c_i for playing arm i . In the above coins example, K was equal to 2 and $c_i = 1$ for all i . In the example corresponding to the clinical trials, c_i denotes the cost of testing treatment i on a patient.

We assume that $0 = a_1 \leq a_2 \dots, a_K = 1$ by suitable scaling. The arms are assumed to be independent, so that the distributions for any i are independent of the corresponding distributions for an $i' \neq i$. The distributions R_i are *not known* to the player, but the player has some idea of the *prior distribution* \mathcal{R}_i from which the corresponding R_i is drawn. Note that R_i are drawn and fixed before the player gets to make a move. Whenever the arm i is played, the player gets an outcome that depends on the true fixed distribution R_i . In case of the coin toss, the R_i will be the probability vector $(1 - p_i, p_i)$ corresponding to coin i and we were told of the distribution of p_i . If the number of times an arm can be played is unlimited, then the exact distribution R_i can be determined by observing the fraction of times a_1, a_2, \dots, a_K are obtained. In such a setting the player will play a few arms and then settle on the best arm eventually. In this setting the reasonable optimization measure is the “infinite horizon discounted reward” and in this setting optimum strategies, which are also simple to characterize, are known [13, 32, 3]. These strategies unfortunately are *not* optimal if there is a limit on the number of plays on an arm. This budgeted aspect is the key issue in the design of experiments and active model selection.

An alternate view of the above is a process where playing the arm i the player resolves a (current) prior distribution to a posterior distribution, which is the current prior for the state the player arrives at. For example, for the case of a coin, on observing a head on tossing coin i , we can expect that the reward, the probability p_i , to be larger. We make the standard assumption that the priors can be efficiently updated. Examples include the well-known Dirichlet priors [18] over multinomial distributions, and arbitrary discrete priors over arbitrary discrete distributions over

the K reward values. It is also typical to assume K is a small constant, and the total number of plays is polynomial (since the policy needs to execute in poly-time). These assumptions make the problem no less intractable.

Any policy for the budgeted multi-armed bandit problem described above consists of two phases:

1. In the “exploration” (or clinical trial phase), a cost budget C is given. The arms are played (or the treatments are tested on patients) in some adaptively determined order (possibly dependent on the outcome of the previous plays), as long as the total cost of playing does not exceed C . When the budget C is exhausted, each arm has been played a certain number of times, and the outcome \bar{s}_i for the plays for any arm i (or the results of the trials for the treatment i) yield the posterior distribution $\mathcal{R}_i(\bar{s}_i)$ on the possible reward distribution R_i for this arm conditioned on the outcome \bar{s}_i of the exploration. Denote by \bar{s} the entire set of outcomes of all the plays of all the arms; given that arms are independent, if \bar{s}_i is the projection of \bar{s} to the outcomes of the i^{th} arm then $\mathcal{R}_i(\bar{s}_i) = \mathcal{R}_i(\bar{s})$.
2. After the exploration phase is over, the policy moves into the exploitation phase where it has to choose one arm irrevocably to play forever (this corresponds to choosing one treatment). The “exploitation gain” of the policy is the expected reward of this arm (or the expected efficacy of the chosen treatment). Clearly, at the end of the exploration phase, the policy will choose that arm i such that $\mu_i(\bar{s}) = \mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})}[\mathbf{E}[R_i]]$ is maximum, where the outer expectation is over choosing a reward distribution R_i according to the posterior distribution $\mathcal{R}_i(\bar{s})$, and \bar{s} is the outcome of the entire exploration. Since the distributions $\mathcal{R}_i(\bar{s})$ depend on the adaptive strategy used in the exploration phase, the exploitation gain is clearly a function of the strategy used in the exploration phase.

The **goal** is to devise an adaptive budget-constrained exploration strategy whose expected exploitation gain (over all possible \bar{s}) is maximized. Computing the optimal solution of the budgeted multi-armed bandit problem is NP-HARD even when at most one play is allowed per arm [25, 9], and even with unit cost ($c_i = 1$) for any play [14].

Our Contribution: In this paper, we present a 4-approximation to the budgeted multi-armed bandit problem. We present a unified solution framework for general budgeted learning problems where the exploitation gain could involve any separable non-negative concave utility function over the rewards of the arms subject to a packing constraint. The concave utility assumption models decreasing marginal utilities in resource allocation, and captures natural extensions such as choosing the top- m best arms for exploitation (corresponding to choosing the m best treatments), etc. Consider for instance, a pharmaceutical company that wants to decide how to allocate a different fixed resource for production (which is separate from the C for trials) among the competing treatments. The benefit (to the company) of a treatment i is a non-negative function $g(x_i, r_i)$ of the allotted resource x_i , and effectiveness r_i . This function is concave and non-decreasing in x_i (decreasing marginal benefit), and non-decreasing in r_i . The distribution of the r_i can be resolved during exploration (clinical trial phase) before deciding how to allocate x_i among competing treatments. Our framework handles these type of extensions which arise frequently in optimizing system performance (e.g. resource allocation for fairness [24]), as long as the arms are independent.

Our framework also extends naturally to solve Lagrangean variants where the optimization objective or net profit, is the *difference* between the exploitation gain and the total cost spent in exploration. There is no explicit budget constraint. Note that the net profit could be negative if the exploration cost outweighs the reward of the best arm – our framework handles such difficulties.

The policies produced by our algorithm are natural – they are semi-adaptive policies, where the play of an arm depends on the outcomes for that arm, but the arms are explored in a fixed order and never revisited.

Techniques: Our policy design is via a linear programming formulation over the state space of individual arms – since the state space of any single arm is polynomially bounded (refer Section 2 for details) and the arms are independent, we achieve polynomial sized formulation. We then bring to bear techniques from stochastic packing literature, particularly the work on adaptivity gaps [11, 12, 10] to develop a novel LP rounding scheme. Our overall technique can be thought of as “LP rounding via stochastic packing” – finding this connection between budgeted learning and stochastic packing by designing simple LP rounding policies for a very general class of budgeted learning problems represents the key contribution of this work.

The work of Dean, Goemans, and Vondrák [11] shows constant factor adaptivity gaps for knapsack and related packing problems with stochastic item sizes when items need to be *irrevocably* committed to the knapsack before their true size is revealed. In contrast, the budgeted multi-armed bandit problem does not have irrevocable commitment – after an arm has been explored many times, it can be discarded if no benefit is observed; an arm can be revisited multiple times during exploration, etc. In fact, the strategy needs to be adaptive (refer Appendix A). Despite these significant differences between the problems, we show that the idea of irrevocable commitment is a powerful rounding technique, yielding simple to analyze semi-adaptive policies.

Related Work: The budgeted multi-armed bandit problem is a special case of *active learning* [8, 2, 19, 27], and also a special case of the Design of Experiments (DOE) problem in Statistics literature (refer [28] for a survey). This problem also falls in the framework of *action-evaluation* in meta-reasoning [9]. Computing the optimal solution of the budgeted multi-armed bandit problem is NP-HARD even when at most one play is allowed per arm [25, 9], and with unit costs [14].

Several heuristics have been proposed for the budgeted multi-armed bandit problem by Schneider and Moore [29]. Extensive empirical comparison of these heuristics in the context of the coins problem ($K = 2$ and 0/1 rewards) is presented by Madani *et al.* [25], and it appears that an adaptive policy which makes the next play decision based on both (1) the remaining budget, and (2) the outcomes of the plays so far, outperform heuristics which base their decisions on other information.

Prior work [23, 14, 16], considered the problem of *model-driven optimization*, where full information was revealed in a single “probe”. This is a special case of the budgeted multi-armed bandit where the priors are point masses, so that any R is some value a_j with probability 1. In [14, 16], we showed that there existed non-adaptive algorithms (which ignores the outcomes \bar{s} for the plays, but took the outcome in consideration in the subsequent optimization) that perform up to a constant factor of the best adaptive strategy. This allowed us greedy algorithms based on submodularity. Similar greedy algorithms were proposed in [23] for the problem of minimizing residual entropy. However, these sub-modularity based techniques do not extend to general priors, since the adaptivity gap between ignoring and considering \bar{s} can be worse by a factor $\Omega(n)$ – we present such an example for $c_i = 1$ and *i.i.d* priors in Appendix A.

The Lagrangean variant with point priors and the net profit being the difference between the value of the highest posterior expected reward arm and the exploration cost is considered in [17], where a 1.25 approximation is presented. The techniques are again specific to point priors and do not extend to either general priors.

There is a rich literature on policies for stochastic packing [21, 15, 11, 12] as well as stochastic scheduling [26, 31]. LP formulations over the state space of outcomes exist for multi-stage stochastic optimization with recourse [22, 30, 7]. However, the significant difference is that in order to achieve

poly-sized formulation, we do not sample the joint space of outcomes – since our budget constraint runs across stages, our scenarios would be trajectories [20] which are hard to sample. We instead use the independence of the arms to decompose the state space. In that sense, our LP relaxation is akin to the LP relaxations for multi-class queueing systems [4, 5, 6]; however, unlike our relaxation, there are no provable guarantees on the quality of the multi-class queueing relaxations.

2 The State Space

Definition 1. For arm i , the true reward distribution R_i is distributed according to the prior \mathcal{R}_i . Let Θ_i be the distribution of $\mathbf{E}[R_i]$ as R_i is chosen from the prior \mathcal{R}_i . The distributions for different arms are assumed to be independent. The R_i are over K values $0 \leq a_1 \cdots \leq a_K \leq 1$.

Given an outcome \bar{s} for the exploration, let $\mathcal{R}_i(\bar{s})$ denote the posterior distribution of the reward R_i for arm i conditioned on the projection to arm i of the outcome \bar{s} . Let $\Theta_i(\bar{s})$ denote the corresponding distribution of $\mathbf{E}[R_i]$. Given the outcome \bar{s} , the exploitation phase chooses that arm i whose $\mathbf{E}[\Theta_i(\bar{s})]$ is maximum. The exploitation gain is the expectation of this quantity over all outcomes \bar{s} .

Key Assumptions: Recall that number of possible outcomes on playing an arm once is K . Let h be the maximum number of times any particular arm can be played. For each arm, the specification of the policy of that arm depends at least on the observed rewards *for that arm*. The space of possible observations (or states) for a **single arm** (see more below) has size at most $\Sigma = \binom{h+K}{K}$. We assume Σ is polynomially bounded, and the running time of the algorithms we present will depend polynomially on Σ . This follows for instance if h is polynomially bounded, and K is a constant. The value of Σ is indeed polynomially bounded in most previous work on budgeted learning [25, 29, 14, 16]. For instance, the coins problem [25], corresponds to $K = 2$ and h being polynomially bounded. The model-driven optimization framework [14, 16] corresponds to $h = 1$ and K being polynomially bounded. For these problems, note that the joint state space of all n arms has size $\Omega(\Sigma^n)$ which is exponentially large – in fact, these problems are NP-HARD [9, 25, 14] even for h being a constant.

The following theorem implies that even if h and K are arbitrary, Σ can be made poly-logarithmic by only losing a small additive term in approximation ratio. It is proved in Appendix B

Theorem 2.1. Let G_C denote the exploitation gain of the optimal policy with cost budget C .

1. For $\epsilon > 0$, let $G_C(\epsilon)$ denote the gain of the optimal policy which the same cost budget which plays any arm at most $N_\epsilon = \frac{12 \log n}{\epsilon^2}$ times. For sufficiently small constant ϵ , we have $G_C(\epsilon) \geq G_C - 16\epsilon$.
2. For any $\epsilon > 0$, let $\mathcal{K} = 1/\epsilon$. Let $G_C(\epsilon)$ denote the gain of the optimal policy when the space of outcomes for playing any arm is discretized so that if $a_j \in [l\epsilon, (l+1)\epsilon)$ for $l \in \{0, 1, \dots, \mathcal{K}-1\}$, then $a_j \leftarrow (l+1)\epsilon$, so that each distribution R_i is now discrete over \mathcal{K} values. Then, $G_C(\epsilon) \geq G_C - \epsilon$.

The above theorem implies that with loss of additive $O(\epsilon)$ in the exploitation gain of the optimal solution, we can set $h = O(\frac{\log n}{\epsilon^2})$ and $K = 1/\epsilon$, so that $\Sigma = O((\frac{\log n}{\epsilon^2})^{1/\epsilon})$ is poly-logarithmic.

We will also assume that conditioned on a new observed reward, the prior $\mathcal{R}_i(\bar{s})$ can be updated in polynomial time. This is true if the priors \mathcal{R} are the standard Dirichlet priors [18] on multinomial distributions; or if the priors \mathcal{R} and distributions R are discrete distributions of polynomial size.

To characterize the state space further, focus on a particular arm i and consider the behavior of the optimal strategy on this arm. When the optimal solution plays this arm, then based on

the outcome, the optimum may try other arms and then revisit this arm to try again. The *state* of arm i is characterized by the multi-set of reward outcomes observed for this arm. This set of observed outcomes for this arm uniquely defines the posterior distribution $\mathcal{R}_i(\bar{s})$ and hence the posterior mean. Notice that the order of the outcomes are not relevant since the outcomes are being drawn *i.i.d.* from an underlying distribution R_i . Therefore, the states for arm i describe a labeled directed acyclic graph (DAG) \mathbf{T}_i – conditioned on being at a node (or state) $u \in \mathbf{T}_i$ the *multi-set* of outcomes along all paths from the initial state (or root) of arm i to the state u are the same.

Definition 2. *The resolution DAG \mathbf{T}_i for arm i is a rooted directed acyclic graph with out-degree K for each node. The nodes correspond to states of the arm. The K out-edges from every internal node $u \in \mathbf{T}_i$ are labeled a_1, a_2, \dots, a_K respectively, and corresponding to the outcomes of the arm being played when its state is u . Let $l(e)$ denote the label on edge e .*

Let ρ_i denote the root node of DAG \mathbf{T}_i . Each node $u \in \mathbf{T}_i$ encodes a possible state of arm i . Suppose this node is at depth d and let $\{b_1, b_2, \dots, b_d\} \in \{a_1, a_2, \dots, a_K\}^d$ denote the edge-labels encountered along some path from ρ_i to u . Then the node u corresponds to arm i being played d times and the rewards observed for the d plays being b_1, b_2, \dots, b_d in arbitrary order.

The size of DAG \mathbf{T}_i is at most $\Sigma = \binom{h+K}{K}$ which as assumed above is polynomially bounded. Conditioned on reaching a node u , the outcomes of the arm i , $\bar{s}_i = b_1, b_2, \dots, b_d$ are determined. We will use $\mathcal{R}_i(u)$ to denote the corresponding $\mathcal{R}_i(\bar{s})$. Likewise we also obtain the posterior distribution $\Theta_i(u)$ on the possible $\mathbf{E}[R_i]$ values.

Definition 3. *For $u \in \mathbf{T}_i$, let $\zeta_i(u) = \mathbf{E}[\Theta_i(u)]$. We will overload terminology and term $\zeta_i(u)$ the “reward” of arm i given its state is u . The root node ρ_i has reward $\zeta_i(\rho_i) = \mathbf{E}[\Theta_i]$ since $\mathcal{R}_i(\rho_i) = \mathcal{R}_i$.*

Definition 4. *Let $D(u)$ denote the children of node u in \mathbf{T}_i . These correspond to the next states when arm i is played and its initial state was u . Suppose edge e from node u to node $v \in D(u)$ is labeled with $l(e) = a_j \in \{a_1, a_2, \dots, a_K\}$. Then, the edge e has “traversal probability” $\mathbf{p}_{uv} = \Pr[l(e)|\mathcal{R}_i(u)]$. This is the conditional probability that when arm i is played and initially had state $u \in \mathbf{T}_i$, the observed outcome is $l(e) = a_j$.*

The next claim re-states the definition of Bayesian inference in a form more useful to us:

Claim 2.2. $\zeta_i(u) = \sum_{v \in D(u)} l(u, v) \times \mathbf{p}_{uv}$. Furthermore, $\zeta_i(u) = \sum_{v \in D(u)} \zeta_i(v) \times \mathbf{p}_{uv}$

Illustrative Example: Consider the case where $K = 2$, and the rewards for the states are 1 and 0. In this case, any reward distribution R is characterized by the single parameter $q = \Pr[R = 1] = \mathbf{E}[R] = \Theta$. It is standard in Bayesian statistics [25] to assume that the prior in this case is a beta distribution. We assume now that the prior distribution \mathcal{R}_i of Θ_i (or q) follows a beta distribution $B(\alpha_1, \alpha_2)$ for positive integers α_1, α_2 . This corresponds to having observed $(\alpha_1 - 1)$ 0’s and $(\alpha_2 - 1)$ 1’s. The p.d.f. for this distribution is $c\Theta^{\alpha_1-1}(1-\Theta)^{\alpha_2-1}$, where c is the normalizing constant. We have $\mu_i = \mathbf{E}_{\mathcal{R}_i}[\Theta_i] = \frac{\alpha_1}{\alpha_1 + \alpha_2}$. Furthermore, if the arm is played once and reward 1 is observed (which happens with probability $\frac{\alpha_1}{\alpha_1 + \alpha_2}$), the posterior distribution of Θ becomes $B(\alpha_1 + 1, \alpha_2)$. Similarly, if the play results in 0, the posterior distribution becomes $B(\alpha_1, \alpha_2 + 1)$. Note that $B(1, 1)$ is the uniform distribution.

Therefore, for DAG \mathbf{T}_i , the root ρ has $\mathcal{R}_i = B(\alpha_1, \alpha_2)$, and $\zeta_i(\rho) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$. This node has two children: (i) The child u corresponding to playing the arm and observing reward 1 has $\mathcal{R}_i(u) = B(\alpha_1 + 1, \alpha_2)$, $\zeta_i(u) = \frac{\alpha_1 + 1}{\alpha_1 + \alpha_2 + 1}$, and $\mathbf{p}_{\rho u} = \frac{\alpha_1}{\alpha_1 + \alpha_2}$; and (ii) the child v corresponding to playing the arm and observing reward 0 has $\mathcal{R}_i(v) = B(\alpha_1, \alpha_2 + 1)$, $\zeta_i(v) = \frac{\alpha_2}{\alpha_1 + \alpha_2 + 1}$, and $\mathbf{p}_{\rho v} = \frac{\alpha_2}{\alpha_1 + \alpha_2}$.

3 Linear Program Formulation

Since our problem is a special case of finite-horizon Markov Decision Processes, standard dynamic programming techniques solve it. However, this requires computing an action for each possible state of the system. As mentioned above, the joint state space has size $\Omega(\Sigma^n)$. We show a polynomial size linear programming relaxation below with $O(nK\Sigma)$ variables and constraints. The catch of course is that the LP is only a relaxation and does not directly yield a feasible policy.

Consider any adaptive policy and its associated decision tree. For a trajectory (or decision path) t in this tree, and for some arm i and $u \in \mathbf{T}_i$, let $A_u(t)$ be a 0/1 random variable which is 1 if $u \in t$. Let $w_u = \mathbf{E}[A_u(t)]$ where the expectation is over all trajectories t . Therefore, w_u denotes the probability that during the execution of the policy, arm i enters state $u \in \mathbf{T}_i$. Similarly, let $B_u(t)$ be a 0/1 r.v. which is 1 if on trajectory t , state $u \in \mathbf{T}_i$ is visited and arm i is played in this state, and let r.v. $D_u(t)$ be 1 if arm i in state $u \in \mathbf{T}_i$ is chosen for exploitation. Then $z_u = \mathbf{E}[B_u(t)]$ denotes the probability that the state of arm i is u and the policy plays arm i in this state, and $x_u = \mathbf{E}[D_u(t)]$ denotes the probability that the policy chooses the arm i in state u for exploitation. Note that since $B_u(t)$ and $D_u(t)$ correspond to mutually exclusive events, we have $B_u(t) + D_u(t) \leq A_u(t)$. Consider the following LP which has three variables w_u, x_u , and z_u for each arm i and each $u \in \mathbf{T}_i$.

$$\begin{aligned} & \text{Maximize} && \sum_{i=1}^n \sum_{u \in \mathbf{T}_i} x_u \zeta_i(u) \\ & \sum_{i=1}^n c_i (\sum_{u \in \mathbf{T}_i} z_u) & \leq & C \\ & \sum_{i=1}^n \sum_{u \in \mathbf{T}_i} x_u & \leq & 1 \\ & \sum_{v: u \in D(v)} z_v \mathbf{p}_{vu} & = & w_u \quad \forall i, u \in \mathbf{T}_i \setminus \{\rho_i\} \\ & x_u + z_u & \leq & w_u \quad \forall u \in \mathbf{T}_i, \forall i \\ & x_u, z_u, w_u & \in & [0, 1] \quad \forall u \in \mathbf{T}_i, \forall i \end{aligned}$$

Let γ^* be the optimal LP value, and OPT be the exploitation gain of the optimal adaptive policy.

Claim 3.1. $OPT \leq \gamma^*$.

Proof. We show that the w_u, z_u, x_u are feasible for the constraints of the LP. The first two constraints follow by linearity of expectation over the space of trajectories, since on any trajectory, the exploration cost is at most C and at most one arm is chosen for exploitation respectively. Conditioned on reaching $v \in \mathbf{T}_i$ and playing arm i , the next state is $u \in \mathbf{T}_i$ with probability precisely \mathbf{p}_{vu} . This implies the set of constraints: $\sum_{v: u \in D(v)} z_v \mathbf{p}_{vu} = w_u$. The constraint $x_u + z_u \leq w_u$ follows by linearity of expectation on $B_u(t) + D_u(t) \leq A_u(t)$. Again, by linearity of expectation over the trajectory space, the exploitation gain of the adaptive optimal solution is precisely $\sum_{i=1}^n \sum_{u \in \mathbf{T}_i} x_u \zeta_i(u)$. This proves the claim. \square

Since the size of each DAG \mathbf{T}_i is at most $\Sigma = \binom{h+K}{K}$. The LP therefore has $O(n\Sigma)$ variables and $O(nK\Sigma)$ constraints. Since we assume Σ is polynomially bounded, the LP can be solved in polynomial time. Using Theorem 2.1, even if h and K are arbitrary making Σ super-polynomial, the state space can be reduced to $h = O(\frac{\log n}{\epsilon^2})$ and $K = \frac{1}{\epsilon}$ (so that Σ is poly-logarithmic) while losing an *additive* $O(\epsilon)$ in the exploitation gain of the optimal solution. We leave open the question of achieving bounds with multiplicative error guarantees for arbitrary h and K . This would require succinct representation of the state-space of a single arm, and hence new techniques.

4 The Exploration Policy

The optimal LP solution clearly does not directly correspond to a feasible policy since the variables do not faithfully capture the joint evolution of the states of different arms. Below, we present an interpretation of the LP solution, and show how it can be converted to a feasible approximately optimal policy.

Let $\langle w_u^*, x_u^*, z_u^* \rangle$ denote the optimal solution to the LP. Assume w.l.o.g. that $w_{\rho_i}^* = 1$ for all trees \mathbf{T}_i . Ignoring the first two constraints of the LP for the time being, the remaining constraints encode a separate policy for each arm as follows: Consider any arm i in isolation. The play starts at state ρ_i . The arm is played with probability $z_{\rho_i}^*$, so that state $u \in \mathbf{T}_i$ is reached with probability $z_{\rho_i}^* \mathbf{P}_{\rho_i u}$. At state ρ_i , with probability $x_{\rho_i}^*$, the play stops and arm i is chosen for exploitation. The events involving playing the arm and choosing for exploitation are disjoint. Similarly, conditioned on reaching state $u \in \mathbf{T}_i$, with probabilities z_u^*/w_u^* and x_u^*/w_u^* , arm i is played and chosen for exploitation respectively. This yields a policy \mathcal{A}_i for arm i which is described in Figure 1. The policy \mathcal{A}_i sets $\mathcal{E}_i = 1$ if on termination, arm i was chosen for exploitation. If $\mathcal{E}_i = 1$ at state $u \in \mathbf{T}_i$, then $X_i = \zeta_i(u)$ denotes the reward of the state where the policy terminated, in other words, the exploitation gain. Therefore, \mathcal{E}_i and X_i are random variables whose values depend on the execution trajectory of \mathcal{A}_i . For policy \mathcal{A}_i , it is easy to see by induction that if state $u \in \mathbf{T}_i$ is reached by the policy with probability w_u^* , then state $u \in \mathbf{T}_i$ is reached *and* arm i is played with probability z_u^* . Let r.v. C_i denote the cost of executing policy \mathcal{A}_i .

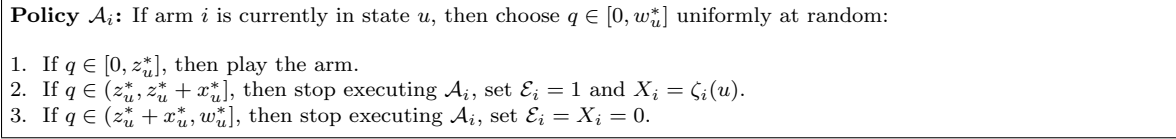


Figure 1: The Policy \mathcal{A}_i .

Let \mathcal{A} denote the exploration policy that is obtained by executing each \mathcal{A}_i independently in succession. Since policy \mathcal{A}_i is obtained by considering arm i in isolation, \mathcal{A} is **not a feasible policy** for the following reasons: (i) The cost $\sum_i C_i$ spent exploring all the arms need not be at most C in the worst case, and (ii) It could happen that for several arms i , \mathcal{E}_i is set to 1, which implies several arms could be chosen simultaneously for exploitation.

However, all is not lost. First note that the r.v. X_i, C_i, \mathcal{E}_i for different i are independent. Furthermore, it is easy to see using the first two constraints and objective of the LP formulation that \mathcal{A} is feasible in a certain expected sense:

1. $\mathbf{E}[C_i] = c_i \sum_{u \in \mathbf{T}_i} z_u^*$ so that $\sum_i \mathbf{E}[C_i] \leq C$.
2. $\mathbf{E}[\mathcal{E}_i] = \sum_{u \in \mathbf{T}_i} x_u^*$ so that $\sum_i \mathbf{E}[\mathcal{E}_i] \leq 1$.
3. $\mathbf{E}[X_i] = \sum_{u \in \mathbf{T}_i} x_u^* \zeta_i(u)$ so that $\sum_i \mathbf{E}[X_i] = \gamma^*$.

Based on the above, we show that policy \mathcal{A} can be converted to a feasible policy using ideas from the adaptivity gap proofs for stochastic packing problems [11, 12, 10]. We treat each policy \mathcal{A}_i as an item which takes up cost C_i , has size \mathcal{E}_i , and profit X_i . These items need to be placed in a knapsack – placing item i corresponds to exploring arm i according to policy \mathcal{A}_i . This placement is an irrevocable decision, and after the placement, the values of C_i, \mathcal{E}_i, X_i are revealed. We need $\sum_i C_i$ for items placed so far should be at most C . Furthermore, the placement (or exploration) stops the first time some \mathcal{E}_i is set to 1, and uses arm i is used for exploitation (obtaining gain or profit X_i). Since only one $\mathcal{E}_i = 1$ event is allowed before the play stops, this yields the "size constraint" $\sum_i \mathcal{E}_i \leq 1$. The knapsack therefore has both cost and size constraints, and the goal

is to sequentially and irrevocably place the items in the knapsack, stopping when the constraints would be violated. The goal is to choose the order to place the items in order to maximize the expected profit, or the exploitation gain. This is a two-constraint stochastic packing problem. The LP solution implies that the expected values of the random variables satisfy the packing constraints.

We show that the “start-deadline” framework in [10] can be adapted to show that there is a fixed order of exploring the arms according to the \mathcal{A}_i which yields gain at least $\gamma^*/4$. There is one subtle point – the profit (or gain) is also a random variable correlated with the size and cost. Furthermore, the “start deadline” model in [10] would also imply the final packing could violate the constraints by a small amount. We get around this difficulty by presenting an algorithm GREEDYORDER that explicitly obeys the constraints, but whose analysis will be coupled with the analysis of a simpler policy GREEDYVIOLATE which exceeds the budget. The central idea would be that although the benefit of the current arm has not been “verified”, the alternatives have been ruled out.

4.1 The Rounding Algorithm

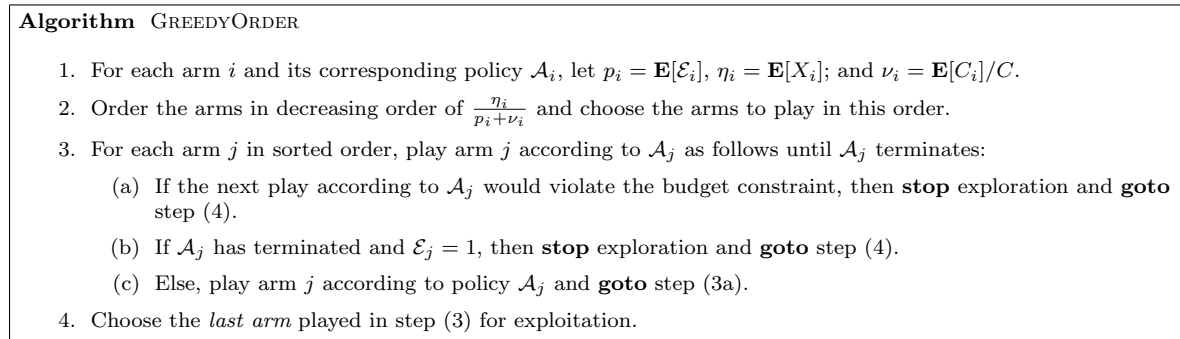


Figure 2: The GREEDYORDER policy.

The GREEDYORDER policy is shown in Figure 2. Note that step (4) makes the above policy online – no arm is ever revisited. For the purpose of analysis, we first present an infeasible policy GREEDYVIOLATE which is simpler to analyze. The quantities p_i, η_i, ν_j are the same as before. The algorithm is the same as GREEDYORDER except for step (3), which we outline in Figure 3.

In GREEDYVIOLATE, the cost budget is checked only *after* fully executing a policy \mathcal{A}_j . Therefore, the policy could violate the budget constraint by at most the total exploration cost c_{\max} of one arm. It is easy to ensure in the LP formulation that $c_{\max} \leq C$.

Theorem 4.1. GREEDYVIOLATE spends cost at most $C + c_{\max}$ and achieves gain $\frac{OPT}{4}$.

Proof. We have $\gamma^* = \sum_i \eta_i$, and $\sum_i p_i \leq 1$. Since $\sum_i \mathbf{E}[C_i] \leq C$, we have $\sum_i \nu_i \leq 1$. We note that the random variables corresponding to different i are independent.

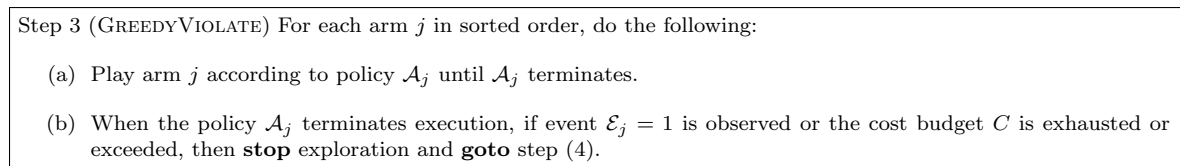


Figure 3: The GREEDYVIOLATE policy.

We therefore have $\sum_i(p_i + \nu_i) \leq 2$. Re-number the arms according to the sorted ordering so that the first arm played is numbered 1. Let k denote the smallest integer such that $\sum_{i=1}^k(p_i + \nu_i) \geq 1$. Thus, $\sum_{i=1}^k \eta_i \geq \frac{1}{2}\gamma^*$.

Let $W_i = \sum_{j=1}^i \mathcal{E}_j$, and let $w_i = \mathbf{E}[W_i] = \sum_{j=1}^i p_j$. Let Γ_j be the exact cost seen on playing arm j . Thus $\mathbf{E}[\Gamma_j] = C_j$. Let $M_i = \sum_{j=1}^i \Gamma_j / C$ and let $m_i = \mathbf{E}[M_i] = \sum_{j=1}^i C_j / C = \sum_{j=1}^i \nu_j$. Let $m_0 = w_0 = 0$.

For $i \leq k$, we have $m_{i-1} + w_{i-1} \leq 1$. Let \mathcal{Y}_i denote the event that GREEDYVIOLATE plays arm i . This corresponds to $W_{i-1} = 0$ and $M_{i-1} < 1$, which translates to $W_{i-1} + M_{i-1} < 1$. By Markov's inequality, we have $\Pr[\mathcal{Y}_i] = \Pr[W_{i-1} + M_{i-1} < 1] \geq 1 - w_{i-1} - m_{i-1}$.

If arm i is played, it yields reward X_i that directly contributes to the exploitation gain. Since X_i is independent of \mathcal{Y}_i , the expected gain of GREEDYVIOLATE can be bounded by linearity of expectation as follows.

$$\text{Gain of GREEDYVIOLATE} = \mathcal{G} \geq \sum_{i=1}^k (1 - m_{i-1} - w_{i-1}) \eta_i$$

We now follow the proof idea in [10]. Consider the arms $1 \leq i \leq k$ as deterministic items with item i having profit $v_i = \frac{\eta_i}{\gamma^*/2}$ and size $s_i = \nu_i + p_i$. We therefore have $\sum_{i=1}^k v_i \geq 1$ and $\sum_{i=1}^{k-1} s_i \leq 1$. Also note that $m_i + w_i = \sum_{j=1}^i s_j$.

Suppose these items are placed into a knapsack of size 1 in decreasing order of $\frac{v_i}{s_i}$ with the last item possibly being fractionally placed. This is the same ordering that the algorithm uses to play the arms. Let $\Phi(r)$ denote the profit when size of the knapsack filled is $r \leq 1$. We have $\Phi(1) \geq 1$ because $\sum_{i=1}^k \eta_i \geq \frac{1}{2}\gamma^*$. Plot the function $\Phi(r)$ as a function of r . This plot connects the points $\{(0, 0), (m_1 + w_1, v_1), (m_2 + w_2, v_1 + v_2), \dots, (1, \Phi(1))\}$. This function is concave, therefore the area under the curve is at least $\frac{\Phi(1)}{2} \geq 1/2$. However, the area under this curve is at most

$$v_1 + v_2(1 - m_1 - w_1) + \dots + v_k(1 - m_{k-1} - w_{k-1}) \leq \frac{\mathcal{G}}{\gamma^*/2}.$$

Therefore, $\frac{\mathcal{G}}{\gamma^*/2} \geq \frac{1}{2}$. Since $OPT \leq \gamma^*$, \mathcal{G} is at least $\frac{OPT}{4}$. \square

Theorem 4.2. *The GREEDYORDER policy with cost budget C achieves gain at least $\frac{OPT}{4}$.*

Proof. Consider the GREEDYVIOLATE policy. This policy could exceed the cost budget because the budget was checked only at the end of execution of policy \mathcal{A}_i for arm i . Now suppose the play for arm i reaches state $u \in \mathbf{T}_i$, and the next decision of GREEDYVIOLATE involves playing arm i and this would exceed the cost budget. The GREEDYVIOLATE policy continues to play arm i according to \mathcal{A}_i and when the play is finished, it checks the budget constraint, realizes that the budget is exhausted, stops, and chooses arm i for exploitation. Suppose the policy was modified so that instead of the decision to play arm i further at state u , the policy instead checks the budget, realizes it is not sufficient for the next play, stops, and chooses arm i for exploitation. This new policy is precisely GREEDYORDER.

Note now that conditioned on reaching node u with the next decision of GREEDYVIOLATE being to play arm i , so that the policies GREEDYVIOLATE and GREEDYORDER diverge in their next action, both policies choose arm i for exploitation. By Claim 2.2, the gain from choosing arm i for exploitation at node u is the same as the expected gain from playing the arm further and then choosing it for exploitation. Therefore, the expected gain of both policies is identical, and the theorem follows. \square

5 Lagrangean Gain

The nice aspect of the proof of Theorem 4.1 is that it does not necessarily require $X_i \geq 0$. As long as $\mathbf{E}[X_i] \geq 0$, the proof holds. This has implications when we are considering the the Lagrangean variant of the problem, where the net profit is the exploitation gain *minus* the total cost spent in exploration. The goal now is to design an adaptive exploration phase such that the expected net profit is maximized. The variables are identical to the previous formulation, but there is no budget constraint. The linear program relaxation is below:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^n \sum_{u \in \mathbf{T}_i} (x_u \zeta_i(u) - c_i z_u) \\ \sum_{i=1}^n \sum_{u \in \mathbf{T}_i} x_u & \leq 1 \\ \sum_{v:u \in D(v)} z_v \mathbf{P}_{vu} & = w_u \quad \forall i, u \in \mathbf{T}_i \setminus \{\rho_i\} \\ x_u + z_u & \leq w_u \quad \forall u \in \mathbf{T}_i, \forall i \\ x_u, z_u, w_u & \in [0, 1] \quad \forall u \in \mathbf{T}_i, \forall i \end{aligned}$$

Let OPT = optimal net profit and γ^* = optimal LP solution. The next is similar to Claim 3.1.

Claim 5.1. $OPT \leq \gamma^*$.

From this LP optimum $\langle w_u^*, x_u^*, z_u^* \rangle$, the policy \mathcal{A}_i is constructed as described in Figure 1, and the quantities $\mathcal{E}_i, C_i, X_i, p_i, \nu_i, \eta_i$ are obtained as described in the beginning of Section 4. Let $r, v, Y_i = X_i - C_i$ denote the net profit of playing arm i according to \mathcal{A}_i .

Claim 5.2. For any arm i , $\mathbf{E}[Y_i] = (\eta_i - \nu_i) \geq 0$.

Proof. For each i , since all $\zeta_i(u) \geq 0$, setting $x_{\rho_i} \leftarrow \sum_{u \in \mathbf{T}_i} x_u$, $w_{\rho_i} \leftarrow 1$, and $z_u \leftarrow 0$ for $u \in \mathbf{T}_i$ yields a feasible non-negative solution. The LP optimum will therefore guarantee that the term $\sum_{u \in \mathbf{T}_i} (x_u^* \zeta_i(u) - c_i z_u^*) \geq 0$. Therefore, $\mathbf{E}[Y_i] = \sum_{u \in \mathbf{T}_i} (x_u^* \zeta_i(u) - c_i z_u^*) \geq 0$ for all i . \square

The GREEDYORDER policy orders the arms in decreasing order of $\frac{\eta_i - \nu_i}{p_i}$, and plays them according to their respective \mathcal{A}_i until some $\mathcal{E}_i = 1$.

Theorem 5.3. The expected net profit of GREEDYORDER is at least $OPT/2$.

Proof. The LP solution yields $\sum_i \mathbf{E}[\mathcal{E}_i] = \sum_i p_i \leq 1$ and $\sum_i \mathbf{E}[Y_i] = \sum_i (\eta_i - \nu_i) = \gamma^*$. Re-number the arms according to the sorted ordering of $\frac{\eta_i - \nu_i}{p_i}$ so that the first arm played is numbered 1.

Let $W_i = \sum_{j=1}^i \mathcal{E}_j$, and let $w_i = \mathbf{E}[W_i] = \sum_{j=1}^i p_j$. Let $w_0 = 0$. For $i \leq n$, we have $w_{i-1} \leq 1$. Let \mathcal{Z}_i denote the event that GREEDYORDER plays arm i . This corresponds to $W_{i-1} = 0$. By Markov's inequality, we have $\Pr[\mathcal{Z}_i] = \Pr[W_{i-1} < 1] \geq 1 - w_{i-1}$.

If arm i is played, it yields net profit $Y_i = X_i - C_i$. This implies the net profit of GREEDYORDER is $\sum_i \mathcal{Z}_i Y_i$. Since Y_i is independent of \mathcal{Z}_i , and since Claim 5.2 implies $\mathbf{E}[Y_i] \geq 0$, the expected net profit \mathcal{G} of GREEDYORDER can be bounded by linearity of expectation as follows.

$$\mathcal{G} = \sum_i \Pr[\mathcal{Z}_i] \mathbf{E}[Y_i] \geq \sum_i (1 - w_{i-1}) (\eta_i - \nu_i)$$

We now follow the proof idea in [10]. Consider the arms $1 \leq i \leq n$ as deterministic items with item i having profit $v_i = \frac{\eta_i - \nu_i}{\gamma^*}$ and size $s_i = p_i$. We therefore have $\sum_i v_i = 1$ and $\sum_i s_i \leq 1$. Also note that $w_i = \sum_{j=1}^i s_j$. Using the same proof idea as in Theorem 4.1, it is easy to see that $\frac{\mathcal{G}}{\gamma^*} \geq \frac{1}{2}$. Since $OPT \leq \gamma^*$, \mathcal{G} is at least $\frac{OPT}{2}$. \square

6 Concave Utility Functions

The above framework in fact solves the more general problem of maximizing any concave stochastic objective function over the rewards of the arms subject to a (deterministic) packing constraint. In what follows, we extend our arguments in the previous section to develop approximation algorithms for all positive concave utility maximization problems in this exploration-exploration setting. Suppose arm i has a gain function $g_i(y, r)$ where $y \in [0, 1]$ denotes the weight assigned to it in the exploitation phase, and r is the reward on playing that arm. For fixed r , the gain function is an arbitrary positive non-decreasing concave function of y . Furthermore, for a fixed y , this function is non-decreasing in r . Given a realization \bar{s} of outcomes, suppose the posterior reward distribution of arm i is $\mathcal{R}_i(\bar{s}_i)$. Then, if y_i is the weight assigned to this arm in the exploitation phase, the contribution of this arm to the exploitation gain is $\mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})} [\mathbf{E}_{r \in R_i} [g_i(y_i, r)]]$. The assignment of weights is subject to a deterministic packing constraint $\sum_i \sigma_i y_i \leq B$, where $\sigma_i \in [0, B]$. Therefore, for a given outcome \bar{s} of exploration, the exploitation gain is given by the convex program $\mathcal{P}(\vec{\mathcal{R}}(\bar{s}))$:

$$\begin{aligned} \max \quad & \sum_{i=1}^n \mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})} [\mathbf{E}_{r \in R_i} [g_i(y_i, r)]] \\ \text{s.t.} \quad & \sum_{i=1}^n \sigma_i y_i \leq B, \forall i \ y_i \in [0, 1] \end{aligned}$$

The goal as before is to design an adaptive exploration phase so that $\mathbf{E}_{\bar{s}}[\mathcal{P}(\vec{\mathcal{R}}(\bar{s}))]$ is maximized, where the expectation is over the outcomes \bar{s} of exploration and cost of exploring \bar{s} is at most C .

- For the maximum reward problem the $\mathcal{P}(\vec{\mathcal{R}}(\bar{s}))$ is $\max \sum_{i=1}^n \mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})} [\mathbf{E}_{r \in R_i} [y_i r]]$ s.t. $\sum_{i=1}^n y_i \leq 1, \forall i \ y_i \in [0, 1]$. The $g_i(y, r) = yr$ for all i and $\sigma_i = 1$. Since the objective is linear and since $\mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})} [\mathbf{E}_{r \in R_i} [r]] = \mathbf{E}[\Theta_i(\bar{s})]$, the above is equivalent to: $\max \sum_{i=1}^n y_i \mathbf{E}[\Theta_i(\bar{s})]$ subject to $\sum_{i=1}^n y_i \leq 1$ and $y_i \in [0, 1]$. This is simply $\max_i \mathbf{E}[\Theta_i(\bar{s})]$.
- Suppose we wish to choose the m best rewards, the program $\mathcal{P}(\vec{\mathcal{R}})$ is given by:

$$\begin{aligned} \max \quad & \sum_{i=1}^n \mathbf{E}_{R_i \in \mathcal{R}_i(\bar{s})} [\mathbf{E}_{r \in R_i} [y_i r]] \\ \text{s.t.} \quad & \sum_{i=1}^n y_i \leq m, \forall i \ y_i \in [0, 1] \end{aligned}$$

This chooses those m arms with largest $\mathbf{E}[\Theta_i]$ after the exploration phase. Note that we can also conceive of a scenario where the c_i correspond to cost of “pilot studies” and each treatment i requires cost σ_i for large scale studies. This would lead us to a KNAPSACK type problem where σ_i are now the “sizes”.

6.1 Linear Program

The DAGs \mathbf{T}_i , the probabilities \mathbf{p}_{uv} , and the posteriors $\mathcal{R}_i(u)$ are defined just as in Section 2. For small constant $\epsilon > 0$, let $L = \frac{n}{\epsilon}$. Discretize the domain $[0, 1]$ in multiples of $1/L$. For $l \in \{0, 1, \dots, L\}$, let $\zeta_i(u, l) = \mathbf{E}_{R_i \in \mathcal{R}_i(u)} [\mathbf{E}_{r \in R_i} [g(l/L, r)]]$. This corresponds to the contribution of arm i to the exploitation gain on allocating weight $y_i = l/L$.

Define the following linear program:

$$\text{Max} \quad \sum_{i=1}^n \sum_{u \in \mathbf{T}_i} \sum_{l=0}^L x_{ul} \zeta_i(u, l)$$

Policy \mathcal{A}_i : If arm i is currently in state u , choose $q \in [0, w_u^*]$ u.a.r. and do one of the following:

1. If $q \in [0, z_u^*]$, **then** play the arm.
2. **else** stop executing \mathcal{A}_i . Find the smallest $l \geq 0$ such that $q \leq z_u^* + \sum_{k=0}^l x_{uk}^*$. Set $\mathcal{E}_i = \frac{l}{L}$ and $X_i = \zeta_i(u, l)$.

Figure 4: The policy \mathcal{A}_i for concave gain functions.

$$\begin{aligned}
\sum_{i=1}^n c_i \left(\sum_{u \in \mathbf{T}_i} z_u \right) &\leq C \\
\sum_{i=1}^n \sigma_i \left(\sum_{u \in \mathbf{T}_i} \sum_{l=0}^L l x_{ul} \right) &\leq BL(1 + \epsilon) \\
\sum_{v: u \in D(v)} z_v \mathbf{P}_{vu} &= w_u \quad \forall i, u \in \mathbf{T}_i \setminus \{\rho_i\} \\
z_u + \sum_{l=0}^L x_{ul} &\leq w_u \quad \forall u \in \mathbf{T}_i, \forall i \\
w_u, x_{ul}, z_u &\in [0, 1] \quad \forall u \in \mathbf{T}_i, \forall i, l
\end{aligned}$$

Let γ^* be the optimal LP value and OPT = value of the optimal adaptive exploration policy.

Lemma 6.1. $OPT \leq \gamma^*$.

Proof. In the optimal solution, let w_u denote the probability that the policy reaches state $u \in \mathbf{T}_i$, and let z_u denote the probability of reaching state $u \in \mathbf{T}_i$ and playing arm i in this state. For $l \geq 1$, let x_{ul} denote the probability of stopping exploration at $u \in \mathbf{T}_i$ and allocating weight $y_i \in (\frac{l-1}{L}, \frac{l}{L}]$ to arm i . All the constraints are straightforward, except the constraint involving B . Observe that if the weight assignments y_i in the optimal solution were rounded up to the nearest multiple of $1/L$, then the total size of any assignment increases by at most ϵB since all $s_i \leq B$. Therefore, this constraint is satisfied. Using the same rounding up argument, if the weight satisfies $y_i \in (\frac{l-1}{L}, \frac{l}{L}]$, then the contribution of arm i to the exploitation gain is upper bounded by $\zeta_i(u, l)$ since the function $g_i(y, r)$ is non-decreasing in y for every r . Therefore, the proof follows. \square

6.2 Exploration Policy

Let $\langle w_u^*, x_{ul}^*, z_u^* \rangle$ denote the optimal solution to the LP . Assume $w_{\rho_i}^* = 1$ for all i . Also w.l.o.g, $z_u^* + \sum_{l=0}^L x_{ul}^* = w_u^*$ for all $u \in \mathbf{T}_i$. The LP solution yields a natural (infeasible) exploration policy \mathcal{A} consisting of one independent policy \mathcal{A}_i per arm i . Policy \mathcal{A}_i is described in Figure 4.

The policy \mathcal{A}_i is independent of the states of the other arms. It is easy to see by induction that if state $u \in \mathbf{T}_i$ is reached by the policy with probability w_u^* , then state $u \in \mathbf{T}_i$ is reached *and* arm i is played with probability z_u^* . Let random variable C_i denote the cost of executing \mathcal{A}_i . Denote this overall policy \mathcal{A} – this corresponds to one independent decision policy \mathcal{A}_i (determined by $\langle w_u^*, x_{ul}^*, z_u^* \rangle$) per arm. It is easy to see that the following hold for \mathcal{A} :

1. $\mathbf{E}[C_i] = c_i \sum_{u \in \mathbf{T}_i} z_u^*$ so that $\sum_i \mathbf{E}[C_i] \leq C$.
2. $\mathbf{E}[\mathcal{E}_i] = \frac{1}{L} \sum_{u \in \mathbf{T}_i} \sum_{l=0}^L l x_{ul}^*$
 $\Rightarrow \sum_i \sigma_i \mathbf{E}[\mathcal{E}_i] \leq B(1 + \epsilon)$.
3. $\mathbf{E}[X_i] = \sum_{u \in \mathbf{T}_i} \sum_{l=0}^L x_{ul}^* \zeta_i(u, l)$
 $\Rightarrow \sum_i \mathbf{E}[X_i] = \gamma^*$.

The GREEDYORDER policy is presented in Figure 5. We again use an infeasible policy GREEDYVIOLATE which is simpler to analyze. The algorithm is the same as GREEDYORDER except for step (3), where violation of $\sum_i \sigma_i \mathcal{E}_i \leq B$ or the budget constraint are only checked after the policy \mathcal{A}_j terminates.

Theorem 6.2. GREEDYVIOLATE spends cost at most $C + c_{\max}$ and has gain $\frac{OPT}{8}(1 - \epsilon)$.

Algorithm GREEDYORDER

1. For each arm i and its corresponding policy \mathcal{A}_i , let $p_i = \frac{\sigma_i \mathbf{E}[\mathcal{E}_i]}{B}$, $\eta_i = \mathbf{E}[X_i]$; and $\nu_i = \mathbf{E}[C_i]/C$.
2. Order the arms in decreasing order of $\frac{\eta_i}{p_i + \nu_i}$ and choose the arms to play in this order.
3. For each arm j in sorted order, play arm j according to \mathcal{A}_j as follows until \mathcal{A}_j terminates:
 - (a) If the next play according to \mathcal{A}_j would violate the budget constraint, then set $\mathcal{E}_j \leftarrow 1$, **stop** exploration, and **goto** step (4).
 - (b) When \mathcal{A}_j has terminated, if $\sum_i \sigma_i \mathcal{E}_i \geq B$, then **stop** exploration and **goto** step (4).
 - (c) Else, play arm j according to policy \mathcal{A}_j and **goto** step (3a).
4. **Exploitation:** Scale down \mathcal{E}_i by a factor of 2.

Figure 5: The GREEDYORDER policy for concave functions.

Proof. We have $\gamma^* = \sum_i \eta_i$, and $\sum_i p_i \leq 1 + \epsilon$. Also $C_i \leq C$ for all i , enforced by constraints in the LP and $\sum_i \mathbf{E}[C_i] \leq C$. We therefore have $\sum_i \nu_i \leq 1$. This implies $\sum_i (p_i + \nu_i) \leq 2 + \epsilon$. Now using the same proof as Theorem 4.1, we obtain the gain \mathcal{G} of GREEDYVIOLATE according to the weight assignment \mathcal{E}_i at the end of Step (3) is at least $\frac{OPT}{4}(1 - \epsilon)$. This weight assignment could be infeasible because of the last arm, so that the \mathcal{E}_i only satisfy $\sum_i \sigma_i \mathcal{E}_i \leq 2B$. This is made feasible in Step (4) by scaling all \mathcal{E}_i down by a factor of 2. Since the functions $g_i(y, r)$ are concave in y , the exploitation gain reduces by a factor of 1/2 because of scaling down. \square

Theorem 6.3. *The GREEDYORDER policy with cost budget C achieves gain at least $\frac{OPT}{8}(1 - \epsilon)$.*

Proof. Consider the GREEDYVIOLATE policy. Now suppose the play for arm i reaches state $u \in \mathbf{T}_i$, and the next decision of GREEDYVIOLATE involves playing arm i and this would exceed the cost budget. Conditioned on this next decision, GREEDYORDER sets $\mathcal{E}_i = 1$ and stops exploration. In this case, the exploitation gain of GREEDYORDER from arm i is at least the expected exploitation gain of GREEDYVIOLATE for this arm. Therefore, for the assignments at the end of Step (3), the gain of GREEDYORDER is at least $\frac{OPT}{4}(1 - \epsilon)$. Since Step (4) scales the \mathcal{E} 's down by a factor of 2, the theorem follows. \square

Acknowledgement: We would like to thank Jen Burge, Vincent Conitzer, Ashish Goel, Ronald Parr, Fernando Pereira, and Saswati Sarkar for helpful discussions.

References

- [1] N. Abe and M. K. Warmuth. On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning*, 9:205–260, 1992.
- [2] Q. An, H. Li, X. Liao, and L. Carin. Active feature acquisition with POMDP models. *Submitted to Pattern Recognition Letters*, 2006.
- [3] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2001.
- [4] D. Bertsimas, D. Gamarnik, and J. Tsitsiklis. Performance of multiclass markovian queueing networks via piecewise linear Lyapunov functions. *Annals of Applied Probability*, 11(4):1384–1428, 2002.
- [5] D. Bertsimas and J. Nino-Mora. Conservation laws, extended polymatroids and multi-armed bandit problems: A unified polyhedral approach. *Math. of Oper. Res.*, 21(2):257–306, 1996.
- [6] D. Bertsimas, I. Paschalidis, and J. N. Tsitsiklis. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *Annals of Applied Probability*, 4(1):43–75, 1994.
- [7] M. Charikar, C. Chekuri, and M. Pál. Sampling bounds for stochastic optimization. In *APPROX-RANDOM*, pages 257–269, 2005.
- [8] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press, 1995.
- [9] V. Conitzer and T. Sandholm. Definition and complexity of some basic metareasoning problems. In *IJCAI*, pages 1099–1106, 2003.
- [10] B. Dean. *Approximation Algorithms for Stochastic Scheduling Problems*. PhD thesis, MIT, 2005.
- [11] B. C. Dean, M. X. Goemans, and J. Vondrak. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 208–217, 2004.
- [12] B. C. Dean, M. X. Goemans, and J. Vondrák. Adaptivity and approximation for stochastic packing problems. In *Proc. 16th ACM-SIAM Symp. on Discrete algorithms*, pages 395–404, 2005.
- [13] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in statistics (European Meeting of Statisticians)*, 1972.
- [14] A. Goel, S. Guha, and K. Munagala. Asking the right questions: Model-driven optimization using probes. In *Proc. ACM Symp. on Principles of Database Systems*, 2006.
- [15] A. Goel and P. Indyk. Stochastic load balancing and related problems. In *Proc. Symp. on Foundations of Computer Science*, 1999.
- [16] S. Guha and K. Munagala. Model driven optimization using adaptive probes. *Proc. ACM-SIAM Symp. on Discrete Algorithms*, 2007.

- [17] S. Guha, K. Munagala, and S. Sarkar. Jointly optimal probing and transmission strategies for multi-channel wireless systems. *Submitted to IEEE Transactions on Information Theory*, 2006. Available at <http://www.cs.duke.edu/~kamesh/partialinfo.pdf>.
- [18] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, 20(3):197–243, 1995.
- [19] R. Karp and R. Kleinberg. Noisy binary search and its applications. In *Proc. 18th ACM-SIAM Symp. on Discrete Algorithms (SODA 2007)*, pages 881–890, 2007.
- [20] M. J. Kearns, Y. Mansour, and A. Y. Ng. Approximate planning in large POMDPs via reusable trajectories. In *NIPS*, pages 1001–1007, 1999.
- [21] J. Kleinberg, Y. Rabani, and É. Tardos. Allocating bandwidth for bursty connections. *SIAM J. Comput.*, 30(1), 2000.
- [22] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. on Optimization*, 12(2):479–502, 2002.
- [23] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. *Proc. 21st Conf. on Uncertainty in Artificial Intelligence*, 2005.
- [24] S. H. Low and D. E. Lapsley. Optimization flow control-I: Basic algorithm and convergence. *IEEE/ACM Trans. Netw.*, 7(6):861–874, 1999.
- [25] O. Madani, D. J. Lizotte, and R. Greiner. Active model selection. In *UAI '04: Proc. 20th Conf. on Uncertainty in Artificial Intelligence*, pages 357–365, 2004.
- [26] R. H. Mohring, A. S. Schulz, and M. Uetz. Approximation in stochastic scheduling: the power of LP-based priority policies. *J. ACM*, 46(6):924–942, 1999.
- [27] A. Moore, J. Schneider, J. Boyan, and M. Lee. Q2: Memory-based active learning for optimizing noisy continuous functions. *International Conference on Machine Learning*, 1998.
- [28] R. H. Myers and D. C. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments (2nd ed.)*. Wiley, 2002.
- [29] J. Schneider and A. Moore. Active learning in discrete input spaces. *The 34th Interface Symposium, Montreal, Quebec*, 2002.
- [30] D. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as discrete optimization. In *Proc. 45th Symp. on Foundations of Computer Science*, pages 228–237, 2004.
- [31] M. Skutella and M. Uetz. Scheduling precedence-constrained jobs with stochastic processing times on parallel machines. In *SODA '01: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 589–590, 2001.
- [32] J. N. Tsitsiklis. A short proof of the Gittins index theorem. *Annals of Applied Probability*, 4(1):194–199, 1994.

A Adaptivity Gap Example

A non-adaptive strategy allocates a fixed budget to each arm in advance. It then explores the arms according to these budgets (ignoring the outcome of the plays in choosing the next arm to explore), and at the end of exploration, chooses the best arm for exploitation. This is termed an *allocational strategy* in [25]. We present an example with unit costs where an adaptive strategy that dynamically allocates the budget achieves far better exploitation gain than a non-adaptive strategy.

Theorem A.1. *The adaptivity gap of the budgeted multi-armed bandit problem is $\Omega(\sqrt{n})$. Furthermore, even if we allow the non-adaptive exploration to use $\gamma > 1$ times the exploration budget, the adaptivity gap remains $\Omega(\sqrt{n/\gamma})$.*

Proof. Let $K = 3$ and let $a_1 = 0$, $a_2 = 1/n^9$ and $a_3 = 1$. Let $q = 1/\sqrt{n}$. All priors \mathcal{R}_i are i.i.d, and each has 3 distributions R_1, R_2, R_3 . R_1 is the deterministic value a_1 , R_2 is deterministically a_2 and R_3 is a_3 w.p. q and a_2 w.p. $1 - q$. In each prior \mathcal{R}_i , we have $\Pr[R_1] = 1 - q$, $\Pr[R_2] = q(1 - q)$ and $\Pr[R_3] = q^2$. All $c_i = 1$ and the total budget is $C = 5n$.

We first show that the adaptive policy chooses an arm with reward distribution R_3 with constant probability. This policy first plays each arm once and discards all arms with observed reward a_1 . With probability at least $1/2$, there are at most $2/q$ arms which survive, and at least one of these arms has underlying reward distribution R_3 . If more arms survive, choose any $2/q$ arms. The policy now plays each of the $2/q$ arms $2\sqrt{n}$ times. The probability that an arm with distribution R_3 yields reward a_3 on some play is at least once is $1 - (1 - q)^{2/q} \approx \Theta(1)$. In this case, it chooses the arm with reward distribution R_3 for exploitation. Since this happens w.p. at least a constant, the expected exploitation gain is $\Theta(q)$. Note that this is best possible to within constant factors, since $\mathbf{E}[R_3] = \Theta(q)$.

Now consider any non-adaptive policy. With probability $1 - 1/n^{\Theta(1)}$, there are at most $2 \log n$ arms with reward distribution R_3 , and at least $1/(2q)$ arms with reward distribution R_2 . Let $r \gg 2 \log n$. The strategy allocates at most $5r$ plays to at least $n(1 - 1/r)$ arms – call this set of arms T . With probability $(1 - 1/r)^{2 \log n} = \Omega(1 - (2 \log n)/r)$, all arms with reward distribution R_3 lie in this set T . For any of these arms played $O(r)$ times, with probability $1 - O(qr)$, all observed rewards will have value a_2 . This implies with probability $1 - O(qr)$, all arms with distribution R_3 yield rewards a_2 , and so do $\Omega(1/(2q))$ arms with distributions R_2 . Since these appear indistinguishable to the policy, it can at best choose one of these at random, obtaining exploitation gain $\frac{q \log n}{2(1/q)} = O(q^2 \log n)$. Since this situation happens with probability $1 - O(\log n/r)$, and with the remaining probability the exploitation gain is at most q , the strategy therefore has expected exploitation gain $O(q \log n(\frac{1}{r} + q))$. This implies the adaptivity gap is $\Omega(1/q) = \Omega(\sqrt{n})$ if we set $r = 1/q$.

Now suppose we allow the budget to be increased by a factor of $\gamma > 1$. Then the strategy would allocate at most $5\gamma r$ plays to at least $n(1 - 1/r)$ arms. By following the same argument as above, the expected gain is $O(q \log n(\frac{1}{r} + q\gamma))$. This proves the second part of the theorem. \square

B Additive Approximation

Recall that the observed outcomes form the set $0 \leq a_1 \leq a_2 \leq \dots \leq a_K = 1$.

Theorem B.1 ((Theorem 2.1)). *Let G_C denote the exploitation gain of the optimal policy with cost budget C .*

1. For $\epsilon > 0$, let $G_C(\epsilon)$ denote the gain of the optimal policy which the same cost budget which plays any arm at most $N_\epsilon = \frac{12 \log n}{\epsilon^2}$ times. For sufficiently small constant ϵ , we have $G_C(\epsilon) \geq G_C - 16\epsilon$.
2. For any $\epsilon > 0$, let $\mathcal{K} = 1/\epsilon$. Let $G_C(\epsilon)$ denote the gain of the optimal policy when the space of outcomes for playing any arm is discretized so that if $a_j \in [l\epsilon, (l+1)\epsilon)$ for $l \in \{0, 1, \dots, \mathcal{K}-1\}$, then $a_j \leftarrow (l+1)\epsilon$, so that each distribution R_i is now discrete over \mathcal{K} values. Then, $G_C(\epsilon) \geq G_C - \epsilon$.

Proof. Part (1): Let OPT denote the original optimal policy. First modify it to policy OPT' which stops playing an arm after it has been played at most N_ϵ times. This modification is performed as follows: After an arm has been played N_ϵ times, the decision tree of OPT on that arm is simulated, but the arm is not actually played (so that the observed reward of the arm does not contribute to the outcome of OPT' but contributes to the outcome for OPT). This couples the behavior of this new policy OPT' with the original policy OPT . Note that OPT' is not an optimal policy, but simply a modification of OPT . We prove the theorem on the policy OPT' .

The proof is not straightforward because the sample average of the plays is not the same as the posterior mean, the latter depending on *both* the prior distribution and the sample average. The proof idea will be to argue over a high probability portion of the prior distribution and show that for this portion, the sample average is roughly the same the the posterior mean.

Let \mathcal{S} denote the set of all outcomes of OPT , *i.e.*, the leaves of its decision tree. Denote a leaf node, let $\vec{s} = \{s_1, s_2, \dots, s_n\} \in \mathcal{S}$ where s_i denotes the sequence of observed values for arm i . Let \mathcal{T} denote the set of all outcomes for OPT' . Let $\Theta_i(\vec{t})$ and $\Theta_i(\vec{s})$ denote the posterior distributions of the mean reward given leaf nodes \vec{t} and \vec{s} respectively. Let $\mu_i(\vec{s}) = \mathbf{E}[\Theta_i(\vec{s})]$ denote the expected posterior value of the reward for arm i for leaf node \vec{s} . For $\vec{t} \in \mathcal{T}$, let $\mu_i(\vec{t}) = \mathbf{E}[\Theta_i(\vec{t})]$ denote the expected posterior mean reward for arm i in OPT' .

For any leaf node $\vec{s} \in \mathcal{S}$, let $tr(\vec{s}) = \{s'_1, s'_2, \dots, s'_k\}$ denote the set of outcomes restricted to the first N_ϵ plays. The function tr maps a set of leaves in \mathcal{S} to one leaf in \mathcal{T} . Note that if $|s_i| \leq N_\epsilon$, then $\Theta_i(\vec{s}) = \Theta_i(tr(\vec{s}))$. For outcome vector $\vec{t} \in \mathcal{T}$, let $\mathcal{O}(\vec{t}) = \{\vec{s} \in \mathcal{S} | tr(\vec{s}) = \vec{t}\}$.

Recall that Θ_i denotes the prior distribution of the mean reward of arm i . Let $\mu_i = \mathbf{E}[\Theta_i]$. For any interval $I \subseteq [0, 1]$, let $f_i(I)$ denote the probability mass of prior Θ_i in the interval I .

Let $M = \frac{1}{\epsilon}$. Split $[0, 1]$ into consecutive disjoint intervals of size ϵ and denote these intervals B_1, B_2, \dots, B_M .

Definition 5. Interval B_j is “good” for arm i if $f_i(B_j) \geq \frac{1}{n^2}$, and “bad” otherwise.

Clearly, the total probability mass of Θ_i in bad intervals is at most $\frac{M}{n^2}$.

Definition 6. 1. Let $I_x = [x - 3\epsilon, x + 3\epsilon]$, $L_x = [x - 5\epsilon, x + 5\epsilon]$, $J_x = [x - 7\epsilon, x + 7\epsilon]$, and $H_x = [x - 8\epsilon, x + 8\epsilon]$.

2. For leaf node $\vec{t} \in \mathcal{T}$ where arm i has been played N_ϵ times, let $m_i(\vec{t})$ denote the sample average of the N_ϵ outcomes.
3. A value x is termed “good” for arm i if $f_i(I_x) > 1/n^2$ and “bad” otherwise.
4. Leaf node $\vec{t} \in \mathcal{T}$ is “good” if for all arms i , either $|t_i| < N_\epsilon$, or $|t_i| = N_\epsilon$ and $m_i(\vec{t})$ is “good” for arm i . The node is otherwise termed “bad”.

We will use the following well-known bound extensively:

Lemma B.2 (Hoeffding Bound). Let X_1, X_2, \dots, X_q be independent 0/1 random variables with $\mathbf{E}[X_i] = p_i$. Let $X = \frac{1}{q} \sum_i X_i$ and $p = \mathbf{E}[X] = \frac{1}{q} \sum_i p_i$. Then, $\Pr[|X - p| > t] \leq 2 \exp(-\frac{t^2 q}{3p})$.

Claim B.3. The probability over the decision tree \mathcal{T} that $\vec{t} \in \mathcal{T}$ is “bad” is $O(1/n^2)$.

Proof. Suppose $m_i(\vec{t}) \in B_j$. Clearly, if $f_i(I_{m_i(\vec{t})}) \leq 1/n^2$, then B_{j-1}, B_j, B_{j+1} are all “bad” intervals. Now consider the process by which $m_i(\vec{t})$ is generated. With probability at least $1 - M/n^2$, the mean μ_i for arm i is chosen from a “good” interval B_k , and in this case, with probability at least $1 - P_\epsilon$, the sample average $m_i(\vec{t})$ is within the interval $[\mu_i - \epsilon, \mu_i + \epsilon]$. In this case, $m_i(\vec{t})$ is “good”. By Hoeffding bounds, the probability P_ϵ is at most:

$$P_\epsilon \leq 2e^{-\frac{12 \log n}{\epsilon^2} \frac{\epsilon^2}{3}} = O\left(\frac{1}{n^4}\right)$$

By union bounds, the probability that an observed $m_i(\vec{t})$ is “bad” is at most $M/n^2 + 1/n^4 = O(1/n^2)$.

Recall that a leaf node $\vec{t} \in \mathcal{T}$ is “good” if for all i , either $|t_i| < N_\epsilon$ or if $|t_i| = N_\epsilon$, then $m_i(\vec{t})$ is “good”. By union bounds, the probability that a leaf node \vec{t} is good is therefore at least $1 - O(1/n)$. \square

Definition 7. Let \mathcal{E} denote the set of all “good” outcomes in \mathcal{T} .

For $\vec{t} \in \mathcal{E}$, if $|t_i| < N_\epsilon$, then $\mu_i(\vec{s}) = \mu_i(\vec{t})$ for every $\vec{s} \in \mathcal{O}(\vec{t})$. Else if $|t_i| = N_\epsilon$, then we have the following claim:

Claim B.4. For $\vec{t} \in \mathcal{E}$ and $|t_i| = N_\epsilon$, the following hold for the posterior distribution $\Theta_i(\vec{t})$:

- (i) $\Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})}] = O(1/n^2)$.
- (ii) $\mu_i(\vec{t}) \in H_{m_i(\vec{t})}$.
- (iii) $\Pr_{\vec{s} \in \mathcal{O}(\vec{t})}[|\mu_i(\vec{t}) - \mu_i(\vec{s})| \geq 15\epsilon] = O(1/n^2)$.

Proof. Part (i): Let A denote the random variable corresponding to the sample average for N_ϵ plays for arm i . For any $x \in [0, 1]$, by Hoeffding’s bounds, we have:

$$\Pr[A \in L_x | \Theta_i(\vec{t}) \notin J_x] \leq P_\epsilon \leq \frac{1}{n^4}$$

Again, by Hoeffding bounds:

$$\Pr[A \in L_x] \geq f_i(I_x) - o(1/n^2)$$

If $x \in [0, 1]$ is “good”, then $f_i(I_x) \geq 1/n^2$. Combining the above inequalities, we have for any “good” x :

$$\Pr[A \in L_x] = \Omega(1/n^2)$$

The next inequality is trivial: $\Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})}] \leq 1$.

Since $\vec{t} \in \mathcal{E}$, this implies $m_i(\vec{t})$ is “good”. Setting $x = m_i(\vec{t})$ and combining the above inequalities via Bayes’ rule, we have for any $\vec{t} \in \mathcal{E}$:

$$\begin{aligned} \Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})} | A = m_i(\vec{t})] &\leq \Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})} | A \in L_{m_i(\vec{t})}] \\ &\leq \frac{\Pr[A \in L_{m_i(\vec{t})} | \Theta_i(\vec{t}) \notin J_{m_i(\vec{t})}] \times \Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})}]}{\Pr[A \in L_{m_i(\vec{t})}]} \\ &= O(1/n^2) \end{aligned}$$

Part (ii): At leaf node $\vec{t} \in \mathcal{E}$, by Part (i), we have $\Pr[\Theta_i(\vec{t}) \notin J_{m_i(\vec{t})}] = O(1/n^2)$. Since $\Theta_i(\vec{t}) \in [0, 1]$ and $\mu_i(\vec{t}) = \mathbf{E}[\Theta_i(\vec{t})]$, the claim follows.

Part (iii): Consider the worst-case scenario where OPT plays arm i infinitely many times beyond \vec{t} and resolves the mean exactly. Even in this case, by Part (i), we must have $\Pr_{\vec{s} \in \mathcal{O}(\vec{t})} [|m_i(\vec{t}) - \mu_i(\vec{s})| \geq 7\epsilon] = O(1/n^2)$. We also have by Part (ii) that $|\mu_i(\vec{t}) - m_i(\vec{t})| \leq 8\epsilon$. Combining these two inequalities yields the proof. \square

The above implies that for every $\vec{t} \in \mathcal{E}$, we have $\Pr_{\vec{s} \in \mathcal{O}(\vec{t})} [|\mu_i(\vec{s}) - \mu_i(\vec{t})| \geq 15\epsilon] = O(1/n^2)$ for all arms i . Therefore, we have $\Pr_{\vec{s} \in \mathcal{O}(\vec{t})} [|\max_i \mu_i(\vec{s}) - \max_i \mu_i(\vec{t})| \geq 15\epsilon] = O(1/n)$ by union bounds. We also have $\Pr[\mathcal{E}] = 1 - O(1/n)$. Therefore, $\Pr_{\vec{t} \in \mathcal{T}, \vec{s} \in \mathcal{O}(\vec{t})} [|\max_i \mu_i(\vec{s}) - \max_i \mu_i(\vec{t})| \geq 15\epsilon] = O(1/n)$.

Let \mathcal{G} denote the random variable corresponding to the gain of the original policy OPT and \mathcal{G}' denote the same for the modified policy OPT' . We have $G_C = \mathbf{E}[\mathcal{G}]$ and $G_C(\epsilon) = \mathbf{E}[\mathcal{G}']$. From the above, we have $\Pr[|\mathcal{G} - \mathcal{G}'| \geq 15\epsilon] = O(1/n)$, where the probability is over $\vec{t} \in \mathcal{T}$, and $\vec{s} \in \mathcal{O}(\vec{t})$. Since $|\mathcal{G} - \mathcal{G}'| \leq 1$ always, we have $G_C(\epsilon) \geq G_C - 16\epsilon$.

Part (2): We now show that the space of rewards can be discretized to $O(1/\epsilon)$ values by losing an additional additive $O(\epsilon)$ in the exploitation gain. Let $\mathcal{K} = 1/\epsilon$. The space of outcomes for playing any arm is discretized so that if $a_j \in [l\epsilon, (l+1)\epsilon)$ for $l \in \{0, 1, \dots, \mathcal{K} - 1\}$, then $a_j \leftarrow (l+1)\epsilon$. Each distribution R_i is now discrete over \mathcal{K} values. Note that this new discretized instance stochastically dominates the original instance. If the decision tree of the original instance is used for the new instance, the exploitation gain is at least as large. Now, the optimal exploitation strategy for the new instance yields a feasible solution to the original instance whose exploitation gain is only ϵ smaller (since all reward values in the new instance are within additive ϵ of the original instance). This shows that the discretization loses at most an additive ϵ in the exploitation gain. \square