The Irrevocable Multi-Armed Bandit Problem

Vivek F. Farias [*] Ritesh Madan [†]

22 June 2008
Revised: 16 June 2009

Abstract

This paper considers the multi-armed bandit problem with multiple simultaneous arm pulls and the additional restriction that we do not allow recourse to arms that were pulled at some point in the past but then discarded. This additional restriction is highly desirable from an operational perspective and we refer to this problem as the 'Irrevocable Multi-Armed Bandit' problem. We observe that natural modifications to well known heuristics for multi-armed bandit problems that satisfy this irrevocability constraint have unsatisfactory performance, and thus motivated introduce a new heuristic: the 'packing' heuristic. We establish through numerical experiments that the packing heuristic offers excellent performance, even relative to heuristics that are not constrained to be irrevocable. We also provide a theoretical analysis that studies the 'price' of irrevocability i.e. the performance loss incurred in imposing the constraint we propose on the multi-armed bandit model. We show that this performance loss is uniformly bounded for a general class of multi-armed bandit problems, and also indicate its dependence on various problem parameters. Finally, we obtain a computationally fast algorithm to implement the packing heuristic; the algorithm renders the packing heuristic computationally cheaper than methods that rely on the computation of Gittins indices.

1. Introduction

Consider the operations of a 'fast-fashion' retailer such as Zara or H&M. Such retailers have developed and invested in merchandise procurement strategies that permit lead times for new fashions as short as two weeks. As a consequence of this flexibility, such retailers are able to adjust the assortment of products offered on sale at their stores to quickly adapt to popular fashion trends. In particular, such retailers use weekly sales data to refine their estimates of an item's popularity, and based on such revised estimates weed out unpopular items, or else re-stock demonstrably popular ones on a week-by-week basis. In view of the great deal of a-priori uncertainty in the popularity of a new fashion and the speed at which fashion trends evolve, the fast-fashion operations model is highly desirable and emerging as the de-facto operations model for large fashion retailers.

Among other things, the fast-fashion model relies crucially on an effective technology to learn from purchase data, and adjust product assortments based on such data. Such a technology must strike a balance between 'exploring' potentially successful products and 'exploiting' products that

[*]Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, email :vivekf@mit.edu

[†]Qualcomm-Flarion Technologies, email :rkmadan@stanfordalumni.org

are demonstrably popular. A convenient mathematical model within which to design algorithms capable of accomplishing such a task is that of the multi-armed bandit. While we defer a precise mathematical discussion to a later section, a multi-armed bandit consists of multiple (say n) 'arms', each corresponding to a *Markov Decision Process*. As a special case, one may think of each arm as an independent Bernoulli random variable with an uncertain bias specified via some prior distribution. At each point in time, one may 'pull' up to a certain number of arms (say $k < n$) simultaneously. For each arm pulled, we modify our estimate of its bias based on its realization and earn a reward proportional to its realization. We neither learn about, nor earn rewards from arms that are not pulled. The multi-armed bandit problem requires finding a policy that adaptively selects k arms to pull at every point in time with an objective of maximizing total expected reward earned over some finite time horizon or alternatively, discounted rewards earned over an infinite horizon or perhaps, even long term average rewards.

The multi-armed bandit model while general and immensely useful, fails to capture an important restriction one faces in several applications. In particular, in a number of applications, the act of 'pulling' an arm that has been pulled in the past but discarded in favor of another arm is undesirable or unacceptable. Ignoring references to the extant literature for now, examples of such applications include:

1. Fast Fashion: The fixed costs associated with the introduction of a new product make frequent changes in the assortment of products offered undesirable. More importantly, fast fashion retailers rely heavily on discouraging their customers from delaying/ strategizing on the timing of their purchase decisions. They accomplish this by adhering to strict restocking policies; reintroduction of an old product is undesirable from this viewpoint.

2. Call-Center Hiring: Given the rich variety of tasks call-center workers might face, recent research has raised the possibility of 'data-driven' hiring/ staff allocation decisions at call-centers. In this setting, the act of discarding an arm is equivalent to a firing or reassignment decision; it is clear that such decisions are difficult to reverse.

3. Clinical Trials: A classical application of the bandit model, the act of discarding an arm in this setting is equivalent to the discontinuation of trials on a particular treatment. The ethical objections to administering treatments that may be viewed as inferior play a critical role in the design of such trials and it is reasonable to expect that re-starting trials on a procedure after a hiatus might well raise ethical concerns.

This paper considers the multi-armed bandit problem with an additional restriction: we require that decisions to remove an arm from the set of arms currently being pulled be 'irrevocable'. That is, we do not allow recourse to arms that were pulled at some point in the past but then discarded. We refer to this problem as the *Irrevocable Multi-Armed Bandit Problem*. We introduce a novel heuristic we call the 'packing' heuristic for this problem. The packing heuristic establishes a static ranking of bandit arms based on a measure of their potential value relative to the time required to realize that value, and pulls arms in the order prescribed by this ranking. For an arm currently being pulled, the heuristic may either choose to continue pulling that arm in the next time step or else discard the arm in favor of the next highest ranked arm not currently being pulled. Once discarded, an arm will *never* be chosen again hence satisfying the irrevocability constraint. We demonstrate via computational experiments that the use of the packing heuristic incurs a small

performance loss relative to an optimal bandit policy without the irrevocability constraint. In greater detail, the present work makes the following contributions:

- We introduce the irrevocable multi-armed bandit problem and develop a heuristic for its solution motivated by recent advances in the study of stochastic packing. We present a computational study which demonstrates that the performance of the packing heuristic compares favorably with a computable upper bound on the performance of any (potentially non-irreovocable) multi-armed bandit policy. We compare the performance of the packing heuristic with that of a heuristic originally proposed by Whittle (which is not irrevocable) and a natural irrevocable version of Whittle's heuristic. We find that the packing heuristic offers substantial performance gains over the irrevocable version of Whittle's heuristic we consider. Further we observe that the number of 'revocations' of an arm under Whittle's heuristic is substantial while offering only a modest improvement over the packing heuristic.

- We present a theoretical analysis to bound the performance loss incurred relative to an optimal policy with no restrictions on irrevocability. We characterize a general class of bandits for which this 'price of irrevocability' is uniformly bounded. We show that this class of bandits admits the 'learning' applications we have alluded to thus far. For bandits within this class, we show that the packing heuristic earns expected rewards that are always within a factor of $1/8$ of an optimal policy for the classical multi-armed bandit. We present stronger bounds by allowing for a dependence on problem parameters such as the number of bandits and the degree of parallelism (i.e. the ratio k/n). For instance, we show that in a natural scaling regime first proposed by Whittle, the above bound can be improved by a factor of 2. These bounds imply, to the best of our knowledge, the *first* performance bounds for an important general class of bandit problems with multiple simultaneous pulls over a finite time horizon. We introduce a mode of analysis distinct from the 'mean-field' techniques used for bandit problems with the long run average reward criterion [1]; these latter techniques do not apply to finite horizon problems.

  An additional outcome of this analysis is that we establish a precise connection between stochastic packing problems and the multi-armed bandit problem; we anticipate that this connection can serve as a useful tool for the further design and analysis of algorithms for bandit problems.

- In the interest of practical applicability, we develop a fast, essentially combinatorial implementation of the packing heuristic. Assuming that an individual arm has $O(\Sigma)$ states, and given a time horizon of T steps, an optimal solution to the multi-armed bandit problem under consideration requires $O(\Sigma^n T^n)$ computations. The main computational step in the packing heuristic calls for the one time solution of a linear program with $O(n\Sigma T)$ variables, whose solution via a generic LP solver requires $O(n^3\Sigma^3 T^3)$ computations. We develop an algorithm that solves this linear program in $O(n\Sigma^2 T \log T)$ steps by solving a sequence of dynamic programs for each bandit arm. The technique we develop here is potentially of independent interest for the solution of 'weakly coupled' optimal control problems with coupling constraints that must be met in expectation. Employing this solution technique, our heuristic

[1] we discuss why the average reward criterion is uninteresting for Bayesian learning problems in Section 1.1

requires a total of $O(n\Sigma^2 \log T)$ computations per time step amortized over the time horizon. In contrast, Whittle's heuristic (or the irrevocable version of that heuristic we consider) requires $O(n\Sigma^2 T \log T)$ computations per time step. Given the substantial amount of research that has been dedicated to simply calculating Gittins indices (in the context of Whittle's heuristic) rapidly, this is a notable contribution. More importantly, we establish that the packing heuristic is computationally attractive.

1.1. Relevant Literature

The multi-armed bandit problem has a rich history, and a number of excellent references (such as Gittins (1989)) provide a thorough treatment of the subject. Our consideration of the 'irrevocable' multi-armed bandit problem stems from a number of applications of the bandit framework alluded to earlier. Caro and Gallien (2007) have considered using the multi-armed bandit for the assortment design problem faced by fast fashion retailers. Pich and Van der Heyden (2002) emphasize the importance of not allowing for 'repeat' products in an assortment in that setting. Arlotto et al. (2009) consider the application of the multi-armed bandit model in the context of ascertaining the suitability of individuals for a given task at a call-center. The methodology suggested by the authors respects the irrevocability constraint studied here and is similar to the irrevocable version of Whittle's heuristic we examine. This constraint is quite natural to their setting as firing decisions are difficult to reverse. Finally, there is a very large and varied literature on the design of clinical trials and we make no attempt to review that here. 'Ethical' experimentation policies are an overriding theme of much of the work in this area; see Anscombe (1963) for an early treatment on the subject and Armitage et al. (2002) for a more recent overview.

There has been a great deal of work on heuristics for the multi-armed bandit problem. In the case where $k = 1$, that is, allowing for a single arm to be pulled in a given time step, Gittins and Jones (1974) developed an elegant index based policy that was shown to be optimal for the problem of maximizing discounted rewards over an infinite horizon. Their index policy is known to be suboptimal if one is allowed to pull more than a single arm in a given time step. Whittle (1988) developed a simple index based heuristic for a more general bandit problem (the 'restless' bandit problem) allowing for multiple arms to be pulled in a given time step. While his original paper was concerned with maximizing long-term average rewards, his heuristic is easily adapted to other objectives such as discounted infinite horizon rewards or expected rewards over a finite horizon (see for instance Bertsimas and Nino-Mora (2000); Caro and Gallien (2007)). It is important to note, however, that much of the extant performance analysis for bandit problems (beyond the initial work of Gittins and Jones (1974)) in general, and Whittle's heuristic in particular, focuses on bandits with a *single recurrent class* and the *average cost/reward criterion*. In that setting Weber and Weiss (1990) presented a set of technical conditions that guarantee that Whittle's heuristic is asymptotically optimal (in a regime where n and k go to infinity keeping n/k constant) for the general restless bandit problem and further, that Whittle's heuristic is *not* optimal in general. The conditions proposed by Weber and Weiss (1990) are non-trivial to verify as they require checking the global stability of a system of non-linear differential equations. In addition there is a vast amount of work that analyzes special applications of the bandit model (for instance in scheduling, or queueing problems) which we do not review here.

While the assumption of a single recurrent class and the average reward criterion permits

performance analysis (via certain mean-field approximation techniques), such a setting immediately rules out a vast number of interesting bandit problems, including most learning applications. In addition to the fact that the assumption of a single recurrent class does not hold, the average reward criterion is *too coarse* for such applications: very crudely, optimal policies for this criterion do not face the problem of carefully allocating an 'exploration budget' across arms. More precisely, any policy with 'vanishing regret' (Lai and Robbins (1985)) is optimal for the average reward criterion. A relatively recent paper by Glazebrook and Wilkinson (2000) establishes that a Whittle-like heuristic for *irreducibile* multi-armed bandits and the discounted infinite horizon criterion approaches the optimal policy at a uniform rate as the discount factor approaches unity; this is a regime where the average cost and discounted cost criteria effectively co-incide. Moreover, the requirement of irreducibility again rules out Bayesian learning applications.

There is thus little available in the way of general performance analyses for the bandit problem with multiple simultaneous plays under either the discounted infinite horizon or finite time horizon criteria. Since the packing heuristic is certainly feasible for the multi-armed bandit problem, we believe that the present work offers the *first* performance bounds for an important general class of multi-armed bandit problems with the finite time horizon criterion and multiple simultaneous arm plays.

The packing heuristic policy builds upon recent insights on the 'adaptivity' gap for stochastic packing problems. In particular, Dean et al. (2008) recently established that a simple static rule (Smith's rule) for packing a knapsack with items of fixed reward (known a-priori), but whose sizes were stochastic and unknown a-priori was within a constant factor of the optimal adaptive packing policy. Guha and Munagala (2007) used this insight to establish a similar static rule for 'budgeted learning problems'. In such a problem one is interested in finding a coin with highest bias from a set of coins of uncertain bias, assuming one is allowed to toss a *single* coin in a given time step and that one has a finite budget on the number of such experimental tosses allowed. Our work parallels that work in that we draw on the insights of the stochastic packing results of Dean et al. (2008). In addition, we must address two significant hurdles - correlations between the total reward earned from pulls of a given arm and the total number of pulls of that arm (these turn out not to matter in the budgeted learning setting, but are crucial to our setting), and secondly, the fact that multiple arms may be pulled simultaneously (only a single arm may be pulled at any time in the budgeted learning setting). Finally, a working paper (Bhattacharjee et al. (2007)), brought to our attention by the authors of that work considers a variant of the budgeted learning problem of Guha and Munagala (2007) wherein one is allowed to toss multiple coins simultaneously. While it is conceivable that their heuristic may be modified to apply to the multi-armed bandit problem we address, the heuristic they develop is also *not* irrevocable.

Restricted to learning applications, our work takes an inherently Bayesian view of the multi-armed bandit problem. It is worth mentioning that there are a number of non-parametric formulations to such problems with a vast associated literature. Most relevant to the present model are the papers by Anantharam et al. (1987a,b) that develop simple 'regret-optimal' strategies for multi-armed bandit problems with multiple simultaneous plays. One could easily imagine imposing a similar 'irrevocability' restriction in that setting and it would be interesting to design algorithms for such a problem.

The remainder of this paper is organized as follows. Section 2 presents the irrevocable multi-

armed bandit model. Section 3 develops the packing heuristic. Section 4 introduces a structural property for bandit arms we call the 'decreasing returns' property. It is shown that bandits for learning applications possess this property. That section then establishes that the price of irrevocability for bandits possessing the decreasing returns property is uniformly bounded and develops stronger performance bounds in interesting asymptotic parameter regimes. Section 5 presents very encouraging computational experiments for large scale bandit problems drawn from an interesting generative family. In the interest of implementability, Section 6 develops a combinatorial algorithm for the fast computation of packing heuristic policies for multi-armed bandits. Section 7 concludes with a perspective on interesting directions for future work.

2. The Irrevocable Multi-Armed Bandit Model

We consider a multi-armed bandit problem with multiple simultaneous 'pulls' permitted at every time step and 'irrevocability' restrictions. A single bandit arm (indexed by i) is a Markov Decision Process (MDP) specified by a state space $\mathcal{S}_i$, an action space, $\mathcal{A}_i$, a reward function $r_i : \mathcal{S}_i \times \mathcal{A}_i \to \mathbb{R}_+$, and a transition kernel $P_i : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{S}_i \to [0,1]$; $P_i(x_i, a_i, y_i)$ is thus the probability that employing action a_i on arm i while it is in state x_i will lead to a transition to state y_i. Given the state and action for an arm i at some time t, the evolution of the state for that arm over the subsequent time step is independent of the other arms.

Every bandit arm is endowed with a distinguished 'idle' action ϕ_i. Should a bandit be idled in some time period, it yields no rewards in that period and transitions to the same state with probability 1 in the next period. More precisely,

$$\begin{aligned} r_i(s_i, \phi_i) &= 0, \quad \forall s_i \in \mathcal{S}_i, \\ P_i(s_i, \phi_i, s_i) &= 1, \quad \forall s_i \in \mathcal{S}_i. \end{aligned}$$

We consider a bandit problem with n arms. The *only* action available at arms that were idled in the prior time step but pulled at some point in the past is the idle action; that is, the decision to idle an arm pulled in the previous time step is 'irrevocable'. Should an action other than the idle action be selected at an arm, we refer to such a selection as a 'pull' of that arm. That is, any action $a_i \in \mathcal{A}_i \setminus \{\phi_i\}$ would be considered a pull of the ith arm. In each time step one must select a subset of up to $k(\leq n)$ arms to pull. One is forced to pick the idle action for the remaining $n-k$ arms. We wish to find an action selection (or control) policy that maximizes expected rewards earned over T time periods. Our problem may be cast as an optimal control problem. In particular, we define as our *state-space* the set $\mathcal{S} = \prod_i \mathcal{S}_i$ and as our *action space*, the set $\mathcal{A} = \prod_i \mathcal{A}_i$. We let $\mathcal{T} = \{0, 1, \ldots, T-1\}$. We understand by s_i, the ith component of $s \in \mathcal{S}$ and similarly let a_i denote the ith component of $a \in \mathcal{A}$. We define a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$, given by $r(s,a) = \sum_i r_i(s_i, a_i)$ and a system transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, given by $P(s, a, s') = \Pi_i P_i(s_i, a_i, s'_i)$.

We now formally develop what we mean by a feasible control policy. Let X_0 be a random variable that encapsulates any endogenous randomization in selecting an action, and define the filtration generated by X_0 and the history of visited states and actions by

$$\mathcal{F}_t = \sigma(X_0, (s^0), (s^1, a^0), \ldots, (s^t, a^{t-1})),$$

where s^t and a^t denote the state and action at time t, respectively. We assume that $\mathbb{P}(s^{t+1} = s' | s^t = s, a^t = a, H_t = h_t) = P(s, a, s')$ for all $s, s' \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}$ and any $\mathcal{F}_t$-measurable random variable H_t. A *feasible* policy simply specifies a sequence of $\mathcal{A}$-valued actions $\{a^t\}$ adapted to $\mathcal{F}_t$ and satisfying:

$$a_i^t = \phi_i \text{ if } a_i^{t-1} = \phi_i \text{ and } \exists t' < t \text{ with } a_i^{t'} \neq \phi_i \text{ (Irrevocability)}$$

and

$$\sum_i \mathbf{1}_{\{a_i^t \neq \phi_i\}} \leq k. \text{ (At most } k \text{ simultaneous pulls).}$$

In particular, such a policy may be specified by a collection of $\sigma(X^0)$ measurable, $\mathcal{A}$-valued random variables, $\{\mu(s^0, \ldots, s^t, a^0, \ldots, a^{t-1}, t)\}$, one for each possible state-action history of the system. We let M denote the set of all such policies μ, and denote by $J^\mu(s, 0)$ the expected value of using policy μ starting in state s at time 0; in particular

$$J^\mu(s, 0) = E\left[\sum_{t=0}^{T-1} r(s^t, a^t) \,\middle|\, s^0 = s\right],$$

where $a^t = \mu(s^0, \ldots, s^t, a^0, \ldots, a^{t-1}, t)$.

Our goal is to compute an optimal feasible policy. In particular, we would like to find a policy μ^* that achieves

$$J^{\mu^*}(s, 0) = \sup_{\mu \in M} J^\mu(s, 0).$$

3. The Packing Heuristic

The irrevocable multi-armed bandit problem defined above does not appear to admit a tractable optimal solution. As such, this section focuses on developing a heuristic for the problem that we will subsequently demonstrate offers excellent performance and admits uniform performance guarantees.

We begin this section with an overview of our proposed heuristic: Assume we are given some set of policies, one for each individual arm, $\bar{\mu}_i : \mathcal{S}_i \to \Delta_{\mathcal{A}_i}$ (where $\Delta_{\mathcal{A}_i}$ is the $|\mathcal{A}_i|$ dimensional unit simplex). Notice that unlike the optimal policy for the (irrevocable) multi-armed bandit problem, this set is a tractable object since each policy in specified as a function of the state of a single arm. Consider applying policy $\bar{\mu}_i$ to the ith arm in isolation over T time periods. Let $\nu_i(\bar{\mu}_i)$ be the expected reward garnered from the arm and let $\psi_i(\bar{\mu}_i)$ be the expected number of times the arm was pulled over this T period horizon. Next, consider the problem of finding individual arm policies $\bar{\mu}_i$ so as to solve the following problem:

$$\begin{array}{ll} \max_{\bar{\mu}_i, i=1,2,\ldots,n} & \sum_i \nu_i(\bar{\mu}_i) \\ \text{s. t.} & \sum_i \psi_i(\bar{\mu}_i) \leq kT \end{array}$$

The above program can be expressed as a linear program with a tractable number of variables and constraints. Moreover, the value of this program provides an upper bound on the performance of an optimal policy for the classical multi-armed bandit problem. In fact, we will derive this program by considering a natural relaxation of the classical multi-armed bandit problem. Although we do not

show it, Whittle's heuristic may be viewed as a policy motivated by the dual of this program. Given a solution, $\bar{\mu}^*$, to the above program, the policy we propose – the 'packing heuristic' – operates roughly as follows:

- Sort the arms in decreasing order of the ratio $\nu_i(\bar{\mu}_i^*)/\psi_i(\bar{\mu}_i^*)$.
- Select the top k arms according to this ranking and select actions for these arms according to their respective policies $\bar{\mu}_i^*$. Should the policy for a specific arm choose not to pull that arm, discard it and replace it with the highest ranked arm from the set of arms that have not been selected yet. Once all arms are set to be pulled, let time advance.
- In the tth time step repeat the above procedure starting with the set of arms pulled in the $t-1$st time step. If an arm is discarded its place is taken by the highest ranked arm according to our initial ranking from among the arms not selected yet; discarded arms can thus never be re-introduced.

In the remainder of this section, we rigorously develop the heuristic described above. We begin by considering the classical multi-armed bandit problem which may be viewed as a relaxation of the irrevocable problem and describe a (standard) linear program for its solution.

3.1. Computing an Optimal Policy without restrictions on Irrevocability

It is useful to consider the classical multi-armed bandit problem (without the irrevocability constraint) in designing policies for the irrevocable multi-armed bandit problem. In finding an optimal policy for the classical multi-armed bandit problem, it suffices to restrict attention to Markovian policies. A Markovian policy in this case is specified as a collection of independent $\mathcal{A}$ valued random variables $\{\mu(s,t)\}$ each measurable with respect to $\sigma(X_0)$, satisfying $\sum_i \mathbf{1}_{\{\mu(s,t)_i \neq \phi_i\}} \leq k$, for all s,t. In particular, assuming the system is in state s at time t, such a policy selects an action a^t as the random variable $\mu(s,t)$, independent of past states and actions.

We denote an optimal Markovian policy for the classical bandit problem by μ^*_{UB} and let $J^*_{\text{UB}}(s,0)$ denote the value garnered under this policy starting in state s at time t. Now, $J^*_{\text{UB}}(s,0) \geq J^*(s,0)$ since a feasible policy for the irrevocable bandit problem is clearly feasible for the classical bandit problem. The policy μ^*_{UB} may be found via the solution of the following linear program, $LP(\tilde{\pi}_0)$, specified by a parameter $\tilde{\pi}_0 \in \Delta_{\mathcal{S}}$ that determines the distribution of arm states at time $t = 0$. Here, $\mathcal{A}^{\text{feas}} = \{a \in \mathcal{A} : \sum_i \mathbf{1}_{\{a_i \neq \phi_i\}} \leq k\}$.

$$
\begin{array}{lll}
\max. & \sum_t \sum_{s,a} \pi(s,a,t) r(s,a), & \\
\text{s. t.} & \sum_a \pi(s,a,t) = \sum_{s',a'} P(s',a',s)\pi(s',a',t-1), & \forall t > 0, s \in \mathcal{S}, \\
& \pi(s,a,t) = 0, & \forall s,t,a \notin \mathcal{A}^{\text{feas}} \\
& \sum_a \pi(s,a,0) = \tilde{\pi}_0(s), & \forall s \in \mathcal{S}, \\
& \pi \geq 0. &
\end{array}
$$

where the variables are the state-action frequencies $\pi(s,a,t)$, which give the probability of being in state s at time t and choosing action a. The first set of constraints in the above program simply enforce the dynamics of the system, while the second set of constraints enforces the requirement that at most k arms are simultaneously pulled at any point in time.

An optimal solution to the program above may be used to construct a policy μ^*_{UB} that attains expected value $J^*_{UB}(s,0)$ starting at any state s for which $\tilde{\pi}_0(s) > 0$. In particular, given an optimal solution π^{opt} to $LP(\tilde{\pi}_0)$, one obtains such a policy by defining $\mu^*_{\text{UB}}(s,t)$ as a random variable that takes value $a \in \mathcal{A}$ with probability $\pi^{\text{opt}}(s,a,t)/\sum_a \pi^{\text{opt}}(s,a,t)$. By construction, we have $E[J^*_{\text{UB}}(s,0)|s \sim \tilde{\pi}_0] = OPT(LP(\tilde{\pi}_0))$. Efficient solution of the above program is not a tractable task and we next consider making a further relaxation: as opposed to allowing up to k pulls in a given time step, we require that this constraint only be met in expectation.

3.2. A Further Relaxation

Consider the following relaxation of the program $LP(\tilde{\pi}_0)$, $RLP(\tilde{\pi}_0)$:

$$\begin{array}{ll}
\max. & \sum_i \sum_t \sum_{s_i,a_i} \pi_i(s_i,a_i,t) r_i(s_i,a_i), \\
\text{s. t.} & \sum_{a_i} \pi_i(s_i,a_i,t) = \sum_{s'_i,a'_i} P_i(s'_i,a'_i,s_i)\pi_i(s'_i,a'_i,t-1), \quad \forall t>0, s_i \in \mathcal{S}_i, i, \\
& \sum_i \left[T - \sum_{s_i} \sum_t \pi_i(s_i,\phi_i,t)\right] \leq kT, \\
& \sum_{a_i} \pi_i(s_i,a_i,0) = \sum_{\bar{s}:\bar{s}_i=s_i} \tilde{\pi}_0(\bar{s}), \\
& \pi \geq 0,
\end{array}$$

where $\pi_i(s_i,a_i,t)$ is the probability of the ith bandit being in state s_i at time t and choosing action a_i.

The program above relaxes the requirement that up to k arms be pulled in a given time step; instead we now require that over the entire horizon at most kT arms are pulled in *expectation*, where the expectation is over policy randomization and state evolution. The first set of equality constraints enforce individual arm dynamics whereas the first inequality constraint enforces the requirement that at most kT arms be pulled in expectation over the entire time horizon. The following lemma makes the notion of a relaxation to $LP(\tilde{\pi}_0)$ precise; the proof may be found in the appendix.

Lemma 1. $OPT(RLP(\tilde{\pi}_0)) \geq OPT(LP(\tilde{\pi}_0))$

Given an optimal solution $\bar{\pi}$ to $RLP(\tilde{\pi}_0)$, one may consider the policy μ^R, that, assuming we are in state s at time t, selects a random action $\mu^R(s,t)$, where $\mu^R(s,t) = a$ with probability $\prod_i \left(\bar{\pi}_i(s_i,a_i,t)/\sum_{a_i} \bar{\pi}_i(s_i,a_i,t)\right)$ independent of the past. Noting that the action for each arm i is chosen independently of all other arms, we use $\mu^R_i(s_i,t)$ to denote the induced policy for arm i. Assume for convenience that $\tilde{\pi}_0$ is degenerate and puts mass 1 on a single starting state, say s, we have by construction, $J^{\mu^R}(s,0) = OPT(RLP(\tilde{\pi}_0))$. Moreover, we have that μ^R satisfies the constraint

$$E\left[\sum_{t=0}^{T-1} \sum_i \mathbf{1}_{\{\mu^R_i(s^t_i,t) \neq \phi_i\}} \Big| s^0 = s\right] \leq kT,$$

where the expectation is over random state transitions and endogenous policy randomization. Note that μ^R is not necessarily feasible; we ultimately require a policy that entails at most k arm pulls in any time step and is irrevocable. We will next use μ^R to construct such a feasible policy.

3.3. The Packing Heuristic

In what follows we will assume for convenience that $\tilde{\pi}_0$ is degenerate and puts mass 1 on a single starting state. That is, $\tilde{\pi}_0(s_i) = 1$ for some $s_i \in \mathcal{S}_i$ for all i. We first introduce some relevant notation. Given an optimal solution $\bar{\pi}$ to $RLP(\tilde{\pi}_0)$, define the value generated by arm i as the random variable

$$R_i = \sum_{t=0}^{T-1} r_i(s_i^t, \mu_i^R(s_i^t, t)),$$

and the 'active time' of arm i, T_i as the total number of pulls of arm i entailed under that policy

$$T_i = \sum_{t=0}^{T-1} \mathbf{1}_{\{\mu_i^R(s_i^t,t) \neq \phi_i\}}.$$

The expected value of arm i, $E[R_i] = \sum_{s_i,a_i,t} \bar{\pi}_i(s_i, a_i, t) r_i(s_i, a_i)$, and the expected active time $E[T_i] = \sum_{s_i,a_i,t:a_i \neq \phi_i} \bar{\pi}_i(s_i, a_i, t)$. We will assume in what follows that $E[T_i] > 0$ for all i; otherwise, we simply consider eliminating those i for which $E[T_i] = 0$. We will also assume for analytical convenience that $\sum_i E[T_i] = kT$. Neither assumption results in a loss of generality.

To motivate our policy we begin with the following analogy with a packing problem: Imagine packing n objects into a knapsack of size B. Each object i has size ψ_i and value ν_i. Moreover, we assume that we are allowed to pack fractional quantities of an object into the knapsack and that packing a fraction α of the ith object requires space $\alpha\psi_i$ and generates value $\alpha\nu_i$. An optimal policy is then given by the following greedy procedure: select objects in decreasing order of the ratio ν_i/ψ_i and place them in to the knapsack to the extent that there is room available. If one had more than a single knapsack and the additional constraint that an item could not be placed in more than a single knapsack, then the situation is more complicated. One may consider a greedy procedure that, as before, considers items in decreasing order of the ratio ν_i/ψ_i and places them (possibly fractionally) in sequence, into the least loaded of the bins at that point. This generalization of the greedy procedure for the simple knapsack is suboptimal, but still a reasonable heuristic.

Thus motivated, we begin with a loose high level description of our control policy, which we call the 'packing' heuristic. We think of each bandit arm i as an 'item' of value $E[R_i]$ with size $E[T_i]$, where $E[R_i]$ and $E[T_i]$ are obtained through the solution of $RLP(\tilde{\pi}_0)$ as described above. For the purposes of this explanation *alone*, we will assume for convenience that should policy μ^R call for an arm that was pulled in the past to be idled, it will never again call for that arm to be pulled; we will momentarily remove that assumption. Our control policy will operate as follows: we will order arms in decreasing order of the ratio $E[R_i]/E[T_i]$. We begin with the top k arms according to this ordering. For each such arm we will select an action according to the policy specified for that arm by μ_i^R; should this policy call for the arm to be idled, we discard that arm and will never again consider pulling it. We replace the discarded arm with the next available arm (in order of initial arm rankings) and select an action for the arm according to μ^R. We repeat this procedure until we have selected non-idle actions for up to k arms (or no arms are available). We then let time advance, earn rewards, and repeat the procedure described above until the end of the time horizon.

Algorithm 1 describes the packing heuristic policy precisely, addressing the fact that μ_i^R may call for an arm to be idled but then pulled in some subsequent time step.

Algorithm 1 The Packing Heuristic

Renumber bandits so that $\frac{E[R_1]}{E[T_1]} \geq \frac{E[R_2]}{E[T_2]} \cdots \geq \frac{E[R_n]}{E[T_n]}$. Index bandits by variable i.
$l_i \leftarrow 0, a_i \leftarrow \phi_i$ for all i, $s \sim \tilde{\pi}_0(\cdot)$
{The 'local time' of every arm is set to 0 and its designated action to the idle action. An initial state is drawn according to the initial state distribution $\tilde{\pi}_0$.}
$J \leftarrow 0$ {Total reward earned is initialized to 0.}
$\mathbb{X} \leftarrow \{1, 2, \ldots, k\}, \mathbb{A} \leftarrow \{k+1, \ldots, n\}, \mathbb{D} = \emptyset$.
{Initialize the set of active ($\mathbb{X}$), available ($\mathbb{A}$), and discarded ($\mathbb{D}$) arms.}
for $t = 0$ to $T - 1$ **do**
  while there exists an arm $i \in \mathbb{X}$ with $a_i = \phi_i$ **do** {Select up to k arms to pull.}
    Select an $i \in \mathbb{X}$ with $a_i = \phi_i$
    {In what follows, either select an action for arm i or else discard it.}
    while $a_i = \phi_i$ and $l_i < T$ **do** {Attempt to select a pull action for arm i}
      Select $a_i \propto \bar{\pi}_i(s_i, \cdot, l_i)$ {Select an action according to the solution to $RLP(\bar{\pi})$.}
      $l_i \leftarrow l_i + 1$ {Increment arm i's local time.}
    end while
    if $l_i = T$ and $a_i = \phi_i$ **then** {Discard arm i and activate next highest ranked arm available.}
      $\mathbb{X} \leftarrow \mathbb{X} \setminus \{i\}, \mathbb{D} \leftarrow \mathbb{D} \cup \{i\}$ {Discard arm i.}
      if $\mathbb{A} \neq \emptyset$ **then** {There are available arms.}
        $j \leftarrow \min \mathbb{A}$ {Select highest ranked available arm.}
        $\mathbb{X} \leftarrow \mathbb{X} \cup \{j\}, \mathbb{A} \leftarrow \mathbb{A} \setminus \{j\}$ {Add arm to active set.}
      end if
    end if
  end while
  for Every $i \in \mathbb{X}$ **do** {Pull selected arms.}
    $s_i \sim P(s_i, a_i, \cdot)$
    {Pull arm i; select next arm i state according to its transition kernel assuming the use of action a_i.}
    $J \leftarrow J + r_i(s_i, a_i)$ {Earn rewards.}
    $a_i \leftarrow \phi_i$
  end for
end for

In the event that we placed no restriction on the time horizon (i.e. we ignored the upper limit on t in line 5 of the algorithm), we have by construction, that the expected total reward earned under the above policy is precisely $OPT(RLP(\tilde{\pi}_0))$; subsequent analysis will, in a sense, quantify the loss due to the fact that we do not count rewards earned by the algorithm beyond $t = T-1$. In essence, $RLP(\tilde{\pi}_0)$ prescribes a policy wherein each arm generates a total reward with mean $E[R_i]$ using an expected total number of pulls $E[T_i]$, independent of other arms. Our algorithm may be visualized as one which 'packs' as many of the pulls of various arms possible in a manner so as to meet feasibility constraints.

In the Sec. 5, we present a comprehensive computational study of the packing heuristic we have proposed. The study establishes that the packing heuristic offers performance levels within about 10% of an upper bound on the performance of an optimal policy for the *classical* multi-armed bandit problem. The computational study also establishes that Whittle's heuristic for the corresponding classical multi-armed bandit problems entails a large number of 'revocations' while yielding only a marginal performance improvement over the packing heuristic. Finally, we also consider a natural 'irrevocable' modification of Whittle's heuristic which we show performs poorly relative to the packing heuristic. Before we launch into these computational experiments however, we present a theoretical analysis of the performance loss incurred in using the packing heuristic.

4. The Price of Irrevocability

This section establishes upper bounds on the performance loss incurred in using the irrevocable packing heuristic relative to an upper bound on the performance of an optimal policy for the classical multi-armed bandit problem. We restrict attention to a class of bandits whose arms satisfy a certain 'decreasing returns' property; as we will subsequently discuss, this class subsumes an important canonical family of bandit problems related to learning applications. We establish that the packing heuristic always earns expected rewards that are within a factor of 1/8 of an optimal scheme for such problems. We sharpen our analysis for problems in an asymptotic regime first proposed by Whittle, where the number of bandits n is increased while keeping the ratio k/n constant. In that regime, we present a performance guarantee that depend on the 'degree of parallelism' in the problem, i.e. the ratio k/n and also a substantially improved uniform guarantee.

Our analysis provides the first performance bounds for a general class of bandits with multiple simultaneous plays and the finite horizon criterion. Prior analyses have typically focused on irreducible bandits and the infinite horizon criterion (which rule out applications to learning problems, for instance); the mean field analyses used there do not apply here. Our analysis sheds light on the structural properties and operating regimes for which the packing heuristic is likely to offer a viable solution to the irrevocable multi-armed bandit problem. Our methods make a precise connection between stochastic packing problems and multi-armed bandit problems, and in doing so open up new avenues for the design and analysis of multi-armed bandit algorithms.

In what follows we first specify the decreasing returns property and explicitly identify a class of bandits that possess this property. We then present our performance analysis which will proceed as follows: we first consider pulling bandit arms *serially*, i.e. at most one arm at a time, in order of their rank and show that the total reward earned from bandits that were first pulled within the first $kT/2$ pulls is at least within a factor of 1/8 of an optimal policy; this factor can be improved

to 1/4 in a certain asymptotic regime. Our uniform bound relies on the static ranking of bandit arms used, and a symmetrization idea exploited by Dean et al. (2008) in their result on stochastic packing where rewards are statistically independent of item size. In contrast to that work, we must address the fact that the rewards earned from a bandit are statistically dependent on the number of pulls of that bandit and to this end we exploit the decreasing returns property that establishes the nature of this correlation. We then show via a combinatorial sample path argument that the expected reward earned from bandits pulled within the first $T/2$ time steps of the packing heuristic i.e., with arms being pulled in parallel, is at least as much as that earned in the setting above where arms are pulled serially, thereby establishing our first performance guarantee. Our analysis in Whittle's regime uses a similar program but sharper estimates of a number of quantities of interest.

4.1. The Decreasing Returns Property

Define for every i and $l < T$, the random variable

$$L_i(l) = \sum_{t=0}^{l} \mathbf{1}_{\{\mu_i^R(s_i^t,t)\neq\phi_i\}}.$$

$L_i(l)$ tracks the number of times a given arm i has been pulled under policy μ^R among the first $l+1$ steps of selecting an action for that arm. Further, define

$$R_i^m = \sum_{l=0}^{T-1} \mathbf{1}_{\{L_i(l)\leq m\}} r_i(s_i^l, \mu_i^R(s_i^l, l)).$$

R_i^m is the random reward earned within the first m pulls of arm i under the policy μ^R. The decreasing returns property roughly states that the expected incremental returns from allowing an additional pull of a bandit arm are, on average, decreasing. More precisely, we have:

Property 1. *(Decreasing Returns)* $E[R_i^{m+1}] - E[R_i^m] \leq E[R_i^m] - E[R_i^{m-1}]$ *for all* $0 < m < T$.

One useful class of bandits from a modeling perspective that satisfy this property are bandits whose arms yield i.i.d rewards of an a-priori unknown, arm-specific mean. We refer to these as 'learning problems'. The following discussion makes this notion more precise:

4.1.1. Learning problems and the decreasing returns property

We consider the following generic class of 'learning' problems: We have n bandit arms. A pull of the ith arm yields an independent, random, non-negative reward [2] X_i having density (or p.m.f.) $f_{\theta_i}(\cdot)$ where θ_i is an unknown parameter in some set Θ_i. We assume that θ_i is drawn randomly at time 0 according to the density (or p.m.f) g_i, and is independent of all θ_j with $j \neq i$. Our objective is to arrive at an arm selection policy that adaptively selects a subset of k arms to pull at each point in time with a view to maximizing total expected reward earned over T periods. In the interest of tractability, we assume that g_i belongs to some parametric class of functions $\mathcal{G}_i$, a member of which is specified by parameter $s_i \in \mathcal{S}_i$; we make this dependence precise with the notation $g_i^{s_i}$.

[2] we may also assume that the reward earned is $h(X_i)$ where h is a known, non-negative, concave function.

Moreover, we assume that $g_i^{s_i}$ is a conjugate prior for f_{θ_i} for all $s_i \in \mathcal{S}_i$. That is, our posterior on θ_i given an observation X_i remains in $\mathcal{G}_i$.

Learning problems of this type are rather common and fit a number of modeling needs, including for instance, the fast fashion and call-center staffing examples described in the introduction (see Caro and Gallien (2007), Arlotto et al. (2009), and also Section 5 for concrete examples within this framework). In addition, these problems are in a sense the canonical application of the bandit model (see Bellman (1956), Gittins and Wang (1992)). For further applications see the books Bergman and Gittins (1985); Berry and Fristedt (1985).

It is not hard to see that the learning problem we have posed can be cast as a multi-armed bandit problem in the sense of the model in Section 3. In particular, the state space for each arm is simply $\mathcal{S}_i$, with action space $\mathcal{A}_i = \{p_i, \phi_i\}$ consisting of two actions – pull and idle. The transition kernel P_i is specified implicitly by Bayes' rule and the reward function is defined according to:

$$r_i(s_i, p_i) = \int_{x,\theta} x f_{\theta_i}(x) g_i^{s_i}(\theta_i) d\theta_i dx$$

By Bayes' rule, rewards from a given arm (as defined above) will then satisfy the following intuitive property reflecting the consistency of our estimate of the mean reward from a bandit arm:

$$r_i(s_i, p_i) = \sum_{s_i' \in \mathcal{S}_i} P_i(s_i, p, s_i') r_i(s_i', p_i), \quad \forall s_i \in \mathcal{S}_i.$$

In light of the following Lemma, this broad class of learning problems satisfy the decreasing returns property. In particular, we have the following result whose proof may be found in the appendix:

Lemma 2. *Given a multi-armed bandit problem with $\mathcal{A}_i = \{p_i, \phi_i\}$ $\forall i$, and*

$$r_i(s_i, p_i) \geq \sum_{s_i' \in \mathcal{S}_i} P_i(s_i, p_i, s_i') r_i(s_i', p_i), \quad \forall i, s_i \in \mathcal{S}_i,$$

we must have

$$E[R_i^{m+1}] - E[R_i^m] \leq E[R_i^m] - E[R_i^{m-1}]$$

for all $0 < m < T$.

4.2. A Uniform Bound on the Price of Irrevocability

For convenience of exposition we assume that T is even; addressing the odd case requires essentially identical proofs but cumbersome notation.

We re-order the bandits in decreasing order of $E[R_i]/E[T_i]$ as in the packing heuristic. Let us define

$$H^* = \min \left\{ j : \sum_{i=1}^{j} E[T_i] \geq kT/2 \right\}.$$

Thus, H^* is the set of bandits that take up approximately half the budget on total expected pulls. Next, let us define for all $i \leq H^*$, random variables $\tilde{R}_i$ and $\tilde{T}_i$ according to $\tilde{R}_i = R_i, \tilde{T}_i = T_i$ for all $i < H^*$ and $\tilde{R}_{H^*} = \alpha R_{H^*}$ and $\tilde{T}_{H^*} = \alpha T_{H^*}$, where $\alpha = \frac{kT/2 - \sum_{i=1}^{H^*-1} E[T_i]}{E[T_{H^*}]}$.

We begin with a preliminary lemma whose proof may be found in the appendix:

Lemma 3.

$$\sum_{i=1}^{H^*} E[\tilde{R}_i] \geq \frac{1}{2} OPT(RLP(\tilde{\pi}_0)).$$

We next compare the expected reward earned by a certain subset of bandits with indices no larger than H^*. The significance of the subset of bandits we define will be seen later in the proof of Lemma 6 – we will see there that all bandits in this subset will begin operation prior to time $T/2$ in a run of the packing heuristic. In particular, define

$$R_{1/2} = \sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i.$$

Lemma 4.

$$E[R_{1/2}] \geq \frac{1}{4} OPT(RLP(\tilde{\pi}_0)).$$

Proof. We have:

$$\begin{aligned}
E[R_{1/2}] &\overset{(a)}{=} \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} T_j < kT/2\right) E[R_i] \\
&\overset{(b)}{\geq} \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} T_j < kT/2\right) E[\tilde{R}_i] \\
&\overset{(c)}{=} \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} \tilde{T}_j < kT/2\right) E[\tilde{R}_i] \\
&\overset{(d)}{\geq} \sum_{i=1}^{H^*} \left(1 - \frac{\sum_{j=1}^{i-1} E[\tilde{T}_j]}{kT/2}\right) E[\tilde{R}_i] \\
&= \sum_{i=1}^{H^*} E[\tilde{R}_i] - \sum_{i=1}^{H^*} \frac{\sum_{j=1}^{i-1} E[\tilde{T}_j]}{kT/2} E[\tilde{R}_i] \\
&\overset{(e)}{\geq} \sum_{i=1}^{H^*} E[\tilde{R}_i] - \frac{1}{2} \sum_{i=1}^{H^*} \frac{\sum_{j=1, j\neq i}^{H^*} E[\tilde{T}_j]}{kT/2} E[\tilde{R}_i] \\
&\overset{(f)}{\geq} \frac{1}{2} \sum_{i=1}^{H^*} E[\tilde{R}_i] \\
&\overset{(g)}{\geq} \frac{1}{4} OPT(RLP(\tilde{\pi}_0))
\end{aligned}$$

Equality (a) follows from the fact that under policy μ^R, R_i is independent of T_j for $j < i$. Inequality (b) follows from our definition of $\tilde{R}_i$: $\tilde{R}_i \leq R_i$. Equality (c) follows from the fact that by definition $\tilde{T}_i = T_i$ for all $i < H^*$. Inequality (d) invokes Markov's inequality.

Inequality (e) is the critical step in establishing the result and uses the simple symmetrization idea exploited by Dean et al. (2008): In particular, we observe that since $\frac{E[R_i]}{E[T_i]} \leq \frac{E[R_j]}{E[T_j]}$ for $i > j$, it follows that $E[R_i]E[T_j] \leq \frac{1}{2}(E[R_i]E[T_j] + E[R_j]E[T_i])$ for $i > j$. Replacing every term of the form $E[R_i]E[T_j]$ (with $i > j$) in the expression preceding inequality (e) with the upper

bound $\frac{1}{2}(E[R_i]E[T_j] + E[R_j]E[T_i])$ yields inequality (e). Inequality (f) follows from the fact that $\sum_{i=1}^{H^*} E[\tilde{T}_i] = kT/2$ and since $E[R_i] \geq 0$. Inequality (g) follows from Lemma 3. ■

Before moving on to our main Lemma that translates the above guarantees to a guarantee on the performance of the packing heuristic, we need to establish one additional technical fact. Recall that R_i^m is the reward earned by bandit i in the first m pulls of this bandit under policy μ^R. Also, note that $R_i^T = R_i$. Exploiting the assumed decreasing returns property, we have the following Lemma whose proof may be found in the appendix:

Lemma 5. *For bandits satisfying the decreasing returns property (Property 1),*

$$E\left[\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2}\right] \geq \frac{1}{2} E[R_{1/2}].$$

We have thus far established estimates for total expected rewards earned assuming implicitly that bandits are pulled in a serial fashion in order of their rank. The following Lemma connects these estimates to the expected reward earned under the μ^{packing} policy (given by the packing heuristic) using a simple sample path argument. In particular, the following Lemma shows that the expected rewards under the μ^{packing} policy are at least as large as $E\left[\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2}\right]$.

Lemma 6. *Assuming $\tilde{\pi}_0(s) = 1$, we have*

$$J^{\mu^{\text{packing}}}(s, 0) \geq E\left[\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2}\right].$$

Proof. For a given sample path of the system define

$$h = (H^*) \wedge \min\left\{ i : \sum_{j=1}^{i} T_j \geq kT/2 \right\}.$$

On this sample path, it must be that:

$$\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2} = \sum_{i=1}^{h} R_i^{T/2}. \tag{1}$$

We claim that arms $1, 2, \ldots, h$ are all first pulled at times $t < T/2$ under μ^{packing}. Assume to the contrary that this were not the case and recall that arms are considered in order of index under μ^{packing}, so that an arm with index i is pulled for the first time no later than the first time arm l is pulled for $l > i$. Let h' be the highest arm index among the arms pulled at time $t = T/2 - 1$ so that $h' < h$. It must be that $\sum_{j=1}^{h'} T_j \geq kT/2$. But then,

$$H^* \wedge \min\left\{ i : \sum_{j=1}^{i} T_j \geq kT/2 \right\} \leq h'$$

which is a contradiction.

Thus, since every one of the arms $1, 2, \ldots, h$ is first pulled at times $t < T/2$, each such arm may be pulled for at least $T/2$ time steps prior to time T (the horizon). Consequently, we have that the total rewards earned on this sample path under policy μ^{packing} are at least

$$\sum_{i=1}^{h} R_i^{T/2}$$

Using identity (1) and taking an expectation over sample paths yields the result. ■

We are ready to establish our main Theorem that provides a uniform bound on the performance loss incurred in using the packing heuristic policy relative to an optimal policy with no restrictions on exploration. In particular, we have that the price of irrevocability is uniformly bounded for bandits satisfying the decreasing returns property.

Theorem 1. *For multi-armed bandits satisfying the decreasing returns property (Property 1), we have*

$$J^{\mu^{\text{packing}}}(s, 0) \geq \frac{1}{8} J^*(s, 0)$$

Proof. We have from Lemmas 4,5 and 6 that

$$J^{\mu^{\text{packing}}}(s, 0) \geq \frac{1}{8} OPT(RLP(\tilde{\pi}_0))$$

where $\tilde{\pi}_0(s) = 1$. We know from Lemma 1 that $OPT(RLP(\tilde{\pi}_0)) \geq OPT(LP(\tilde{\pi}_0)) = J^*(s, 0)$ from which the result follows. ■

4.3. The Price of Irrevocability in Whittle's Asymptotic Regime

This section considers an asymptotic parameter regime where one may establish a stronger bound than that in Theorem 1. In particular, the regime we will consider is a natural candidate for what one might consider a 'large-scale' problem. We are given an 'unscaled' problem with n_0 arms in which we are allowed up to k_0 simultaneous plays over a time horizon of T. We will assume that each of these n_0 arms have identical specifications and start in identical states; this is not an essential assumption but doing away with it does not permit a clean exposition. We next consider a sequence of problems indexed by N, where the Nth problem has N copies of each of the n_0 arms in the unscaled problem, and we allow Nk_0 simultaneous plays over T time periods. Our goal is to understand the price of irrevocability as N gets large. Notice that this regime is still relevant for learning problems since we are not scaling the time horizon, T, and are not restricting the kernels P_i in any way. This regime is analogous to one considered by Whittle (1988) and Weber and Weiss (1990), albeit for irreducible bandit problems and the *average* reward criterion (which rules out learning applications, for instance). The finite horizon criterion and the fact that the bandits we consider may be (and, for learning applications, will be) non-irreducible rule out the mean-field analysis techniques of Weber and Weiss (1990).

Letting $R_{i,N}$ and $T_{i,N}$ denote the value generated by arm i and its active time (as defined in the previous section) for the Nth problem, the following facts are apparent by our assumption that each bandit is identical:

1. $R_{i,N} \stackrel{d}{=} R_{i,N'}$, and $T_{i,N} \stackrel{d}{=} T_{i,N'}$ for all N, N' and $i \leq \min(n_0 N, n_0 N')$.

2. For every N, the collection of random variables $R_{1,N}, R_{2,N}, \ldots, R_{n_0 N,N}$ are i.i.d. as are the random variables $T_{1,N}, T_{2,N}, \ldots, T_{n_0 N,N}$.

3. $E[T_{i,N}] = k_0 T/n_0$ for all N and $i \leq n_0 N$.

In light of the above facts we will eliminate the subscript N from $R_{i,N}$ and $T_{i,N}$. We note then that for the Nth problem, $OPT(RLP(\tilde{\pi}_0)) = \sum_{i=1}^{Nn_0} E[R_i]$.

We prove two main bounds in this section:

1. We first present a performance guarantee that illustrates a dependence on the ratio k_0/n_0. This ratio may be interpreted as the 'degree of parallelism' inherent to the multi-armed bandit problem at hand.

2. We then prove a performance guarantee that holds in an asymptotic regime where N gets large (but is otherwise uniform over problem parameters). This bound improves the bound in Theorem 1 by a factor of 2.

4.3.1. Impact of the 'Degree of Parallelism' (k_0/n_0)

Let us define the random variable $\tau_N = \min\{j : \sum_{i=1}^{j} T_i \geq k_0 NT\}$. We then have the following Lemma.

Lemma 7. *In the Nth system, all arms with indices smaller than or equal to $\tau_N \wedge n_0 N$ begin operation prior to time T. Moreover, for any $\epsilon > 0$ and almost all $\omega \in \Omega$, $\exists N^{\epsilon}(\omega)$, such that*

$$\tau_N \geq Nn_0 - (1+\epsilon)\sqrt{Nn_0 \log\log Nn_0}$$

for all $N \geq N^{\epsilon}(\omega)$.

The above Lemma is proved in the appendix. The Lemma is remarkable in that it states that for large scale problems (i.e. large N), almost all bandits begin operation prior to the end of the time horizon. We next translate this fact to a bound on performance. To this end, for every N, let $\sigma^N(\cdot)$ denote a (random) permutation of $\{1, \ldots, Nn_0\}$ satisfying $R_i > R_j \implies \sigma(i) < \sigma(j)$. Further, for every $l \leq Nn_0$, define the random variable

$$M_N(l) = \sum_{i:\sigma(i)\leq l} R_i$$

$M_N(l)$ is thus the realized reward of the top l arms assuming the packing heuristic were not terminated at the end of the time horizon. Define for $\alpha \in [0,1]$,

$$\gamma_N(\alpha) = E[M_N(\lceil Nn_0\alpha \rceil)]/E\left[\sum_{i=1}^{Nn_0} R_i\right].$$

Since the R_i are i.i.d random variables, $\lim_N \gamma_N(\alpha) \triangleq \gamma(\alpha)$ is well defined by the law of large numbers and is naturally interpreted as the ratio between the expected contribution of an arm

restricted to realizations that are in the top α fractile and the expected contribution of an arm. We then have the following result theorem indicates the impact of the 'degree of parallelism' in the problem on the performance of the heuristic:

Theorem 2.

$$\lim_{N\to\infty} \frac{J_N^{\mu^{\text{packing}}}(s,0)}{J_N^*(s,0)} \geq 1 - \gamma(\min(k_0/n_0, 1 - k_0/n_0))$$

The above bound (which is established in the appendix) provides an indication of the role played by the ratio k_0/n_0. Loosely, it may be interpreted as stating that the performance loss incurred by the packing heuristic is no more than the relative contribution from the top k_0/n_0 percent of arms. While one may characterize the γ function given the distribution of rewards from a given arm R_i, a fair criticism of this bound is that it is difficult to characterize γ given only primitive problem data. The next bound we provide will be uniform over problems parameters and valid for problems in Whittle's asymptotic regime.

4.3.2. A Uniform Guarantee for Whittle's Regime

We present here a performance guarantee for the asymptotic regime under consideration that depends only on problem primitives (and is, in fact, uniform over this regime). The program we will follow is essentially identical to that we followed for our proof of a uniform bound with the exception of Lemma 4; we prove an alternative result below allowing for a dependence on problem scale N. The proof may be found in the appendix.

Lemma 8. *For the Nth bandit problem, we have:*

$$E[R_{1/2}] \geq \frac{1}{2}(1 - \kappa(N))OPT(RLP(\tilde{\pi}_0)).$$

where $\kappa(N) = O(N^{-1/2+d})$ and $d > 0$ is arbitrary.

Using the result of the previous Lemma (in place of Lemma 4), and Lemmas 5 and 6 we have that

Theorem 3. *For the Nth multi-armed bandit problem,*

$$J^{\mu^{\text{packing}}}(s,0) \geq \frac{1}{4}(1 - \kappa(N))J^*(s,0)$$

where $\kappa(N) = O(N^{-1/2+d})$ for arbitrary $d > 0$ and we assume $s_i = s_j \ \forall i, j$.

This bound while still lose, provides a substantial improvement over the uniform bound in the previous section. Together, Theorems 2 and 3 indicate that the packing heuristic is likely to perform well in Whittle's asymptotic regime.

5. Computational Experiments

This section presents a computational investigation of the performance of the packing heuristic for the irrevocable multi-armed bandit problem with a view to gauge its practical efficacy. We also

examine as an alternative heuristic for the irrevocable bandit problem, a natural modification to Whittle's heuristic. Finally, we examine the performance of Whittle's (non-irrevocable) heuristic itself, paying special attention to the number of arm 'revocations' under that heuristic. In addition, we benchmark the performance of all of these schemes against a computable upper bound on the expected reward for any policy (with no restrictions on revocability); specifically, the bound is given by the objective function of problem $LP(\tilde{\pi}_0)$ in Sec. 3.2 for an optimal solution[3]. We consider a number of large scale bandit problems drawn from a generative family of problems to be discussed shortly, and demonstrate the following:

- The packing heuristic consistently demonstrates performance within about 10 to 20 % of an upper bound on the performance of an optimal policy for the classical multi-armed bandit problem. This upper bound is also an upper bound to the performance of any irrevocable scheme.

- The number of 'revocations' under Whittle's heuristic can be large in a variety of operating regimes. A natural modification to Whittle's heuristic making it feasible for the irrevocable bandit problem typically performs 15 to 20 percent worse than Whittle's heuristic in these regimes. The packing heuristic *can recover a substantial portion of the above gap* (between 50 and 100 %) in most cases.

The Generative Model: We consider multi-armed bandit problems with n arms up to k of which may be pulled simultaneously at any time. The ith arm corresponds to a Binomial(m, P_i) random variable where m is fixed and known, and P_i is unknown but drawn from a Beta(α_i, β_i) prior distribution. Assuming we choose to 'pull' arm i at some point, we realize a random outcome $M_i \in \{0, 1, \ldots, m\}$. M_i is a Binomial(m, P_i) random variable where P_i is itself a Beta(α_i, β_i) random variable. We receive a reward of $r_i M_i$ and update the prior distribution parameters according to $\alpha_i \leftarrow \alpha_i + M_i$, $\beta_i \leftarrow \beta_i + m - M_i$. By selecting the initial values of α_i and β_i for each arm appropriately we can control for the initial level of uncertainty in the value of P_i; by 'level of uncertainty' we mean the co-efficient of variation of P_i which is defined according to $\sigma(P_i)/E[P_i]$. This model is applicable to the dynamic assortment selection problem studied in Caro and Gallien (2007) with each arm representing a product of uncertain popularity and M_i representing the uncertain number of product i sales over a single period in which that product is offered for sale; the only difference with that work is that as opposed to assuming Binomial demand, the authors there assume Poisson demand.

5.1. IID Bandits

We consider bandits with $(n, k) \in \{(500, 75), (500, 125), (100, 15), (100, 25)\}$. These dimensions are representative of large scale applications such as the dynamic assortment problem (see Caro and Gallien (2007)). For each value of (n, k) we consider time horizons $T = 40, 25$ and 10 (again, horizon lengths of 40 and 25 reflect the dynamic assortment applications, assuming weekly restocking decisions). We consider three different values for the coefficient of variation in arm bias: $cv = \{1, 2.5, 4\}$. These coefficients of variation represent respectively, a low, moderate and high degree

[3]The problem $LP(\tilde{\pi}_0)$ for both computation of the upper bound and the packing heuristic is solved with a tolerance of 10^{-6}; the computational algorithm is described in the next section.

CV (cv)	Horizon (T)	Arms (n)	Simultaneous Pulls (k)	Performance: J^μ/J^*			Revocations
				Packing	Whittle Irrev	Whittle	Whittle
High (4)	40	500	125	0.81	0.64	0.89	1685
	40	100	25	0.80	0.64	0.88	340
	40	500	75	0.79	0.68	0.87	723
	40	100	15	0.79	0.68	0.86	149
	25	500	125	0.80	0.68	0.86	1190
	25	100	25	0.80	0.68	0.86	237
	25	500	75	0.78	0.73	0.84	474
	25	100	15	0.78	0.73	0.84	95
	10	500	125	0.79	0.78	0.83	431
	10	100	25	0.79	0.77	0.84	86
	10	500	75	0.78	0.79	0.80	48
	10	100	15	0.78	0.78	0.80	11
Moderate (2.5)	40	500	125	0.87	0.79	0.94	519
	40	100	25	0.85	0.78	0.94	103
	40	500	75	0.86	0.82	0.93	112
	40	100	15	0.85	0.81	0.92	25
	25	500	125	0.85	0.80	0.92	336
	25	100	25	0.84	0.79	0.92	67
	25	500	75	0.83	0.84	0.91	72
	25	100	15	0.82	0.83	0.89	15
	10	500	125	0.82	0.81	0.85	84
	10	100	25	0.82	0.82	0.85	20
	10	500	75	0.80	0.86	0.86	26
	10	100	15	0.80	0.86	0.86	4
Low (1)	40	500	125	0.93	0.95	0.99	60
	40	100	25	0.91	0.93	0.98	14
	40	500	75	0.92	0.99	0.99	17
	40	100	15	0.90	0.99	0.99	3
	25	500	125	0.91	0.97	0.98	34
	25	100	25	0.91	0.95	0.98	9
	25	500	75	0.92	0.98	0.98	19
	25	100	15	0.90	0.98	0.98	3
	10	500	125	0.89	0.95	0.96	19
	10	100	25	0.89	0.95	0.96	3
	10	500	75	0.90	0.96	0.96	11
	10	100	15	0.89	0.96	0.96	2

Table 1: Computational Summary. Each row represents the performance of three different heuristics for $\alpha = 0.2$, $m = 2$, and β chosen to satisfy the corresponding coefficient of variation. Performance for each instance was computed from 3000 simulations of that instance.

of a-priori uncertainty in arm bias (or in the context of the dynamic assortment application, for example, product popularity).

For each combination of the parameters above, we evaluate the packing heuristic, Whittle's heuristic and a natural 'irrevocable' modification to Whittle's heuristic. In particular, this irrevocable modification selects, at every point in time, to pull the k arms with the highest Gittin's index among all arms that are currently active or else, have never been pulled (as opposed to all arms, as would Whittle's heuristic).

We make the following observations:

- **Impact of Initial Coefficient of Variation (cv):** A higher cv represents a high degree of uncertainty in arm bias. Hence, for $cv = 4$, one can potentially gain from exploring a large number of arms before making a decision on the arms to pull for longer periods of time. Thus, Whittle's heuristic has a large number of revocations for higher values of cv, i.e., a large number of arms are discarded and picked again. For an irrevocable heuristic, mistakes – that is, discarding an arm that is performing reasonably in favor of an unexplored arm that turns out to perform poorly – are particularly expensive in such problems. The irrevocable modification to Whittle's heuristic performs poorly because of the restriction that a discarded arm cannot be pulled again; specifically, it loses 20–25% compared to Whittle's heuristic. *In many cases, the packing heuristic is able to recover well over* 50% *of this gap.* This is also true for moderate co-efficients of variation ($cv = 2.5$) in initial arm bias, albeit not for small time horizons (see the point below). For low levels of initial uncertainty in arm bias ($cv = 1$), Whittle's heuristic entails very few revocations and it is thus not surprising that the irrevocable modification to this heuristic also performs well. The packing heuristic is outperformed by both heuristics in this low uncertainty regime, albeit by a few percent; all heuristics are within 11% of an upper bound on achievable performance in the low uncertainty regime.

- **Impact of time horizon (T):** For longer time horizons (T=25 and $T = 40$), it is again reasonable to expect that Whittle's heuristic would entail a large number or revocations (since one may effectively explore all arms before settling on the best). We expect the irrevocable modification to Whittle's heuristic to perform poorly here, as indeed it does. The performance of the packing heuristic is surprisingly consistent, providing a significant advantage over the irrevocable modification to Whittle's heuristic, while losing little in performance relative to Whittle's heuristic. For short time horizons ($T = 10$), all three heuristic are within a few percent of each other with the packing heuristic being dominated by Whittle's heuristic and its irrevocable modification.

To summarize, the packing heuristic provides excellent performance across regimes characterized by a moderate to high degree of uncertainty in initial arm bias and a relatively longer time horizon; it provides a significant improvement over a natural irrevocable modification to Whittle's heuristic in these regimes while being almost competitive with Whittle's heuristic itself. At low levels of uncertainty and/or short time horizons, the packing heuristic is inferior to both Whittle's heuristic as also its irrevocable modification, albeit by a small margin.

Horizon (T)	Arms (n)	Simultaneous Pulls (k)	Performance: J^μ/J^*			Revocations Whittle
			Packing	Whittle Irrev	Whittle	
40	501	125	0.91	0.80	0.92	1983
40	99	25	0.91	0.80	0.92	389
40	501	75	0.88	0.80	0.91	1055
40	99	15	0.88	0.79	0.90	214
25	501	125	0.90	0.83	0.92	1376
25	99	25	0.88	0.82	0.92	264
25	501	75	0.87	0.83	0.90	699
25	99	15	0.88	0.83	0.89	142
10	501	125	0.89	0.90	0.92	322
10	99	25	0.88	0.90	0.91	59
10	501	75	0.85	0.86	0.87	120
10	99	15	0.83	0.88	0.88	26

Table 2: Computational Summary. Each row represents the performance of three different heuristics for $M = 2$, $\alpha/\beta = 0.05$. Each instance consisted of an equal number of bandits with CVs of 1, 2.5, 4.0. Performance for each instance was computed from 3000 simulations of that instance.

5.2. Non IID Bandit Arms

We now consider a model with an equal number of three different categories of bandits: each category has a distinct cv, but the ratio (α/β) is equal across categories, i.e., we have the same initial mean for arm bias P_i for every arm i. The maximum number of arrivals in a given time slot is $m = 1$. The results are summarized in Table 5.2. We see that for moderate to long time horizons, the packing heuristic is effectively *competitive* with Whittle's heuristic even though the latter resorts to a very large number of arm revocations! For these time horizons, the irrevocable modification to Whittle's heuristic is substantially inferior to both the packing heuristic and Whittle's heuristic; this is intuitive given the large number or revocations incurred by Whittle's heuristic for these time horizons. For short time horizons, all three heuristics are quite close, with the packing heuristic being marginally inferior to Whittle's heuristic and its irrevocable modification.

We thus see the same merits for the packing heuristic when the bandit arms are not i.i.d. In fact, the advantages of the packing heuristic are further accentuated in this setting: the heuristic appears to provide levels of performance essentially identical to Whittle's heuristic, although the latter entails a large number of arm revocations.

6. Fast Computation

This section considers the computational effort required to implement the packing heuristic. We develop a computational scheme that makes the packing heuristic substantially easier to implement than popular index heuristics such as Whittle's heuristic and thus establish that the heuristic is viable from a computational perspective.

The key computational step in implementing the packing heuristic is the solution of the linear program $RLP(\tilde{\pi}_0)$. Assuming that $|\mathcal{S}_i| = O(\Sigma)$ and $|\mathcal{A}_i| = O(A)$ for all i, this linear program has $O(nTA\Sigma)$ variables and each Newton iteration of a general purpose interior point method will

require $O\left((nTA\Sigma)^3\right)$ steps. An interior point method that exploits the fact that bandit arms are coupled via a single constraint will require $O(n(TA\Sigma)^3)$ computational steps at each iteration. We develop a combinatorial scheme to solve this linear program that is in spirit similar to the classical Dantzig-Wolfe dual decomposition algorithm. In contrast with Dantzig-Wolfe decomposition, our scheme is efficient. In particular, the scheme requires $O(nTA\Sigma^2 \log(kT))$ computational steps to solve $RLP(\tilde{\pi}_0)$ making it a significantly faster solution alternative to the schemes alluded to above. Equipped with this fast scheme, it is notable that using the packing heuristic requires $O(nA\Sigma^2 \log(kT))$ computations per time step amortized over the time horizon which will typically be substantially less than the $O(nA\Sigma^2 T)$ computations required per time step for index policy heuristics such as Whittle's heuristic.

Our scheme employs a 'dual decomposition' of $RLP(\tilde{\pi}_0)$. The key technical difficulty we must overcome in developing our computational scheme for the solution of $RLP(\tilde{\pi}_0)$ is the non-differentiability of the dual function corresponding to $RLP(\tilde{\pi}_0)$ at an optimal dual solution which prevents us from recovering an optimal or near optimal policy by direct minimization of the dual function.

6.1. An Overview of the Scheme

For each bandit arm i, define the polytope $D_i(\tilde{\pi}_0) \in \mathbb{R}^{|\mathcal{S}_i||\mathcal{A}_i|T}$ of permissible state-action frequencies for that bandit arm specified via the constraints of $RLP(\tilde{\pi}_0)$ relevant to that arm.

A point within this polytope, π_i, corresponds to a set of valid state-action frequencies for the ith bandit arm. With some abuse of notation, we denote the expected reward from this arm under π_i by the 'value' function:

$$R_i(\pi_i) = \sum_{t=0}^{T-1} \pi_i(s_i, a_i, t) r_i(s_i, a_i).$$

In addition denote the expected number of pulls of bandit arm i under π_i by

$$T_i(\pi_i) = T - \sum_{s_i} \sum_{t} \pi_i(s_i, \phi_i, t).$$

We understand that both $R_i(\cdot)$ and $T_i(\cdot)$ are defined over the domain $D_i(\tilde{\pi}_0)$.

We may thus rewrite $RLP(\tilde{\pi}_0)$ in the following form:

$$\begin{array}{ll} \max. & \sum_i R_i(\pi_i), \\ \text{s. t.} & \sum_i T_i(\pi_i) \leq kT. \end{array} \tag{2}$$

The Lagrangian dual of this program is $DRLP(\tilde{\pi}_0)$:

$$\begin{array}{ll} \min. & \lambda kT + \sum_i \max_{\pi_i} \left(R_i(\pi_i) - \lambda T_i(\pi_i)\right), \\ \text{s. t.} & \lambda \geq 0. \end{array}$$

The above program is convex. In particular, the objective is a convex function of λ. We will show that strong duality applies to the dual pair of programs above, so that the optimal solution to the two programs have identical value. Next, we will observe that for a given value of λ, it is simple to compute $\max_{\pi_i} \left(R_i(\pi_i) - \lambda T_i(\pi_i)\right)$ via the solution of a dynamic program over the state

space of arm i (a fast procedure). Finally it is simple to derive useful a-priori lower and upper bounds on the optimal dual solution λ^*. Thus, in order to solve the dual program, one may simply employ a bisection search over λ. Since for a given value of λ, the objective may be evaluated via the solution of n simple dynamic programs, the overall procedure of solving the dual program $DRLP(\tilde{\pi}_0)$ is fast.

What we ultimately require is the optimal solution to the primal program $RLP(\tilde{\pi}_0)$. One natural way we might hope to do this (that ultimately will not work) is the following: Having computed an optimal dual solution λ^*, one may hope to recover an optimal primal solution, π^* (which is what we ultimately want), via the solution of the problem

$$\max_{\pi_i} \left(R_i(\pi_i) - \lambda^* T_i(\pi_i) \right). \tag{3}$$

for each i. This is the typical dual decomposition procedure. Unfortunately, this last step *need not* necessarily yield a feasible solution to $RLP(\tilde{\pi}_0)$. In particular, solving (3) for $\lambda = \lambda^* + \epsilon$ may result in an arbitrarily suboptimal solution for any $\epsilon > 0$, while solving (3) for a $\lambda \leq \lambda^*$ may yield an infeasible solution to $RLP(\tilde{\pi}_0)$. The technical reason for this is that the Lagrangian dual function for $RLP(\tilde{\pi})$ may be non-differentiable at λ^*. These difficulties are far from pathological, and Example 1 illustrates how they may arise in a very simple example.

Example 1. *The following example illustrates that the dual function may be non-differentiable at an optimal solution, and that it is not sufficient to solve* (3) *for $\lambda \leq \lambda^*$ or $\lambda = \lambda^* + \epsilon$ for an $\epsilon > 0$ arbitrarily small. Specifically, consider the case where we have $n = 2$ identical bandits, $T = 1$, and $K = 1$. Each bandit starts in state s, and two actions can be chosen for it, namely, a and the idling action ϕ. The rewards are $r(s, a) = 1$ and $r(s, \phi) = 0$. Thus, $RLP(\tilde{\pi}_0)$ for this specific case is given by:*

$$\begin{aligned} \max. \quad & \pi_1(s, a, 0) + \pi_2(s, a, 0), \\ \text{s. t.} \quad & \pi_1(s, a, 0) + \pi_2(s, a, 0) \leq 1, \end{aligned}$$

where $\pi_i \in D_i(\tilde{\pi}_0)$, $i = 1, 2$. Clearly, the optimal objective function value for the above optimization problem is 1. The Lagrangian dual function for the above problem is

$$\begin{aligned} g(\lambda) &= \lambda + \max_{\pi_1(s,a,0)} \pi_1(s, a, 0)(1 - \lambda) + \max_{\pi_2(s,a,0)} \pi_2(s, a, 0)(1 - \lambda) \\ &= \begin{cases} 2 - \lambda & \lambda \leq 1 \\ \lambda & \lambda > 1 \end{cases} \end{aligned}$$

Not the dual function is minimized at $\lambda^* = 1$, which is a point of non-differentiability. Moreover, solving (3) *at $\lambda^* + \epsilon$ for any $\epsilon > 0$, gives $\pi_1(s, a, 0) = \pi_2(s, a, 0) = 0$ which is clearly suboptimal. Also, a solution for $0 \leq \lambda \leq \lambda^*$ is $\pi_1(s, a, 0) = \pi_2(s, a, 0) = 1$, which is clearly infeasible.*

Notice that in the above example, the average of the solutions to problem (3) for $\lambda = \lambda^* - \epsilon$ and $\lambda = \lambda^* + \epsilon$ *does* yield a feasible, optimal primal solution, $\pi_1(s, a, 0) = \pi_2(s, a, 0) = 1/2$. We overcome the difficulties presented by the non-differentiability of the dual function by computing both upper and lower approximations to λ^*, and computing solutions to (3) for both of these approximations. We then consider as our candidate solution to $RLP(\tilde{\pi}_0)$, a certain convex combination of the two solutions. In particular, we propose algorithm 2, that takes as input the specification of the bandit

and a tolerance parameter ϵ. The algorithm produces a feasible solution to $RLP(\tilde{\pi}_0)$ that is within an additive factor of 2ϵ of optimal.

Algorithm 2 RLP SOLVER

$\lambda^{\text{feas}} \leftarrow r_{\max} + \delta$, for any $\delta > 0, \lambda^{\text{infeas}} \leftarrow 0$.
For all i, $\pi_i^{\text{feas}} \leftarrow \pi_i \in \text{argmax}_{\pi_i} \Big(R_i(\pi_i) - \lambda^{\text{feas}} T_i(\pi_i)\Big)$,
$\pi_i^{\text{infeas}} \leftarrow \pi_i \in \text{argmax}_{\pi_i} \Big(R_i(\pi_i) - \lambda^{\text{infeas}} T_i(\pi_i)\Big)$.
while $\lambda^{\text{feas}} - \lambda^{\text{infeas}} > \frac{\epsilon}{kT}$ **do**
  $\lambda \leftarrow \frac{\lambda^{\text{feas}} + \lambda^{\text{infeas}}}{2}$
  for $i = 1$ to n **do**
    $\pi_i^* \leftarrow \pi_i \in \text{argmax}_{\pi_i} (R_i(\pi_i) - \lambda T_i(\pi_i))$.
  end for
  if $\sum_{i=1}^n T(\pi_i^*) > kT$ **then**
    $\lambda^{\text{infeas}} \leftarrow \lambda, \pi_i^{\text{infeas}} \leftarrow \pi_i^*, \ \forall i$
  else
    $\lambda^{\text{feas}} \leftarrow \lambda, \pi_i^{\text{feas}} \leftarrow \pi_i^*, \ \forall i$
  end if
end while
if $\sum_i T_i(\pi_i^{\text{infeas}}) - T_i(\pi_i^{\text{feas}}) > 0$ **then**
  $\alpha \leftarrow \frac{kT - \sum_i T_i(\pi_i^{\text{feas}})}{\sum_i T_i(\pi_i^{\text{infeas}}) - T_i(\pi_i^{\text{feas}})} \wedge 1$
else
  $\alpha \leftarrow 0$
end if
for $i = 1$ to n **do**
  $\pi_i^{\text{RLP}} \leftarrow \alpha \pi_i^{\text{infeas}} + (1 - \alpha)\pi_i^{\text{feas}}$
end for

It is clear that the bisection search above will require $O(\log(r_{\max} kT/\epsilon))$ steps (where $r_{\max} = \max_{i,s_i,a_i} r(s_i, a_i)$). At each step in this search, we solve n problems of the type in (3), i.e. $\max_{\pi_i} (R_i(\pi_i) - \lambda T_i(\pi_i))$. These subproblems may be reduced to a dynamic program over the state space of a single arm. In particular, we define a reward function $\tilde{r}_i : \mathcal{S}_i \to \mathbb{R}_+$ according to $\tilde{r}_i(s_i, a_i) = r_i(s_i, a_i) - \lambda 1_{a_i \neq \phi_i}$ and compute the value of an optimal policy starting at state $s_{0,i}$ (where s_0 is that state on which $\tilde{\pi}_0$ places mass 1) assuming $\tilde{r}_i$ as the reward function. This requires $O(S^2 AT)$ steps per arm. Thus the RLP Solver algorithm requires a total of $O(nS^2 AT \log(r_{\max} kT/\epsilon))$ computational steps prior to termination. The following theorem, proved in the appendix establishes the quality of the solution produced by the RLP Solver algorithm:

Theorem 4. *RLP Solver produces a feasible solution to $RLP(\tilde{\pi}_0)$ of value at least $OPT(RLP(\tilde{\pi}_0)) - 2\epsilon$.*

The RLP Solver scheme was used for all computational experiments in the previous section. Using this scheme, the largest problem instances we considered were solved in a few minutes on a laptop computer.

7. Concluding Remarks

This paper introduced the 'irrevocable' multi-armed bandit problem as a practical model within which to design policies for a number of interesting learning applications. We have developed a new algorithm for this problem – the packing heuristic – that we have shown performs quite well and is practical for large scale deployment. In particular, we have presented a thorough performance analysis that has yielded uniform approximation performance guarantees as well as guarantees that illustrate a dependence on problem parameters. We have also presented an extensive computational study to support what the theory suggests. In the interest of performance, we have presented a fast implementation of the packing heuristic that is faster than schemes that rely on the computation of Gittins indices.

Perhaps the single most useful outcome of this work has been to show that irrevocability is not necessarily an expensive constraint. This fact is supported by both our theory and computational experiments for a general class of learning applications. While natural 'irrevocable' modifications to schemes that perform well for the classical multi-armed bandit problem (such as Whittles heuristic) may not necessarily achieve this goal, the scheme we provide – the packing heuristic - does.

In addition, the theoretical analysis we provide has indirectly yielded the first performance bounds for an important general class of multi-armed bandit problems that to this point have had surprisingly little theoretical attention. More importantly, the new mode of analysis these problems have called for reveals a tantalizing connection with stochastic packing problems. This paper has furthered that connection.

Moving forward, we anticipate two research directions emerging from the present work. First, it would be interesting to further explore the connection with stochastic packing problems. There exists a vast body of algorithmic work for such problems, and it would be interesting to see what this yields for multi-armed bandit problems. A second direction is exploring the requirement of irrevocability in the non-parametric bandit setting; it is clear that an irrevocable scheme can never be regret optimal. The question to ask is how sub-optimal from a regret perspective can an irrevocable scheme be made in the non-parasmetric setting.

References

Anantharam, V., P. Varaiya, J. Walrand. 1987a. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *Automatic Control, IEEE Transactions on* **32**(11) 968–976.

Anantharam, V., P. Varaiya, J. Walrand. 1987b. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards. *Automatic Control, IEEE Transactions on* **32**(11) 977–982.

Anscombe, F. J. 1963. Sequential medical trials. *Journal of the American Statistical Association* **58**(302) 365–383.

Arlotto, A, S Chick, N. Gans. 2009. Hiring and retention of heterogeneous workers. Working Paper.

Armitage, P., G. Berry, J. N. S. Matthews. 2002. *Statistical methods in medical research*. 4th ed. Wiley-Blackwell.

Bellman, R. E. 1956. A problem in the sequential design of experiments. *Sakhya Ser. A* **30** 221–252.

Bergman, S. W, J.C. Gittins. 1985. *Statistical methods for pharmaceutical research planning*, vol. 67. M. Dekker.

Berry, D. A, B. Fristedt. 1985. *Bandit problems: sequential allocation of experiments*. Monographs on statistics and applied probability, Chapman and Hall.

Bertsimas, D., J. Nino-Mora. 2000. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research* **48**(1) 80–90.

Bhattacharjee, R., A. Goel, S. Khanna, B. Null. 2007. The ratio index for budgeted learning, with applications. *Working Paper* .

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Caro, F., J. Gallien. 2007. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science* **53** 276–292.

Dean, B.C., M.X. Goemans, J. Vondrak. 2008. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Mathematics of Operations Research* **33**(4) 945–964.

Gittins, J, Y. G. Wang. 1992. The learning component of dynamic allocation indices. *The Annals of Statistics* **20**(3) 1625–1636.

Gittins, J.C. 1989. *Multi-armed Bandit Allocation Indices*. Wiley.

Gittins, J.C., D.M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. 241–266.

Glazebrook, K. D., D. J. Wilkinson. 2000. Index-based policies for discounted multi-armed bandits on parallel machines. *The Annals of Applied Probability* **10**(3) 877–896.

Guha, S., K. Munagala. 2007. Approximation algorithms for budgeted learning problems. *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, New York, NY, USA, 104–113.

Lai, T., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematic* **6**(1) 4–22.

Pich, M. T, L. Van der Heyden. 2002. Marks & spencer and zara: Process competition in the textile apparel industry. INSEAD Business Case Study.

Shor, N.Z. 1985. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag.

Weber, R. R., G. Weiss. 1990. On an index policy for restless bandits. *Journal of Applied Probability* **27**(3) 637–648.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability* **25** 287–298.

A. Proofs for Section 3

Lemma 1. $OPT(RLP(\tilde{\pi}_0)) \geq OPT(LP(\tilde{\pi}_0))$

Proof. Let $\hat{\pi}$ be an optimal solution to $LP(\tilde{\pi}_0)$. We construct a feasible solution to $RLP(\tilde{\pi}_0)$ of equal value. In particular, define a candidate solution to $RLP(\tilde{\pi}_0)$, $\bar{\pi}$ according to

$$\bar{\pi}(s_i, a_i, t) = \sum_{\bar{s},\bar{a}:\bar{s}_i=s_i,\bar{a}_i=a_i} \hat{\pi}(\bar{s}, \bar{a}, t)$$

This solution has value precisely $OPT(LP(\tilde{\pi}_0))$. It remains to establish feasibility. For this we first observe that

$$\begin{aligned}
\sum_{s_i',a_i'} P_i(s_i', a_i', s_i)\bar{\pi}_i(s_i', a_i', t-1) &= \sum_{s_i',a_i'} P_i(s_i', a_i', s_i) \sum_{\bar{s},\bar{a}:\bar{s}_i=s_i',\bar{a}_i=a_i'} \hat{\pi}(\bar{s}, \bar{a}, t-1) \\
&= \sum_{s_i',a_i'} P_i(s_i', a_i', s_i) \sum_{\bar{s},\bar{a}:\bar{s}_i=s_i',\bar{a}_i=a_i'} \left(\sum_{\hat{s}_{-i}} \prod_{j\neq i} P(\bar{s}_j, \bar{a}_j, \hat{s}_j) \right) \hat{\pi}(\bar{s}, \bar{a}, t-1) \\
&= \sum_{s',a',\hat{s}:\hat{s}_i=s_i} P(s', a', \hat{s})\hat{\pi}(s', a', t-1) \\
&= \sum_{\hat{s}:\hat{s}_i=s_i,a} \hat{\pi}(\hat{s}, a, t) \\
&= \sum_{a_i} \bar{\pi}_i(s_i, a_i, t)
\end{aligned} \tag{4}$$

Next, we observe that the expected total number of pulls of arm pulls under the policy prescribed by $\hat{\pi}$ is simply

$$\sum_i \sum_{s,t,a:a_i\neq\phi_i} \hat{\pi}(s, a, t)$$

Since the total number of pulls in a given time step under $\hat{\pi}$ is at most k, we have

$$\sum_i \sum_{s,t,a:a_i\neq\phi_i} \hat{\pi}(s, a, t) \leq kT$$

But,

$$\begin{aligned}
\sum_i \sum_{s,t,a:a_i\neq\phi_i} \hat{\pi}(s, a, t) &= \sum_i \sum_t \sum_{s_i,a_i\neq\phi_i} \bar{\pi}(s_i, a_i, t) \\
&= \sum_i \sum_t \left(1 - \sum_{s_i} \bar{\pi}(s_i, \phi_i, t) \right) \\
&= \sum_i \left(T - \sum_{s_i,t} \bar{\pi}(s_i, \phi_i, t) \right),
\end{aligned}$$

so that

$$\sum_i \left(T - \sum_{s_i,t} \bar{\pi}(s_i, \phi_i, t) \right) \leq kT \tag{5}$$

From (4) and (5), $\bar{\pi}$ is indeed a feasible solution to $RLP(\tilde{\pi}_0)$. This completes the proof. ■

B. Proofs for Section 4

Lemma 2. *Given a multi-armed bandit problem with $\mathcal{A}_i = \{p_i, \phi_i\}$ $\forall i$, and*

$$r_i(s_i, p_i) \geq \sum_{s_i' \in \mathcal{S}_i} P_i(s_i, p_i, s_i') r_i(s_i', p_i), \quad \forall i, s_i \in \mathcal{S}_i,$$

we must have

$$E[R_i^{m+1}] - E[R_i^m] \leq E[R_i^m] - E[R_i^{m-1}]$$

for all $0 < m < T$.

Proof. We first introduce some notation. It is clear that the policy μ_i^R induces a Markov process on the state space $\mathcal{S}_i$. We expand this state space, so as to track the total number of arm pulls so that our state space now become $\mathcal{S}_i \times \mathcal{T} \cup \{T\}$. The policy μ_i^R induces a distribution over arm i states for every time $t < T$, which we denote by the variable $\hat{\pi}$. Thus, $\hat{\pi}(s_i, m, t, a_i)$ will denote the probability of being in state (s_i, m) at time t and taking action a_i.

Now,

$$E[R_i^{m+1} - R_i^m] = \sum_{s_i, t<T} \hat{\pi}(s_i, m, t, p_i) r_i(s_i, p_i)$$

and similarly, for $E[R_i^m - R_i^{m-1}]$.

But,

$$\begin{aligned}
\sum_{s_i, t<T} \hat{\pi}(s_i, m, t, p_i) r_i(s_i, p_i) &= \sum_{s_i, t<T-1} \hat{\pi}(s_i, m-1, t, p_i) \left(\sum_{s_i'} P_i(s_i, p_i, s_i') h(s_i', t+1) r_i(s_i', p_i) \right) \\
&\leq \sum_{s_i, t<T-1} \hat{\pi}(s_i, m-1, t, p_i) \left(\sum_{s_i'} P_i(s_i, p_i, s_i') r_i(s_i', p_i) \right) \\
&\leq \sum_{s_i, t<T-1} \hat{\pi}(s_i, m-1, t, p_i) r_i(s_i, p_i) \\
&\leq \sum_{s_i, t<T} \hat{\pi}(s_i, m-1, t, p_i) r_i(s_i, p_i) \\
&= E[R_i^m - R_i^{m-1}]
\end{aligned}$$

where $h(s_i, t) = 1 - \prod_{t'=t}^{T-1} \Pr(\mu_i^R(s_i, t') = \phi_i)$. Here $\prod_{t'=t}^{T-1} \Pr(\mu_i^R(s_i, t') = \phi_i)$ is the probability of never pulling the arm after reaching state s_i at time t so that $h(s_i, t)$ represents the probability of eventually pulling arm i after reaching state s_i at time t. The second inequality follows from the assumption on reward structure in the statement of the Lemma. We thus see that coins satisfy the decreasing returns property. ■

Lemma 4.

$$\sum_{i=1}^{H^*} E[\tilde{R}_i] \geq \frac{1}{2} OPT(RLP(\tilde{\pi}_0)).$$

Proof. For $t \geq 0$, define a function

$$f(t) = \sum_{i=1}^{n} \frac{E[R_i]}{E[T_i]} \left(\left(t - \sum_{j=1}^{i-1} E[T_i] \right)^+ \wedge E[T_i] \right),$$

where $(a \wedge b) = \min(a, b)$ and $(x)^+ = x$ if $x \geq 0$, and 0 otherwise. By construction (i.e. since $\frac{E[R_i]}{E[T_i]}$ is

non-increasing in i), we have that f is a concave function on $[0, kT]$. Now observe that

$$\sum_{i=1}^{H^*} E[\tilde{R}_i] = \sum_{i=1}^{H^*-1} \frac{E[R_i]}{E[T_i]} E[T_i] + \frac{E[R_{H^*}]}{E[T_{H^*}]} \left(kT/2 - \sum_{j=1}^{H^*-1} E[T_i] \right) = f(kT/2).$$

Next, observe that

$$OPT(RLP(\tilde{\pi}_0)) = \sum_{i=1}^{n} \frac{E[R_i]}{E[T_i]} E[T_i] = f(kT).$$

By the concavity of f and since $f(0) = 0$, we have that $f(kT/2) \geq \frac{1}{2} f(kT)$, which yields the result. ■

Lemma 5. *For bandits satisfying the decreasing returns property (Property 1),*

$$E\left[\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2} \right] \geq \frac{1}{2} E[R_{1/2}].$$

Proof. We note that assuming Property 1 implies that $E[R_i^{T/2}] \geq \frac{1}{2} E[R_i]$ for all i. The assertion of the Lemma is then evident – in particular,

$$\begin{aligned} E[R_{1/2}] &= \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} T_j < kT/2 \right) E[R_i] \\ &\leq \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} T_j < kT/2 \right) 2E[R_i^{T/2}] \\ &= 2E\left[\sum_{i=1}^{H^*} \mathbf{1}_{\{\sum_{j=1}^{i-1} T_j < kT/2\}} R_i^{T/2} \right] \end{aligned}$$

where the first and second equality use the fact that R_i and $R_i^{T/2}$ are each independent of T_j for $j \neq i$. ■

Lemma 7. *In the Nth system, all arms with indices smaller than or equal to $\tau_N \wedge n_0 N$ begin operation prior to time T. Moreover, for any $\epsilon > 0$ and almost all $\omega \in \Omega$, $\exists N^\epsilon(\omega)$, such that*

$$\tau_N \geq N n_0 - (1+\epsilon)\sqrt{N n_0 \log\log N n_0}$$

for all $N \geq N^\epsilon(\omega)$.

Proof. We begin with noting that for any $\epsilon > 0$ and almost all $\omega \in \Omega$, $\exists N^\epsilon(\omega)$, such that

$$\sum_{i=1}^{m} T_i \leq m k_0 T/n_0 + (1+\epsilon)\sqrt{m \log\log m} k_0 T/n_0 \tag{6}$$

for all $m \geq N^\epsilon(\omega)$. This is immediate from the Law of the Iterated Logarithm for i.i.d random variables. Now, let us denote $m(N) = N n_0 - (1+\epsilon)\sqrt{N n_0 \log\log N n_0}$. Notice that

$$m(N) k_0 T/n_0 + (1+\epsilon)\sqrt{m(N) \log\log m(N)} k_0 T/n_0 \leq N k_0 T.$$

Thus, by (6), it follows that

$$\sum_{i=1}^{m(N)} T_i \leq N k_0 T$$

eventually. But then it must be that

$$\tau_n \geq m(N)$$

eventually, which yields the second part of the Lemma. Now, for the first part of the Lemma, assume for the sake of contradiction that the highest indexed arm pulled up to time $T-1$ has an index (say, l) smaller than $\tau_N \wedge n_0 N$. That is, this assumption would imply that there exists an $l < \tau_N \wedge n_0 N$ satisfying

$$\sum_{i=1}^{l} T_i \geq N k_0 T.$$

But then, $\tau_N \leq l$ which yields a contradiction and thus the result. ■

Theorem 2.

$$\lim_{N\to\infty} \frac{J_N^{\mu^{\text{packing}}}(s,0)}{J_N^*(s,0)} \geq 1 - \gamma(\min(k_0/n_0, 1-k_0/n_0))$$

Proof. From Lemma 7, it follows that in the Nth system at least $\tau_N \wedge n_0 N$ bandits begin operation prior to time T, i.e. the end of the horizon. Now since at most $k_0 N$ arms can be active at any point in time, it must then be that

$$\begin{aligned} J_N^{\mu^{\text{packing}}}(s,0) &\geq E\left[\sum_{i=1}^{\tau_n \wedge n_0 N} R_i - M_{\tau_n \wedge n_0 N}(k_0 N)\right] \\ &\geq E\left[\sum_{i=1}^{\tau_n \wedge n_0 N} R_i - M_{n_0 N}(k_0 N)\right] \end{aligned}$$

We also have

$$\begin{aligned} J_N^{\mu^{\text{packing}}}(s,0) \geq E\left[\sum_{i=1}^{k_0 N} R_i\right] &\geq E\left[\sum_{i=1}^{n_0 N} R_i - M_{n_0 N}((n_0-k_0)N)\right] \\ &\geq E\left[\sum_{i=1}^{\tau_n \wedge n_0 N} R_i - M_{n_0 N}((n_0-k_0)N)\right] \end{aligned}$$

Now, $\lim_N \frac{E\left[\sum_{i=1}^{\tau_N \wedge n_0 N} R_i\right]}{E\left[\sum_{i=1}^{n_0 N} R_i\right]} = 1$ (by Lemma 7, and since the R_i are bounded) and $J_N^*(s,0) \leq E\left[\sum_{i=1}^{n_0 N} R_i\right]$. Thus, the above bounds, and the fact that $\gamma(\cdot)$ is non-decreasing yield:

$$\lim_{N\to\infty} \frac{J_N^{\mu^{\text{packing}}}(s,0)}{J_N^*(s,0)} \geq 1 - \gamma(k_0/n_0 \wedge 1 - k_0/n_0).$$

■

Lemma 7. *For the Nth bandit problem, we have:*

$$E[R_{1/2}] \geq \frac{1}{2}(1-\kappa(N))OPT(RLP(\tilde{\pi}_0)).$$

where $\kappa(N) = O(N^{-1/2+d})$ and $d > 0$ is arbitrary.

Proof. Observe that for the Nth problem, $H^* = \left\lceil \frac{k_0 NT/2}{k_0 T/n_0} \right\rceil = \left\lceil \frac{N n_0}{2} \right\rceil$. We then have for the Nth problem

and an arbitrary $\epsilon > 0$,

$$E[R_{1/2}] = \sum_{i=1}^{H^*} \Pr\left(\sum_{j=1}^{i-1} T_j < k_0 NT/2\right) E[R_i]$$
$$\geq \sum_{i=1}^{\left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor} \Pr\left(\sum_{j=1}^{i-1} T_j < k_0 NT/2\right) E[R_i].$$

Now, for $i \leq \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor$, we have that $\frac{k_0 NT/2}{ik_0 T/n_0} \geq 1+\epsilon$. Now, for $i \leq \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor$, we have:

$$\begin{aligned}
\Pr\left(\sum_{j=1}^{i} T_j < k_0 NT/2\right) &= 1 - \Pr\left(\sum_{j=1}^{i} T_j \geq k_0 NT/2\right) \\
&\geq 1 - \Pr\left(\sum_{j=1}^{\left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor} T_j \geq k_0 NT/2\right) \\
&\geq 1 - \Pr\left(\sum_{j=1}^{\left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor} T_j \geq \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor k_0 T(1+\epsilon)/n_0\right) \\
&= 1 - \Pr\left(\sum_{j=1}^{\left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor} T_j - E[T_j] \geq \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor \epsilon k_0 T/n_0\right) \\
&\geq 1 - \exp\left(-2k_0^2/n_0^2 \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor \epsilon^2\right)
\end{aligned}$$

where the third inequality is Hoeffding's inequality for independent random variables.

Thus,

$$\begin{aligned}
E[R_{1/2}] &\geq \left(1 - \exp\left(-2k_0^2/n_0^2 \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor \epsilon^2\right)\right) \sum_{i=1}^{\left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor} E[R_i] \\
&\geq \left(1 - \exp\left(-2k_0^2/n_0^2 \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor \epsilon^2\right)\right) \left(\frac{1}{1+\epsilon}\right) \left(\frac{Nn_0 - 2}{Nn_0 + 2}\right) \sum_{i=1}^{H^*} E[R_i] \\
&\geq \left(1 - \exp\left(-2k_0^2/n_0^2 \left\lfloor \frac{Nn_0}{2(1+\epsilon)} \right\rfloor \epsilon^2\right)\right) \left(\frac{1}{1+\epsilon}\right) \left(\frac{Nn_0 - 2}{Nn_0 + 2}\right) \frac{OPT(RLP(\tilde{\pi}_0))}{2}
\end{aligned}$$

where the second inequality follows from the fact that $H^* = \left\lceil \frac{Nn_0}{2} \right\rceil$. By setting $\epsilon = N^{-1/2+d}$ for $d > 0$, arbitrarily small, we have the result.

■

C. Proof of Theorem 4

The following lemma shows that the optimal objective function value of the dual is equal to $OPT(RLP(\tilde{\pi}_0))$. In particular, it shows that Slater's constraint qualification condition holds (see, for example, Boyd and Vandenberghe (2004)).

Lemma 9. *$OPT(RLP(\tilde{\pi}_0)) = OPT(DRLP(\tilde{\pi}_0))$. That is, strong duality holds.*

Proof. To show this, it is sufficient to show that there is a strictly feasible solution to (2), i.e., the inequality is satisfied strictly. This is straightforward – in particular, for each bandit i, set $\pi_i(s_i, \phi_i, t) = \tilde{\pi}_{0,i}(s_i)$ for all s_i and t, where $\tilde{\pi}_{0,i}(s_i)$ is the probability of bandit i starting in state s_i. Set $\pi_i(s_i, a_i, t) = 0$ for $a_i \neq \phi_i$ for all s_i, t. These state action frequencies belong to $D_i(\tilde{\pi}_0)$, and also give $T_i(\pi_i) = 0$. ■

We denote $R^* = OPT(RLP(\tilde{\pi}_0)) = OPT(DRLP(\tilde{\pi}_0))$. Also, define the following set of total running times for all bandits corresponding to a dual variable λ:

$$\mathcal{T}(\lambda) = \left\{ \sum_i T_i(\pi_i) \,\middle|\, \pi_i \in \operatorname*{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda T_i(\pi_i)\right), \ \forall i \right\}.$$

Lemma 10. *If $0 \leq \lambda_1 < \lambda_2$, then*

$$\min \mathcal{T}(\lambda_1) \geq \max \mathcal{T}(\lambda_2).$$

Proof. We denote the objective function in $DRLP(\tilde{\pi}_0)$, i.e., the dual function by:

$$g(\lambda) = \lambda kT + \sum_i \max_{\pi_i} \left(R_i(\pi_i) - \lambda T_i(\pi_i)\right).$$

The slack in the total running time constraint $\sum_i T_i(\pi_i) \leq kT$, i.e. $kT - \sum_i T(\pi_i)$, is a subgradient of g for any π such that $\pi_i \in \operatorname{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda T_i(\pi_i)\right)$ (see Shor (1985)). Thus, the set of subgradients of the dual function g at λ are given by

$$\partial g(\lambda) = \{kT - t : t \in \mathcal{T}(\lambda)\}.$$

Then, since g is a convex function, it follows that for $0 \leq \lambda_1 < \lambda_2$,

$$kT - t_1 \leq kT - t_2, \qquad \forall\ t_1 \in \mathcal{T}(\lambda_1),\ t_2 \in \mathcal{T}(\lambda_2).$$

The lemma then follows. ■

(π^*, λ^*) is an optimal solution for the primal and dual problems if and only if (see, for example, Boyd and Vandenberghe (2004))

$$\begin{aligned} &\pi_i^* \in \operatorname*{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda^* T_i(\pi_i)\right), \\ &\text{either} \quad \lambda^* > 0 \text{ and } \sum_i T_i(\pi_i^*) = kT, \quad \text{or} \quad \lambda^* = 0 \text{ and } \sum_i T_i(\pi_i^*) \leq kT. \end{aligned} \tag{7}$$

We prove the correctness of the *RLP* Solver algorithm separately for the cases when $\lambda^* = 0$ is optimal, and when any optimal solution satisfies $\lambda^* > 0$. We denote the values of the bounds on the dual variable that are computed by the *last iteration* of the *RLP* solver algorithm by λ^{feas} and λ^{infeas}. Recall that,

$$\begin{aligned} \pi_i^{\text{feas}} &\in \operatorname*{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda^{\text{feas}} T_i(\pi_i)\right), \\ \pi_i^{\text{infeas}} &\in \operatorname*{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda^{\text{infeas}} T_i(\pi_i)\right). \end{aligned}$$

We introduce some additional notation:

$$\begin{aligned} T^{\text{feas}} &= \sum_i T_i(\pi_i^{\text{feas}}), \qquad R^{\text{feas}} = \sum_i R_i(\pi_i^{\text{feas}}), \\ T^{\text{infeas}} &= \sum_i T_i(\pi_i^{\text{infeas}}), \quad R^{\text{infeas}} = \sum_i R_i(\pi_i^{\text{infeas}}). \end{aligned}$$

Thus,

$$
\begin{aligned}
g(\lambda^{\text{feas}}) &= \lambda^{\text{feas}} kT + R^{\text{feas}} - \lambda^{\text{feas}} T^{\text{feas}}, \\
g(\lambda^{\text{infeas}}) &= \lambda^{\text{infeas}} kT + R^{\text{infeas}} - \lambda^{\text{feas}} T^{\text{infeas}}.
\end{aligned} \tag{8}
$$

Lemma 11. *If (π^*, λ^*) is a solution to (7) with $\lambda^* = 0$, then*

$$R^* - \left(\alpha R^{\text{infeas}} + (1-\alpha) R^{\text{feas}}\right) \leq \epsilon.$$

Proof. If $\lambda^* = 0$, it follows from (7) that there is some $t \in \mathcal{T}(0)$ such that $t \leq kT$. Hence, it follows from Lemma 10 that for any $\lambda > 0$, $\max \mathcal{T}(\lambda) \leq kT$. Hence, Line 11 of the *RLP* solver algorithm is always invoked, and so, the *RLP* solver algorithm converges to

$$\lambda^{\text{infeas}} = 0 \quad \text{and} \quad 0 < \lambda^{\text{feas}} < \epsilon/(kT).$$

Hence, $\pi_i^{\text{infeas}} \in \text{argmax}_{\pi_i}(R_i(\pi_i))$. Also, $g(\lambda)$ is minimized at $\lambda^* = 0$. Hence, it follows from Lemma 9 that

$$R^* = g(0) = \sum_i \max_{\pi_i} R_i(\pi_i) = R^{\text{infeas}}. \tag{9}$$

Since, $\lambda^{\text{feas}} > 0$, it follows from $T^{\text{feas}} \leq kT$. Hence, we now consider the following three cases:

- *Case 1: $T^{\text{infeas}} \leq kT$.*
  Here, $\alpha = 1$, and hence, using (9) it follows that

  $$R^* - \left(\alpha R^{\text{infeas}} + (1-\alpha) R^{\text{feas}}\right) = 0.$$

- *Case 2: $T^{\text{feas}} = kT$.*
  In this case, $(\pi^{\text{feas}}, \lambda^{\text{feas}})$ satisfy the optimality conditions in (7). Thus, $R^{\text{feas}} = R^*$, and so (since $R^{\text{infeas}} = R^*$ by (9))

  $$R^* - \left(\alpha R^{\text{infeas}} + (1-\alpha) R^{\text{feas}}\right) = 0.$$

- *Case 3: $T^{\text{infeas}} > kT > T^{\text{feas}}$.*
  Since, $g(\lambda)$ is minimized at $\lambda = 0$,

  $$
  \begin{aligned}
  R^* = g(0) \leq g(\lambda^{\text{feas}}) &= \lambda^{\text{feas}} kT + R^{\text{feas}} - \lambda^{\text{feas}} T^{\text{feas}} \\
  \Rightarrow R^* - R^{\text{feas}} &\leq \lambda^{\text{feas}} \left(kT - T^{\text{feas}}\right).
  \end{aligned}
  $$

  Since, $R^* = R^{\text{infeas}}$ (from (9)), and using the fact that $0 < \alpha < 1$ when $T^{\text{infeas}} > kT > T^{\text{feas}}$, we have

  $$
  \begin{aligned}
  R^* - \alpha R^{\text{infeas}} - (1-\alpha) R^{\text{feas}} &= (1-\alpha)(R^* - R^{\text{feas}}) \\
  &\leq (1-\alpha)(kT - T^{\text{feas}})\lambda^{\text{feas}} \\
  &\leq kT \lambda^{\text{feas}} \\
  &\leq \epsilon,
  \end{aligned} \tag{10}
  $$

■

Lemma 12. *If every solution to (7) satisfies $\lambda^* > 0$, then*

$$R^* - \left(\alpha R^{\text{infeas}} + (1-\alpha) R^{\text{feas}}\right) \leq 2\epsilon.$$

Proof. The *RLP* solver algorithm is initialized with $\lambda^{\text{infeas}} = 0$. Since, $\lambda^* > 0$, and $(kT) \in \mathcal{T}(\lambda^*)$ ((7)), it follows from Lemma 10 that $\min \mathcal{T}(0) \geq kT$. But $(kT) \notin \mathcal{T}(0)$, else there would be a solution to (7) that

satisfies $\lambda^* = 0$, leading to a contradiction. Thus, $\min \mathcal{T}(0) > kT$, and so lines 8–12 of the *RLP* solver algorithm guarantee that

$$T^{\text{infeas}} > kT. \tag{11}$$

Using an appropriate modification of the optimality conditions in (7) for the case where the horizon is T^{infeas} (instead of kT), we see that R^{infeas} is the maximum reward earned by any policy in $\{\pi : \sum_i T_i(\pi_i) \leq T^{\text{infeas}}\}$. Since, R^* is the maximum reward earned by any policy in $\{\pi : \sum_i T_i(\pi_i) \leq kT < T^{\text{infeas}}\}$,

$$R^{\text{infeas}} \geq R^*. \tag{12}$$

We now argue that $T^{\text{feas}} \leq kT$. The *RLP* solver algorithm is initialized with $\lambda^{\text{feas}} > r_{max}$. Since, $\pi_i^{\text{feas}} \in \text{argmax}_{\pi_i} \left(R_i(\pi_i) - \lambda^{\text{feas}} T_i(\pi_i)\right)$, initially, the optimal policy is to idle at all times. Thus, $T^{\text{feas}} \leq kT$ at initialization; at all other iterations, lines 8–12 of the algorithm ensure that $T^{\text{feas}} \leq kT$.

We now consider the following two cases separately:

- *Case 1:* $T^{\text{feas}} = kT$.
  In this case, $(\pi^{\text{feas}}, \lambda^{\text{feas}})$ satisfy the optimality conditions in (7), and so, $R^{\text{feas}} = R^*$. Now, using (12)

  $$\left(\alpha R^{\text{infeas}} + (1-\alpha) R^{\text{feas}}\right) \geq R^*.$$

- *Case 2:* $T^{\text{feas}} < kT$.
  Note that the *RLP* solver algorithm terminates when

  $$\lambda^{\text{feas}} - \lambda^{\text{infeas}} < \epsilon/(kT). \tag{13}$$

  Now $T^{\text{feas}} < kT$ and $(kT) \in \mathcal{T}(\lambda^*)$. If $\lambda^{\text{feas}} < \lambda^*$, it follows from Lemma 10 that

  $$T^{\text{feas}} \geq \min \mathcal{T}(\lambda^{\text{feas}}) \geq \max \mathcal{T}(\lambda^*) \geq kT,$$

  which is a contradiction. Hence,

  $$\lambda^{\text{feas}} \geq \lambda^*. \tag{14}$$

  Also, since $(kT) \in \mathcal{T}(\lambda^*)$, it follows from Lemma 10 that for any $\lambda > \lambda^*$, $\max \mathcal{T}(\lambda) \leq kT$. So, (11) implies that

  $$\lambda^{\text{infeas}} \leq \lambda^*. \tag{15}$$

  It follows from (13),(14),(15) that

  $$\max\left(0, \lambda^* - \frac{\epsilon}{kT}\right) \leq \lambda^{\text{infeas}} \quad \text{and} \quad \lambda^{\text{feas}} \leq \lambda^* + \frac{\epsilon}{kT}.$$

  Since $g(\lambda)$ is minimized at λ^*, it follows from (8) and strong duality proved in Lemma 9 that

  $$g(\lambda^*) = R^* \leq g(\lambda^{\text{feas}}) = R^{\text{feas}} + \lambda^{\text{feas}}\left(kT - T^{\text{feas}}\right) \leq R^{\text{feas}} + (\lambda^* + \delta)\left(kT - T^{\text{feas}}\right),$$
  $$g(\lambda^*) = R^* \leq g(\lambda^{\text{infeas}}) = R^{\text{infeas}} + \lambda^{\text{infeas}}\left(kT - T^{\text{infeas}}\right) \leq R^{\text{infeas}} + (\lambda^* - \delta)\left(kT - T^{\text{infeas}}\right),$$

  where $\delta = \epsilon/(kT)$. Note that the above inequalities also use $T^{\text{feas}} < kT$ (by assumption) and $T^{\text{infeas}} >$

kT (from (11)). Thus,

$$\begin{aligned} R^* - \alpha R^{\text{infeas}} - (1-\alpha)R^{\text{feas}} &= \alpha(R^* - R^{\text{infeas}}) + (1-\alpha)(R^* - R^{\text{feas}}) \\ &\leq \alpha(\delta - \lambda^*)\left(T^{\text{infeas}} - kT\right) + (1-\alpha)(\lambda^* + \delta)\left(kT - T^{\text{feas}}\right) \\ &= 2\frac{\delta(T^{\text{infeas}} - kT)(kT - T^{\text{feas}})}{T^{\text{infeas}} - T^{\text{feas}}} \\ &\leq 2\delta kT \\ &= 2\epsilon. \end{aligned}$$

■

Theorem 2. *RLP Solver produce a feasible solution to $RLP(\tilde{\pi}_0)$ of value at least $OPT(RLP(\tilde{\pi}_0)) - 2\epsilon$.*

Proof. The result follows from Lemmas 11 and 12, and the fact that $\lambda^* \geq 0$ (from (7)). ■