Multi-armed Bandits with Metric Switching Costs[*]

Sudipto Guha[†]   Kamesh Munagala[‡]

Abstract

In this paper we consider the stochastic multi-armed bandit with metric switching costs. Given a set of locations (arms) in a metric space and prior information about the reward available at these locations, cost of getting a sample/play at every location and rules to update the prior based on samples/plays, the task is to maximize a certain objective function constrained to a distance cost of L and cost of plays C. This fundamental problem models several stochastic optimization problems in robot navigation, sensor networks, labor economics, etc.

In this paper we consider two natural objective functions – future utilization and finite horizon. We develop a common duality-based framework to provide the *first* approximation algorithms in the metric switching cost model, the approximation ratios being small constant factors. Since both problems are Max-SNP hard, this result is the best possible. We also show an "adaptivity" result, namely, there exists a policy which orders the arms and visits them in that fixed order without revisiting any arm and this policy gives at least $\Omega(1)$ fraction reward of the fully adaptive policy.

The overall technique involves a subtle variant of the widely used Gittins index, and the ensuing structural characterizations will be independently of interest in the context of bandit problems with complicated side-constraints.

1 Introduction

A prevalent paradigm in sensor networks is to use and refine crude probabilistic models of sensed data at various nodes via judiciously sensing and transmitting values [17, 16]; such a paradigm is used, for instance, in the ecological monitoring and forecasting application in the Duke forest [19]. Consider running an *extreme value* query in a sensor network [37]: The base station has crude prior models on the values sensed at each node in the network, and wishes to refine these estimates in order to select a node that is sensing extreme values. The key constraint in this process is energy: It costs energy to measure a value at a particular node, and it costs energy to transfer this process to a new node. The goal is to optimally refine the estimates from the perspective of the query, subject to a budget on the energy consumed in this process.

Such an estimation problem with switching costs was traditionally motivated in several contexts: Price-setting under demand uncertainty [35], decision making in labor markets [28, 33, 7], and resource allocation among competing projects [6, 22]. For instance [7, 6], consider a worker who has a choice of working in k firms. A priori, she has no idea of her productivity in each firm. In each time period, her wage at the current firm is an indicator of her productivity there, and partially resolves the uncertainty in the latter quantity. Her expected wage in the period, in turn, depends on the underlying (unknown) productivity value. At the end of each time step, she can change firms in order to estimate her productivity at different places; however at the end of a "trial" period of duration C, she must choose one firm and stick with it. Her payoff is her expected wage in the finally chosen firm. The added twist is that changing firms does not come for free, and incurs a cost to the worker. What should the strategy of the worker be to maximize her (expected) payoff? A similar problem can be formulated from the perspective of a firm trying out different workers with a priori unknown productivity to fill a post – again, there is a cost to switch between workers. *We make the very reasonable assumption that these costs define a metric.* This assumption is also clearly valid in a sensor network where the communication cost between nodes (and hence the cost to transfer the estimation process to a new node) depends on geographic distance.

The above problems can be modeled using the celebrated stochastic multi-armed bandit (MAB) framework [9, 11, 39] with an additional *switching cost* constraint [1, 2, 7, 6, 13]: We are given a bandit with n independent arms[1]. When arm i is played, we observe a reward drawn from an underlying distribution D_i. A priori, this distribution is unknown; however, a prior $\mathcal{D}_i$ is specified over possible distributions. (For instance, if D_i is a Bernoulli $\{0, a\}$ distribution where the probability of a is an unknown value $t \in [0, 1]$, a possible prior $\mathcal{D}_i$ is the Beta distribution over the parameter t.) As the arm is played, the observed rewards resolve the prior into a posterior distribution over possible reward distributions. The state space of the arm $\mathcal{S}_i$ is the set of possible posterior distributions. Suppose the arm is in state $u \in \mathcal{S}_i$, which corresponds to a certain posterior distribution. Conditioned on this, each reward value is observed with a fixed probability, and causes transition to a different state $v \in \mathcal{S}_i$. Let the expected reward observed on playing in state u be R_u. Let $\mathbf{p}_{uv}$ denote the probability of transitioning to state v given a play in state u. We assume the state space $\mathcal{S}_i$ has polynomial size, and further, the that state transitions define a DAG rooted at the state ρ_i corresponding to the prior distribution $\mathcal{D}_i$. Since the states correspond to evolution of a prior distribution, we have a **martingale property** that $R_u = \sum_{v \in \mathcal{S}_i} R_v \mathbf{p}_{uv}$. The mapping from priors to the state space is standard [11], and is illustrated in Section 2.6 in the context of Dirichlet priors.

The cost model is as follows: If arm i is played in state $u \in \mathcal{S}_i$, the *play cost* is c_u. We will assume $c_u = 1$ for the rest of this section so that the plays can be indexed by time; our results hold

---

[1]In the context of a sensor network, think of the nodes as arms, and the values they sense as random variables.

for arbitrary c_u. The arms are located in a metric space with distance function ℓ. If the previous play was for arm i and the next play is for arm j, the *distance cost* incurred is ℓ_{ij}. A *policy* is an adaptive scheme for playing one arm per step, where the decision about which arm to play depends on the current state of each arm (which in turn depends on the reward sequence observed from previous plays for that arm). The start state of each arm is the root node ρ_i. *The goal is to find the policy that optimizes (expected) reward subject to constraints on both the play cost and distance (or switching) cost.*

The key assumption in all multi-armed bandit problems is that *only the state of an arm being played changes; all other arms retain their present state.* The policy therefore defines a decision tree of depth C over the joint state space of the n arms. Let R^t denote the reward observed for the tth play; the goal is to design a policy that optimizes $\mathbf{E}[\sum_{t=1}^{\infty} \beta_t R^t]$. Different bandit objectives differ in the choice of $\vec{\beta}$; these choices are illustrated, for instance, in [10].

Future Utilization. In the *future utilization* version of the problem, $\beta_{C+1} = 1$, and all other β are zero. This corresponds to choosing the best arm (or best sensor node) at the end of the trial period. At each leaf node of the decision tree of the policy, each arm i is in some final state f_i; the payoff of the policy for this leaf node is $\max_i R_{f_i}$. The goal is to compute the policy (or decision tree) for playing the arms whose expected payoff over all leaf nodes is maximized. We again note that for any feasible policy, there are three constraints:

1. The total play cost is at most C in all decision paths. If the budget C is exhausted, the policy must stop.

2. The total distance cost is at most L in all decision paths. The policy pays this cost whenever it switches the arm being played. If the budget L is exhausted, the policy must stop.

3. The policy has poly-size specification, and can be computed in poly-time. The input specification is of size $O(\sum_i |\mathcal{S}_i|^2)$ which corresponds to the $\mathbf{p}_{uv}, R_u, c_u$ values.

We note that the future utilization objective is already NP-Hard to maximize even when all play costs $c_u = 1$ and a single play reveals the full information about the arm [23], even in the absence of distance costs.

Finite Horizon. We also consider the *finite horizon* version, where $\beta_1 = \beta_2 = \cdots = \beta_C = 1$. Here we seek to maximize the sum of the rewards obtained by the policy in all the plays it makes. The finite horizon objective is more standard, especially in the prior-free setting [5, 3, 4].

Discounted Reward. In the discounted reward setting, $\beta_t = \gamma^t$ for some $\gamma \in (0, 1)$. Though we will not consider this version, we note that due to its computational tractability, this is the most widely studied version of the stochastic MAB problem. In particular, the celebrated Gittins index policy [21] is optimal for this variant.

Results and Techniques: In this paper, we provide an $O(1)$ approximation[2] for the metric switching cost version of the multi-armed bandit problem, for both the future and finite horizon measures. The sensor network motivation also suggests that the policy must have compact description for switching between nodes, in particular: *Is there a strategy which is non-adaptive, that*

---

[2] The *approximation ratio* of an algorithm is the worst case over all inputs of the ratio of the value of the solution it constructs to the value of the optimal solution for that instance. This is a standard measure for the performance of polynomial time algorithms for NP-HARD problems; see for instance [38].

is, does not visit the same arm twice and achieves comparable reward to a fully adaptive optimal strategy? We settle this question also in the affirmative, that is even before playing any arm we can fix an ordering and an associated strategy for each arm to explore.

Our key algorithmic technique is to reduce the multi-armed bandit problem, which is stochastic in nature, to a deterministic combinatorial optimization problem, *orienteering* (see Definition 2). In essence, our algorithm proceeds by considering, via the Lagrangean, a space of *relaxed* decision policies that violate the play cost constraint but respect the distance cost constraint. We show that the optimal solution in this relaxed space reduces to an orienteering problem, for which a $2 + \epsilon$ approximation is known using the results from [12, 8, 15]. At this point, we do not use the Lagrangean to solve the original relaxation, but instead, interpret the Lagrangean as yielding an amortized way of accounting for the reward. This interpretation yields a choice of the Lagrange multiplier and a simple policy that sequentially visits and plays the arms. We analyze this policy by converting the amortization of the reward into a *global* accounting scheme that accounts for the reward based on stopping condition rather than arm by arm. This global method shows the right choice of the Lagrange multiplier in retrospect.

For the future utilization version, this technique yields a $4 + \epsilon$-approximation, which reduces to $3 + \epsilon$ in the absence of switching costs. Our algorithmic technique is similar to that used in [27] in the context of the entirely different *restless bandit* problems (where the state of an arm varies even if it is not played) – this showcases the power of this technique in handling diverse stochastic bandit problems even in the presence of arbitrary side constraints. This technique can be used as a generic framework for reducing bandit-type problems to more manageable combinatorial optimization problems.

Related Work. The future utilization problem without switching costs (a.k.a. *budgeted learning*), is well-known in literature [9, 24, 25, 32, 36, 39]. A preliminary version of this paper [25] presents the first 4-approximation algorithm using an LP based approach. Goel *et al.* [24] extend this to define a *ratio index* policy that computes indexes for each arm in isolation and plays the arm with the highest index. They show that this policy is a 4.54 approximation. This is analogous to the famous *Gittins index* [21, 11] that is optimal for the discounted reward setting. We show the precise connection between these index policies and our technique in Section 2.6, and as a side-effect, show that *not only is our technique generalizable to incorporate switching costs (unlike indices), but in addition, the computation time needed for our algorithm is not any worse than the time needed to compute such indices.* We note that since our problems generalize orienteering, which is already Max-SNP hard, constant factor approximations are best possible (see Section 4).

Though it is widely acknowledged [7, 13] that the right model for bandit problems should have a cost for switching arms, the key hurdle is that even for the discounted reward version, index policies stop being optimal. Banks and Sundaram [7] provide an illuminating discussion of this aspect, and highlight the technical difficulty in designing reasonable policies. In fact, *to the best of our knowledge there was no prior theoretical results on stochastic multi-armed bandit with metric switching costs.* We note that this problem is very different from the Lipshitz MAB considered in [30] where the *rewards* of the bandits correlate to their position in the metric space.

The multi-armed bandit has an extensively studied problem since its introduction by Robbins in [34]. From this starting point the literature diverges into a number of (often incomparable) directions, based on the objective and the information available. In context of theoretical results, the typical goal [14, 29, 18] has been to assume that the agent has absolutely no knowledge of ρ_i (model free assumption) and then the task has been to minimize the "regret" or the lost reward, that is comparing the performance of the agent to a superagent who has full knowledge, and plays the best arm from the start. It is easy to show that with sufficiently small play-cost budgets and

with switching costs, the regret against an all-powerful adversary is too large, and hence the need for the Bayesian approach [9, 11, 39] considered here. The difficulty in the Bayesian approach is computational – these problems can be solved by dynamic programming over the joint (product) state space $\prod_i |\mathcal{S}_i|$ which is exponential in the number of arms. We therefore seek approximation algorithms in this setting.

Roadmap: In Section 2, we present a $4 + \epsilon$ approximation to the future utilization measure. In Section 3, we present a $O(1)$ approximation for the finite horizon measure, using a connection to future utilization implicit in [24]. In Section 4, we present evidence that constant factor approximations are the best possible for the problems we consider.

2   Future Utilization Problem: Notation

Recall that we are given n arms. The distance cost of switching from arm i to arm j is $\ell_{ij} \in \mathcal{Z}^+$; this defines a metric. The arm i has a state space $\mathcal{S}_i$, which is a DAG with root (or initial state) ρ_i. If the arm i is played in a state $u \in \mathcal{S}_i$, it transitions to state $v \in \mathcal{S}_i$ with probability $\mathbf{p}_{uv}$; the play costs $c_u \in \mathcal{Z}^+$ and yields expected reward R_u. Since the states correspond to evolution of a prior distribution, we have a **martingale property** that $R_u = \sum_{v \in \mathcal{S}_i} R_v \mathbf{p}_{uv}$.

The system starts at an arm i_0. A policy, given the outcomes of the actions so far (which decides the current states of all the arms), makes one of the following decisions (i) play the arm it is currently on; (ii) play a different arm (paying the switching cost); or (iii) stop and choose the current arm i in state $u \in \mathcal{S}_i$, obtaining its reward R_u as the payoff. Any policy is subject to **rigid** constraints that the total play cost is at most C and the total distance cost is at most L on all decision paths. Although our algorithm will only stop and choose an arm i when it is already at arm i, we allow the policies to choose any arm as long as the arm was visited sometime in the past.

The **goal** is to find the policy with maximum expected payoff. We focus on constructing, in polynomial time, policies with possibly implicit poly-size specification in the input size $O(\sum_{i=1}^{n} |\mathcal{S}_i|^2)$. Let OPT denote both the optimal solution as well as its value.

2.1   Lagrangean Relaxation

We describe a sequence of relaxations to the optimal policy. We first delete all arms j such that $\ell_{i_0 j} > L$. No feasible policy can reach such an arm without exceeding the distance cost budget. Let $\mathcal{P}$ denote the set of policies on the remaining arms that can perform one of three actions: (1) Choose current arm and ignore this arm in the future; (2) Play current arm; or (3) Switch to different (unchosen) arm. Such policies have no constraints on the play costs, but are required to have distance cost L on all decision paths. Further, a policy $P \in \mathcal{P}$ can choose multiple arms, and obtain payoff equal to the sum of the rewards R_u of the corresponding states. Note that any arm can be chosen at most once, and if so, cannot be played in the future. The original problem had a stricter requirement that only one arm could be chosen, and if a choice is made, the plays stop.

Given a policy $P \in \mathcal{P}$ define the following quantities in expectation over the decision paths: Let $I(P)$ denote the expected number of arms finally chosen; $R(P)$ be the expected reward obtained from the chosen arms; $C(P)$ denote the expected play cost. Note that any policy $P \in \mathcal{P}$ needs to have distance cost at most L on *all* decision paths. Consider the following optimization problem:

$$(M1): \qquad \max_{P \in \mathcal{P}} \left\{ R(P) \,\middle|\, \frac{C(P)}{C} + I(P) \leq 2 \right\}$$

Proposition 2.1. *OPT is feasible for $(M1)$.*

4

Proof. We have $I(OPT) = 1$, $C(OPT) \leq C$; since $OPT \in \mathcal{P}$, this shows it is feasible for $(M1)$. $\square$

Let the optimum solution of $(M1)$ be OPT' and the corresponding policy be P^* such that $R(P^*) = OPT' \geq OPT$. Note that P^* need not be feasible for the original problem, since, for instance, it enforces the play cost constraint only in expectation over the decision paths. We now consider the Lagrangean of the above for $\lambda > 0$, and define the problem $M2(\lambda)$:

$$M2(\lambda): \qquad \max_{P \in \mathcal{P}} \ f_\lambda(P) = 2\lambda + \max_{P \in P} \left(R(P) - \lambda \left(\frac{C(P)}{C} + I(P) \right) \right)$$

Definition 2.1. $V(\lambda) = \max_{P \in P} \left(R(P) - \lambda \left(\frac{C(P)}{C} + I(P) \right) \right)$.

We re-iterate that the only constraint on the set of policies $\mathcal{P}$ is that the distance cost is at most L on all decision paths. The critical insight, which explicitly uses the fact that in the MAB the state of an inactive arm does not change, is the following:

Lemma 2.1. *For any $\lambda \geq 0$, given any $P \in \mathcal{P}$, there exists a $P' \in \mathcal{P}$ that never revisits an arm that it has already played and switched out of, such that $f_\lambda(P') \geq f_\lambda(P)$.*

Proof. We will use the fact that $\mathcal{S}_i$ is finite in our proof. Suppose $P \in \mathcal{P}$ revisits an arm. Consider the deepest point in the decision tree P where at decision node α, it is at arm i, took a decision to move to arm j (child decision node β) and later, on one of the decision paths descending from this point, revisits arm i. Call (α, β) to be "offending".

Note that starting at decision node β, the decision tree P_β satisfies the condition that no arm is revisited. Compress all the actions corresponding to staying in arm j starting at β to a single decision "super-node". This "super node" corresponds to the outcome of plays on arm j starting at β. But the *critical* part is that the plays of the arm j does not affect the state of the other arms (due to independence and the fact that inactive arms do not change state). Also the subtrees that were children of this super-node do not have any action related to arm j by assumption. These subtrees are followed with different probabilities that sum up to 1.

Therefore we can choose the subtree T of the super-node which has the maximum value of $f_\lambda(T)$, and use this subtree irrespective of the outcomes of arm j. This yields a new policy P' so that $f_\lambda(P') \geq f_\lambda(P)$, and furthermore, the distance cost of P' is at most L in all decision paths, so that P' is feasible. By the repeated application of the above, the subtree P_β can be changed to a *path* over super-nodes (which correspond to actions at an arm which is never revisited), without decreasing $f_\lambda(P)$. From now on we will refer to P_β as a "path" in this sense.

Now suppose the arm i corresponding to the play at decision node α, which is the parent of decision node β, is played in this path P_β. The consider moving the entire super-node corresponding to arm i in P_β, to just after node α, so that arm i is played according to this super-node before visiting arm j (as dictated by the decision node β). By independence of the arms, the intermediate plays in P_β before reaching arm i do not affect the state of arm i, and hence the move preserves the states of all the arms. Further, the distance cost of the new policy is only smaller, since the cost of switching into and out of arm i is removed. Note that (α, β) is not "offending" any more, and we have not introduced any new offending pairs of decision nodes. By repeated application of the above the proposition follows. $\square$

Note that the above is *not* true for policies restricted to be feasible for $(M1)$. This is because the step where we use sub-tree T regardless of the outcome of the super-node corresponding to j need not preserve the (implicit) constraint $I(P) \leq 2$, since this depends on whether node j is chosen or not by the super-node. The Lagrangean $M2(\lambda)$ makes the overall objective additive in

the (new) objective values of the super-nodes, with the only constraint being that the distance cost is preserved. Since this cost is preserved in each decision branch, it is preserved by using the best sub-tree T regardless of the outcome of the super-node corresponding to j.

Let $\mathcal{P}_i$ denote the set of all policies that play only arm i. Such a policy at any point in time can (i) play the arm; (ii) stop and choose the arm; or (iii) stop and not choose the arm. Note that the policy can choose the arm at most once in any decision path, but is otherwise unconstrained. Further note that expected play cost $C(P)$ and reward $R(P)$ are defined for such a policy, but not distance cost. The start state of the policy is $\rho_i \in \mathcal{S}_i$, and the state space is $\mathcal{S}_i$.

Definition 1. $Q_i(\lambda) = \max_{P \in \mathcal{P}_i} R(P) - \lambda \left(\frac{C(P)}{C} + I(P) \right)$.

Lemma 2.2. $Q_i(\lambda)$ *and the corresponding policy can be computed in time polynomial in* $|\mathcal{S}_i|$.

Proof. This is a straightforward dynamic program. Let $\mathrm{Gain}(u)$ to be the maximum of the objective of the single-arm policy conditioned on starting at $u \in \mathcal{S}_i$. If u has no children, then if we "play" at node u, then there is no further actions for the policy and so the $\mathrm{Gain}(u) = -\lambda \frac{c_u}{C}$ in this case. Stopping and not doing anything corresponds to $\mathrm{Gain}(u) = 0$. "Choosing" this arm corresponds to a gain of $R_u - \lambda$. Therefore we set $\mathrm{Gain}(u) = \max\{0, R_u - \lambda\}$ in this case.

If u had children, playing would correspond to a gain of $-\lambda \frac{c_u}{C} + \sum_v \mathbf{p}_{uv} \mathrm{Gain}(v)$. Choosing the arm would correspond to gain $R_u - \lambda$. Therefore we have:

$$\mathrm{Gain}(u) = \max \left\{ 0, \quad -\lambda \frac{c_u}{C} + \sum_v \mathbf{p}_{uv} gain(v), \quad R_u - \lambda \right\}$$

We have $\mathrm{Gain}(\rho_i) = Q_i(\lambda)$. This will clearly take polynomial time in $|\mathcal{S}_i|$. $\qquad \square$

Definition 2. *In the* orienteering *problem [12, 8, 15], we are given a metric space $G(V, E)$, where each node $v \in V$ has a reward r_v. There is a start node $s \in V$, and a distance bound L. The goal is to find that tour P starting at s, such that $\sum_{v \in P} r_v$ is maximized, subject to the length of the tour being at most L.*

Now recall that the optimal solution to $M2(\lambda)$ is $2\lambda + V(\lambda)$. An immediate consequence of Lemma 2.1 is the following:

Corollary 2.3. *Define a graph $G(V, E)$, where node $i \in V$ corresponds to arm i. The distance between nodes i and j is ℓ_{ij}, and the reward of node i is $Q_i(\lambda)$. The optimum solution $V(\lambda)$ of $M2(\lambda)$ is the optimal solution to the orienteering problem on G starting at node i_0 and respecting rigid distance budget L.*

Proof. Consider any n-arm policy $P \in \mathcal{P}$. By Lemma 2.1, the decision tree of the policy can be morphed into a sequence of "super-nodes", one for playing each arm, such that the decision about which arm to play next is independent of the outcomes for the current arm. The policy maximizing $f_\lambda(P)$ will therefore choose the best policy in $\mathcal{P}_i$ for each single arm as the "super-node" (obtaining objective value precisely $Q_i(\lambda)$), and visit these subject to the constraint that the distance cost is at most L. This is precisely the orienteering problem on the graph defined above. $\qquad \square$

Theorem 2.4. *For some constant $\epsilon > 0$, assume $Q_i(\lambda) \in \left[\epsilon \frac{\lambda}{n}, \lambda \right]$ for all arms i. Then, $V(\lambda)$ has a $(2 + \epsilon)$-approximation (that we will denote $W(\lambda)$) in time polynomial in $\sum_{i=1}^{n} |\mathcal{S}_i|$.*

Proof. This follows from the $(2 + \epsilon)$-approximation algorithm of Chekuri *et al.* [15] for the orienteering problem. Their result holds only when the $Q_i(\lambda)$ are integers in a polynomial range. If the $Q_i(\lambda) \in \left[\epsilon \frac{\lambda}{n}, \lambda \right]$, then round them in powers of $(1 + \epsilon)$ to make them integers, and then apply the factor $2 + \epsilon$ approximation algorithm. $\qquad \square$

2.2 Choosing and Interpreting the Penalty λ

Recall that optimal value to the problem $(M1)$ is at least OPT. We first relate OPT to the optimal value $2\lambda + V(\lambda)$ of the problem $M2(\lambda)$. We ignore the factor $(1 + \epsilon)$ in the result in Theorem 2.4.

Lemma 2.5. *For any $\lambda \geq 0$, we have $2\lambda + V(\lambda) \geq OPT$.*

Proof. This is simply weak duality: For the optimal policy P^* to $(M1)$, we have $I(P^*) + C(P^*)/C \leq 2$. Since this policy is feasible for $M2(\lambda)$ for any $\lambda \geq 0$, the claim follows. $\quad\square$

Lemma 2.6. *For constant $\delta > 0$, we can choose a λ^* in polynomial time so that for the resulting orienteering tour (of value $W(\lambda)$) on the subset S of arms constructed by the approximation algorithm, we have: (1) $\lambda^* \geq \frac{OPT}{4}(1 - \delta)$; and (2) $W(\lambda) = \sum_{i \in S} Q_i(\lambda^*) \geq \frac{OPT}{4}(1 - \delta)$.*

Proof. First note that as λ increases, the value of any policy $P \in \mathcal{P}_i$ reduces, which implies $Q_i(\lambda)$ decreases. For $\lambda \geq \sum_{i=1}^{n} \sum_{u \in \mathcal{S}_i} R_u$, the optimal policy for arm i does not play the arm, so that $Q_i(\lambda) = 0$ for all arms i. For $\lambda = 0$, we have $Q_i(\lambda) > 0$. This implies the following algorithm to find λ^* such that $W(\lambda^*) \approx \lambda^*$. Start with $\lambda = \sum_{i=1}^{n} \sum_{u \in \mathcal{S}_i} R_u$ and decrease it in powers of $(1 + \epsilon)$. For each value of λ, compute all $Q_i(\lambda)$. Throw out all arms with $Q_i(\lambda) < \epsilon\frac{\lambda}{n}$. This decreases the value of any orienteering tour by at most $\epsilon\lambda$. Now, if some arm i reachable from the root has $Q_i(\lambda) > \lambda$, then the value of the tour is larger than λ, so that λ can be increased further. Otherwise, we have $Q_i(\lambda) \in \left[\epsilon\frac{\lambda}{n}, \lambda\right]$ for all i and we can apply Theorem 2.4 to find a 2 approximation to the reward. Let the value of this approximation be $W(\lambda)$. We have: $W(\lambda) \geq \frac{1}{2}(V(\lambda) - \epsilon\lambda)$. By Lemma 2.5:

$$W(\lambda) \geq \frac{1}{2}(V(\lambda) - \epsilon\lambda) \geq \frac{1}{2}(OPT - (2 + \epsilon)\lambda) \quad \Rightarrow \quad (1 + \epsilon/2)\lambda + W(\lambda) \geq \frac{OPT}{2} \quad (1)$$

As λ is decreased, we encounter a point where: $W(\lambda) \leq \lambda$, and for $\lambda' = \lambda(1 - \epsilon)$, we have $W(\lambda') > \lambda'$. Set $\lambda^* = \lambda'$. For these choices, by Equation (1), we must have: $\lambda = \lambda'(1 + \epsilon) \geq \frac{OPT}{4}(1 - \epsilon)$, and $W(\lambda') \geq \frac{OPT}{4}(1 - \epsilon)$. Now choose $\lambda^* = \lambda'$ and δ as a appropriate function of ϵ. $\quad\square$

As in Theorem 2.4, we will ignore the constant $\delta > 0$ in the remaining analysis.

2.3 Amortized Accounting of the Reward

Consider the single-arm policy $P_i^* \in \mathcal{P}_i$ that corresponds to the value $Q_i(\lambda^*)$. This policy performs one of three actions for each state $u \in \mathcal{S}_i$: (i) Play the arm at cost c_u; (ii) Choose the arm in the current state, and stop; or (iii) Stop, but do not choose the arm. For this policy, note that $R(P_i^*) = Q_i(\lambda^*) + \lambda^*(I(P_i^*) + C(P_i^*)/C)$. This implies the reward $R(P_i^*)$ of this policy can be amortized, so that for state $u \in \mathcal{S}_i$, the reward is collected as follows:

1. An upfront reward of $Q_i(\lambda^*)$ when the play for the arm initiates at the root $\rho \in \mathcal{S}_i$.

2. A reward of $\lambda^* c_u/C$ for playing the arm in $u \in \mathcal{S}_i$.

3. A reward of λ^* when the policy stops and chooses the arm in state $u \in \mathcal{S}_i$.

This amortization shows the following lemma. *The same is not true if the policy P_i^* is executed incompletely*, for instance, if it is terminated prematurely.

Lemma 2.7. *If the arm i is played starting at the root $\rho \in \mathcal{S}_i$ according to policy $\mathcal{P}_i^*$, and the reward is generated according to the above amortized method, then the expected reward of the policy is precisely $Q_i(\lambda^*)$.*

2.4 Final Policy and Analysis

Policy for Future utilization

1. Define the problem $(M1)$. Approximately solve the Lagrangean $M2(\lambda^*)$ for λ^* computed in Lemma 2.6. The policy is an orienteering tour of length at most L on a subset S^* of arms, combined with a single arm policy P_i^* for each arm $i \in S^*$ whose value is $Q_i(\lambda^*)$.

2. Traverse the orienteering tour, and for each arm i encountered, play the arm according to policy P_i^*, with the stopping conditions:

   (a) If P_i^* stops and chooses arm i, then the final policy also stops and chooses arm i.
   (b) If P_i^* stops and does not choose arm i, move to the next arm on the orienteering tour.
   (c) If the arms in S^* are exhausted, then stop.
   (d) If the policy runs out of cost budget C, stop and choose the current arm i.
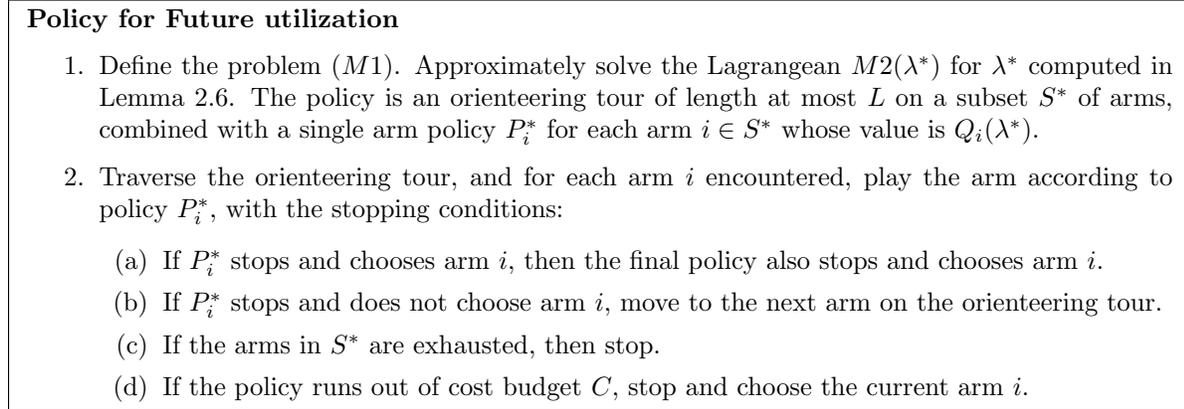

Figure 1: The Final Policy for Future Utilization.

The final policy is now extremely simple, and shown in Figure 1. In order to perform the analysis, first make the following modification to the final policy: If the stopping condition (2d) is encountered and the current arm is i, play the policy P_i^* to completion and then stop and choose arm i. This exceeds the budget C. The original policy differs from the modified policy in that when the budget C is exhausted and the current state is $u \in \mathcal{S}_i$, the original policy chooses the current arm i and obtains reward R_u, while the new policy plays policy P_i^* to completion (and may or may not choose the arm on different decision paths). However, by the **martingale** property of the rewards, the expected future reward of P_i^* even if it chooses the arm on all decision paths is the same as the current reward R_u of state $u \in \mathcal{S}_i$. Therefore, the original policy has at least as much expected reward as the modified policy. Therefore, we will focus on analyzing the modified policy that in Step (2d), plays the current arm i to completion according to policy P_i^*.

Ignoring the ϵ, δ in Theorem 2.4 and Lemma 2.6, we now have:

Lemma 2.8. *The modified final policy, that in Step (2d), plays the current arm i to completion according to policy P_i^*, has reward at least $OPT/4$.*

Proof. For the modified policy, the reward can be accounted for in the amortized sense discussed in Section 2.3. First note that when the policy visits arm i and starts executing the policy P_i^* starting at state $\rho_i \in \mathcal{S}_i$, this policy executes to completion and the execution is independent of the execution of previously played arms. This implies that when the final policy visits arm i, the reward from this arm can be accounted as: Give the system a reward of $Q_i(\lambda^*)$. For every play of cost c, give a reward of $\lambda^* \frac{c}{C}$, and if the arm is chosen (so that the final policy stops and chooses this arm), give a reward of λ^*. It is clear from Lemma 2.7 that the expected reward of the final policy according to this amortization is the same as the expected reward of the finally chosen arm (*i.e.*, the expected reward of the policy). Now, in order to estimate this expected reward, we take a global view of this amortization. Instead of summing the amortized rewards over the random set of arms encountered, we sum these rewards based on the stopping condition the policy encounters. There are three cases:

1. The final policy exhausts all arms. In this case, the accrued amortized reward is at least $\sum_{i \in S} Q_i(\lambda^*) \geq OPT/4$.

2. The final policy runs out of budget C. In that case, since the reward per play is $\frac{\lambda^*}{C}$ per unit cost, the accrued reward is at least $\lambda^* \geq \frac{OPT}{4}$.

8

3. The final policy chooses some arm i. However, the amortized reward per choice is $\lambda^* \geq \frac{OPT}{4}$.

Since the above cases are exhaustive, the modified final policy yields amortized reward at least $\frac{OPT}{4}$, which is the same as the actual reward. $\qquad\square$

As shown above, the reward of the final policy was only larger before we modified Step (2d). Therefore, we have the following theorem:

Theorem 2.9. *For the stochastic multi-armed bandit problem with metric switching costs under the future utilization measure, there exists a poly-time computable ordering of the arms and a policy for each arm, such that a solution which plays the arms using those fixed policies, in that fixed order without revisiting any arm, has reward at least $\frac{1}{4} - \epsilon$ times that of the best adaptive policy, where $\epsilon > 0$ is any small constant.*

The above implies an improved approximation to the *budgeted learning* problem [25, 24], which is the special case where all distances are zero (so that there is no cost for switching between arms).

Corollary 2.10. *For any constant $\epsilon > 0$, the budgeted learning problem has a $3 + \epsilon$ approximation.*

Proof. In this case, the value $V(\lambda)$ of the problem $M2(\lambda)$ is computable in polynomial time, since it is precisely $\sum_{i=1}^{n} Q_i(\lambda)$. Since $2\lambda + V(\lambda) \geq OPT$ by Lemma 2.5, choose λ^* so that $V(\lambda^*) \geq \frac{OPT}{3}$ and $\lambda^* \geq \frac{OPT}{3}$. Now the policy in Figure 1 is trivially a $3 + \epsilon$ approximation. $\qquad\square$

2.5   Extensions

Theorem 2.9 can be easily extended to show some interesting corollaries. We first consider the case of *point priors* and no switching costs. Here, the underlying reward of the arm is deterministic, so that the prior knowledge is a distribution over these values. In this setting, a single play resolves this distribution into a deterministic posterior. Formally, the prior distributional knowledge X_i is a discrete distribution over values $\{a_1^i, a_2^i, \ldots, a_m^i\}$, so that $\Pr[X_i = a_j^i] = p_j^i$ for $j = 1, 2, \ldots, m$. The state space $\mathcal{S}_i$ of the arm is as follows: The root node ρ_i has $r_{\rho_i} = \mathbf{E}[X_i] = \mu_i$. For $j = 1, 2, \ldots, m$, state i_j has $r_{i_j} = a_j^i$, and $\mathbf{p}_{\rho_i i_j} = p_j^i$. Since the underlying reward distribution is simply a deterministic value, the state space is 2-level, defining a star graph with ρ_i being the root, and $i_1, i_2, \ldots, i_m$ being the leaves.

Corollary 2.11. *In the case of point-priors and no switching costs, there is a $3 + \epsilon$ approximate policy that chooses the subset of arms to play in advance, before the results of any play are revealed.*

Proof. The sequential policy constructed in Figure 1 satisfies this property in the case of point priors. In the Lagrangean solution, each arm is either played once or not played at all. Therefore, the sequential policy must necessarily stop after playing the first few arms that cost a total of at most C. If the arms are all played upfront before choosing the final arm, this can only yield larger reward than the reward obtained by the sequential policy. $\qquad\square$

We next consider the case (with switching costs) where the policy is allowed to choose $K > 1$ arms and obtain the sum of the rewards of their final states. In this case OPT would have satisfied $I(P) \leq K$ and the Lagrangean modification would have been $+\lambda(2 - I(P)/K - C(P)/C)$. The remaining policy and analysis would be identical, except for the modification that the sequential policy in Figure 1 stops choosing if it has already chosen K arms. Therefore, we have:

Corollary 2.12. *There is a $4 + \epsilon$ approximation for the future utilization problem where the objective after playing the arms is to choose K arms and obtain the sum of their final rewards.*

We finally note that our analysis assumes that the play costs c_u can depend on the nodes. Suppose now that the cost of making m consecutive plays on an arm is some concave function $f_i(m)$. This is a realistic economies of scale assumption, where performing more tests or experiments with the same setup becomes progressively cheaper, whereas switching to a new setup (or arm) is expensive. We simply make the play cost of states in the mth level in the state space $\mathcal{S}_i$ to be equal to $f_i(m)$. Clearly, the new play cost of a state $u \in \mathcal{S}_i$ is a lower bound on the play cost that OPT would pay if it plays the arm i in state u: this is the actual cost if OPT never switched out of this arm, and is lower than the actual cost if OPT switched out and then back into that arm. Therefore, OPT satisfies the constraint that the total modified play cost on all decision paths is at most C. Since the final policy in Figure 1 does not switch out of an arm, the actual play cost and modified play costs coincide for this policy. Therefore, Theorem 2.9 implies:

Corollary 2.13. *The policy in Figure 1 is a $4 + \epsilon$ approximation to the version of the future utilization problem where the cost of m consecutive plays of arm i is a concave function $f_i(m)$.*

2.6 Discussion

The most widely used policy for the discounted version of the multi-armed bandit problem is the *Gittins index* policy [21]. We consider its adaptation to budgeted learning (no switching costs), *the ratio index*, due to Goel, Khanna, and Null [24]. Recall our interpretation of $Q_i(\lambda)$: Any policy for arm i is given λ per choice on average, and λ/C per play per unit cost on average. The value $Q_i(\lambda)$ is the optimal *excess* reward the policy can earn given these average values, which is $\max_{P \in \mathcal{P}_i} R(P) - \lambda(C(P)/C + I(P))$. This can be generalized to an equivalent definition $Q_i(\lambda, u)$ for policies whose start state is $u \in \mathcal{S}_i$ (instead of being ρ_i). The ratio index for $u \in \mathcal{S}_i$ is:

$$\text{Ratio Index} = \Pi_i(u) = \max\{\lambda \mid Q_i(\lambda, u) > 0\}$$

In other words, unlike our balancing rule, the ratio index is the maximum average reward λ per state, for which the excess reward of the optimal single arm policy with that start state is *zero*. The ratio index policy plays the arm with the highest ratio index every step, and is shown to be a 4.54 approximation for budgeted learning *in the absence of switching costs*. The famous *Gittins index* [21, 11] is similarly defined (refer to the fixed charge problem of Weber [20]) for the discounted version, and is known to be optimal for that version. We note however that for the future utilization version, no index policy can be optimal [32, 24].

The key drawback of these policies is that since the index is defined for each state separately, the execution leads to complicated priorities between arms, so that the algorithms and analyses do not generalize easily when side constraints such as switching costs are present [7, 13]. Our technique provides an alternative method, where a global average reward λ^* is computed by *balancing the total excess reward with the average reward*, and the corresponding single-arm policies are executed sequentially in *arbitrary order*. This has the following advantages:

1. For any λ, computing the value $Q_i(\lambda)$ has running time comparable to computing the ratio index (or Gittins index), since *the dynamic program is the same in either case*.

2. The sequentiality of execution and the indifference to ordering makes it easy to generalize the policies to incorporate switching costs, something that was not previously known for even the discounted reward version where the Gittins index is optimal in the absence of such costs.

3. The $(3 + \epsilon)$ approximation our technique shows for budgeted learning is superior to the 4.54 approximation shown for the ratio index policy [24].

Example of State Space. If the underlying distribution of an arm is Bernoulli$(1, \theta)$, where θ is unknown, its conjugate prior is the Beta distribution. A Beta distribution with parameters $\alpha_1, \alpha_2 \in \{1, 2, \ldots\}$, which we denote $B(\alpha_1, \alpha_2)$ has p.d.f. of the form $c\theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1}$, where c is a normalizing constant. $B(1, 1)$ is the uniform distribution, which corresponds to having no a priori information. The distribution $B(\alpha_1, \alpha_2)$ corresponds to the current (posterior) distribution over the possible values of the bias θ after having observed $(\alpha_1 - 1)$ 0's and $(\alpha_2 - 1)$ 1's. Given this distribution as our belief, the expected value of *theta* is $\frac{\alpha_1}{\alpha_1 + \alpha_2}$.

The state space $\mathcal{S}_i$ is a DAG, whose root ρ_i encodes the initial belief $B(\alpha_1, \alpha_2)$ about θ, so that $r_{\rho_i} = \frac{\alpha_1}{\alpha_1^\rho + \alpha_2}$. When the arm is played in this state, the state evolves depending on the outcome observed – if the outcome is 1, which happens w.p. $\frac{\alpha_1}{\alpha_1 + \alpha_2}$, the child u has belief $B(\alpha + 1, \alpha_2)$, so that $r_u = \frac{\alpha_1 + 1}{\alpha_1 + \alpha_2 + 1}$, and $\mathbf{p}_{\rho u} = \frac{\alpha_1}{\alpha_1 + \alpha_2}$; if the outcome is 0, the child v has belief $B(\alpha_1, \alpha_2 + 1)$, $r_v = \frac{\alpha_1}{\alpha_1 + \alpha_2 + 1}$, and $\mathbf{p}_{\rho v} = \frac{\alpha_2}{\alpha_1 + \alpha_2}$. In general, if the DAG $\mathcal{S}_i$ has depth C (corresponding to playing the arm at most C times), it has $O(C^2)$ states.

Computational Issues. The key bottleneck in the computation is finding the value $Q_i(\lambda)$. As discussed above, this is no harder than computing the ratio index (or Gittins index) for the arm, and takes $O(|\mathcal{S}_i|^2)$ time. Note however that the state space typically corresponds to the evolution of a parametrized prior distribution, so that $|\mathcal{S}_i|$ can be exponential in the implicit representation of this distribution. This is also typically a concern even for the computation of the Gittins index. However, we show below that for the canonical Dirichlet prior over multinomial distributions, the state space can be made poly-bounded in C if we are willing to tolerate an additive ϵ loss in reward. Such guarantees seem possible for other parametrized densities as well. Since this is not the main focus of this paper, we have assumed throughout that Gittins index-type computations can be performed efficiently.

In the case of Dirichlet priors, there are K possible reward values, $0 \leq s_1 \leq s_2 \leq \cdots \leq s_K = 1$. The underlying distribution is an unknown multinomial distribution D over these values, where $\Pr[D = s_j] = x_j$. At the outset, a Dirichlet prior $\mathcal{D}(n_1^0, n_2^0, \ldots, n_K^0)$ is specified over possible $\{(x_1, x_2, \ldots, x_K) \mid \sum_{j=1}^K x_j = 1\}$. Here, each n_i^0 is a positive integer.

For Dirichlet prior $\mathcal{D}(n_1, n_2, \ldots, n_K)$, for any $x_1, x_2, \ldots, x_K$ with $\sum_{j=1}^K x_j = 1$, we have:

$$\Pr\left[\mathcal{D} = (x_1, x_2, \ldots, x_K)\right] \propto \Pi_{j=1}^K x_j^{n_j - 1}$$

A Dirichlet prior evolves in a simple fashion. If the current prior is $\mathcal{D}(n_1, n_2, \ldots, n_K)$, the probability of observing s_j is $n_j / (\sum_{k=1}^K n_k)$, and conditioned on observing s_j, the posterior distribution is $\mathcal{D}(n_1, n_2, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_K)$. This defines the evolution of the state space: The root ρ_i has vector $\vec{N}_0 = (n_1^0, n_2^0, \ldots, n_K^0)$. The root has K children; child j is reached when s_j is observed on playing, which happens w.p. $n_j^0 / (\sum_{k=1}^K n_k^0)$, and resulting the child state corresponds to the posterior $\mathcal{D}(n_1^0, n_2^0, \ldots, n_{j-1}^0, n_j^0 + 1, n_{j+1}^0, \ldots, n_K^0)$. If the current state u corresponds to prior $\mathcal{D}(n_1, n_2, \ldots, n_K)$, the expected reward $R_u = (\sum_{j=1}^K s_j n_j) / (\sum_{k=1}^K n_k)$.

Therefore, over C time steps, the number of possible posterior distributions is the same as the number of multi-sets of size C over the ground set $\{s_1, s_2, \ldots, s_K\}$, which is at most C^K. This is an upper bound on $|\mathcal{S}_i|$, and is polynomial if K is a constant. Therefore, for Beta priors (defined over Bernoulli variables) that correspond to $K = 2$, the state space is polynomial.

If K is large, consider a different ground set of rewards $\{\epsilon, 2\epsilon, 3\epsilon, \ldots, 1\}$ of size $q = \frac{1}{\epsilon}$. For $j = 1, 2, \ldots, q$, let U_j denote the set of original rewards $s_1, s_2, \ldots, s_K$ that fall in the interval $((j - 1)\epsilon, j\epsilon]$. Whenever the original policy observes reward $s_k \in U_j$, pretend that the reward observed was $j\epsilon$. For the prior $\mathcal{D}(n_1^0, n_2^0, \ldots, n_K^0)$, define $m_j^0 = \sum_{s_k \in U_j} n_k^0$, and pretend that the prior

is $\mathcal{D}(m_1^0, m_2^0, \ldots, m_q^0)$. This corresponds to modifying each underlying multinomial distribution so that if the original distribution D had $\Pr[D = s_j] = x_j$, the new distribution D' has $y_k = \Pr[D' = k\epsilon] = \sum_{s_j \in U_k} x_j$. The prior $\mathcal{D}(n_1^0, n_2^0, \ldots, n_K^0)$ over possible $(x_1, x_2, \ldots, x_K)$ now yields in a natural fashion the prior $\mathcal{D}(m_1^0, m_2^0, \ldots, m_q^0)$ over possible $(y_1, y_2, \ldots, y_q)$. In this process, since each multinomial distribution is translated by at most ϵ while the probability distribution over possible distributions is preserved, the reward of the optimal policy is shifted by at most ϵ. The new state space has size $C^{1/\epsilon}$, which is polynomial for any constant $\epsilon > 0$. This implies that by losing an additive ϵ in the reward, the state space can be made poly-bounded for Dirichlet priors.

3 The Finite Horizon Version

In this problem, the description of the state space, distances, and costs are as before. The key difference is in the description of the policy and the objective function: In this case, any policy is faced with three decisions: (i) make a play at the same arm as the previous play (ii) make a play at a different arm and pay the switching cost, or (iii) stop. Note that the key difference from before is that the policy does not need to choose an arm finally. As before, the policy is subject to **rigid** constraints that the total play costs sum to at most C and the total distance costs sum to at most L. At any step when the policy plays arm i in state $u \in \mathcal{S}_i$, it obtains expected reward R_u. The goal is to find the feasible policy whose total expected reward is maximized. Unlike before where the payoff was the reward of the finally chosen arm, in this case, the policy's payoff is the total reward of all the plays it makes. The system starts in arm i_0. Let the optimal policy be OPT.

Assumption. The play costs are uniform for an arm, *i.e.* $c_u = c_i$ for all $i, u \in \mathcal{S}_i$, where $0 \le c_i \le C$.

To solve the finite horizon problem with switching costs, we could use the same argument as for future utilization; however, this would violate the cost constraint by a constant factor, as we cannot use the martingale argument to pre-maturely terminate a policy P_i^* when the cost budget exceeds. We instead use an argument from [24] relating the future and finite horizon measures.

3.1 Defining a Future Utilization Problem

First, separate out arms with $c_i > C/2$. Only one of these arms can be played by any policy; consider the best of these. If this yields more than half the reward of the optimal policy, we are already done, else the remaining arms yield at least half the reward. We therefore assume $c_i \le \frac{C}{2}$ for all arms i.

In the future utilization problem, we *modify* the rewards of the states: A policy obtains modified reward $r_u = \frac{R_u}{c_i}$ if arm i in state $u \in \mathcal{S}_i$ is finally chosen. Define the optimal value of this future utilization problem with rewards r_u and cost budget C as V_C; let the value of the corresponding relaxation $(M1)$ be M_C; and finally let J_C denote the value of the optimal finite horizon solution using the original rewards R_u. Note that by the results in the previous section, we can compute a constant factor approximation to M_C and V_C.

Lemma 3.1 (also in [24]). $V_C \ge \frac{J_C}{C}$.

Proof. Consider the optimal finite horizon policy of value J_C. Consider this as spending play cost continuously and obtaining reward r_u per unit cost in state $u \in \mathcal{S}_i$ continuously. The average reward per unit cost is $\frac{J_C}{C}$. This implies there is a certain cost c^* that the expected reward per unit cost being obtained at the point when c^* cost has been spent is at least $\frac{J_C}{C}$. The future reward policy executes the past reward policy until the next play will exceed c^* cost, and then chooses the

arm that would be played next. This has expected reward V_C of the future utilization policy is therefore at least $\frac{J_C}{C}$. □

3.2 Algorithm and Analysis

Algorithm for Finite Horizon Version

1. Solve the future utilization problem on rewards r_u with play cost budget $C/2$.

2. Execute the future utilization policy.

3. Play the arm chosen by this policy using budget $C/2$. This does not incur extra distance cost.

Figure 2: The policy for finite horizon version.

The overall algorithm is simple and shown in Figure 2. We first have the following lemma:

Lemma 3.2. *The value of the policy in Figure 2 is $\frac{C}{8}M_{C/2}$.*

Proof. In some decision path of the future utilization policy executed in Step (2), suppose arm i in state $u \in \mathcal{S}_i$ is chosen. Then the value gained by the future utilization problem is $r_u = \frac{R_u}{c_i}$. Since all future plays in Step (3) cost the same and since $C \geq 2c_i$, the number of subsequent plays (including the play in Step (2)) is $\lfloor \frac{C}{2c_i} \rfloor \geq \frac{C}{4c_i}$. Using the **martingale** property of the rewards, the expected reward in Step (3) is at least $\frac{C}{2c_i}R_u = \frac{C}{2}r_u$. This implies the reward in Step (3) for the policy is at least $\frac{C}{2}$ times the reward of the future utilization policy executed in Step (2), whose value in turn is at least $\frac{1}{4}M_{C/2}$. Therefore, the value in Step (3) is $\frac{C}{8}M_{C/2}$. □

Lemma 3.3. $M_{C/2} \geq \frac{1}{2}M_C$.

Proof. Consider any policy P feasible for the problem $(M1)$ with cost budget C. Play this policy w.p. $1/2$, and with the remaining probability simply choose the first arm i_0. This policy has reward at least $R(P)/2$, cost $C(P)/2$, and probability of choosing any arm $I(P) = 1$. This implies this is feasible for $(M1)$ with cost budget $C/2$, with half the reward. □

Theorem 3.4. *For the finite horizon measure, the policy in Figure 2 has reward $\frac{C}{16}J_C$. This implies there exists an ordering on arms and a simple policy on arms that gives a $O(1)$ approximation to the reward of a fully adaptive solution.*

Proof. The value of the policy is $\frac{C}{8}M_{C/2}$. From the previous lemma, this is $\frac{C}{16}M_C$. This is $\frac{C}{16}V_C$, since $M_C \geq V_C$. Using Lemma 3.1, this is at least $\frac{C}{16}\frac{J_C}{C}$, as desired. □

The above result can now be simplified when all distances are zero and all play costs are unit. Using the same idea as Corollary 2.10, we have:

Corollary 3.5. *The finite horizon multi-armed bandit problem (which is the finite horizon problem with zero distances and unit play costs) admits to a $12 + \epsilon$ approximation for any constant $\epsilon > 0$.*

4 Lower Bounds

The main goal of the paper was the constant factor analysis, which the following hardness result shows to be best possible:

13

Theorem 4.1. *The future utilization measure under strict distance constraint is Max-SNP hard.*

Proof. We use the fact that orienteering is Max-SNP hard, see [12].

Suppose we could solve the future utilization measure upto a factor $(1 + \epsilon/3)$. Given a gap instance of (max sum) orienteering where node i has reward r_i (let $R = \max_i r_i$); the prior distribution $\mathcal{D}_i$ is that either the arm i gives a fixed reward 1 (which happens with probability $\epsilon r_i/(3nR)$) or gives a fixed reward 0. The play cost is $C = n$.

Note that a single play reveals full information in this case and the play costs are irrelevant. We need to optimize the probability of seeing an arm with reward 1. Suppose a solution visits a set T then that probability is at most $\sum_{i \in T} \epsilon r_i/(3nR) \leq \epsilon/3$. But that probability is also at least $(1 - \epsilon/3) \sum_{i \in T} \epsilon r_i/(3nR)$, which uses the fact that we do not stop at an arm $i \in T$ with probability at least $1 - \epsilon/3$. Therefore we can determine the optimum T for the orienteering problem upto a factor $(1 + \epsilon/3)/(1 - \epsilon/3) < 1 + \epsilon$. □

Note that the reduction in Theorem 4.1 had $C = n$, and the main constraint was the distance constraint L. For the finite horizon version, the exact same reduction shows that when $C \gg L$ is very large the finite horizon measure is Max-SNP hard, because the total reward is (ignoring the lower order terms) the largest reward found in the exploration (the future utilization) times C. Therefore the above result is the best possible as well, if we obey the distance costs.

5 Conclusion

There are several challenging open questions arising from this work; we mention two of them. First, we conjecture that constructing a (possibly adaptive) strategy for the budgeted learning problem, which has no switching costs, is also APX-HARD, *i.e.*, there exists an absolute constant $c > 1$ such that it is NP-HARD to produce a solution which is within factor c times the optimum. Secondly, we have focused exclusively on utility maximization, *i.e.*, maximizing rewards of chosen arms; it would be interesting to explore other objectives, such as minimizing residual information [31].

References

[1] R. Agrawal, M. V. Hegde, and D. Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching costs. *IEEE Transactions on Optimal Control*, 33:899–906, 1988.

[2] M. Asawa and D. Teneketzis. Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, 41(3):328–348, 1996.

[3] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *Proc. of the 1995 Annual Symp. on Foundations of Computer Science*, pages 322–331, 1995.

[6] J. S. Banks and R. K. Sundaram. Denumerable-armed bandits. *Econometrica*, 60(5):1071–1096, 1992.

[7] J. S. Banks and R. K. Sundaram. Switching costs and the gittins index. *Econometrica*, 62(3):687–694, 1994.

[8] N. Bansal, A. Blum, S. Chawla, and A. Meyerson. Approximation algorithms for deadline-tsp and vehicle routing with time-windows. *STOC*, pages 166–174, 2004.

[9] D. A. Berry and B. Fristed. *Bandit problems; Sequential Allocation of Experiments*. Chapman & Hall, New York, 1985.

[10] D.A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, UK, 1985.

[11] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, second edition, 2001.

[12] A. Blum, S. Chawla, D. R. Karger, T. Lane, A. Meyerson, and M. Minkoff. Approximation algorithms for orienteering and discounted-reward TSP. *SIAM J. Comput.*, 37(2):653–670, 2007.

[13] M. Brezzi and T-L. Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108, 2002.

[14] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

[15] C. Chekuri, N. Korula, and M. Pál. Improved algorithms for orienteering and related problems. In *SODA*, pages 661–670, 2008.

[16] D. Chu, A. Deshpande, J. Hellerstein, and W. Hong. Approximate data collection in sensor networks using probabilistic models. In *ICDE*, 2006.

[17] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *VLDB*, 2004.

[18] A. Flaxman, A. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Annual ACM-SIAM Symp. on Discrete Algorithms*, 2005.

[19] P. G. Flikkema, P. K. Agarwal, J. S. Clark, C. S. Ellis, A. Gelfand, K. Munagala, and J. Yang. Model-driven dynamic control of embedded wireless sensor networks. In *Proc. 6th Intl. Conf. on Computational Science (3)*, pages 409–416, 2006.

[20] E. Frostig and G. Weiss. Four proofs of Gittins' multi-armed bandit theorem. *Applied Probability Trust*, 1999.

[21] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in statistics (European Meeting of Statisticians)*, 1972.

[22] J.C. Gittins. *Multi-Armed Bandit Allocation Indices*. Wiley, New York, 1989.

[23] A. Goel, S. Guha, and K. Munagala. Asking the right questions: Model-driven optimization using probes. In *Proc. of the 2006 ACM Symp. on Principles of Database Systems*, 2006.

[24] A. Goel, S. Khanna, and B. Null. The ratio index for budgeted learning, with applications. In *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2009.

[25] S. Guha and K. Munagala. Sequential design of experiments via linear programming. In *CoRR, arxiv:0805.0766*, 2008. Preliminary version in STOC '07.

[26] S. Guha and K. Munagala. Multi-armed bandits with metric switching costs. In *ICALP*, 2009.

[27] S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *CoRR*, abs/0711.3861, 2007. Preliminary version in SODA '09.

[28] B. Jovanovich. Job-search and the theory of turnover. *J. Political Economy*, 87:972–990, 1979.

[29] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. In *Proc. of 16th Conf. on Computational Learning Theory*, 2003.

[30] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandit problems in metric spaces. *Proc. 40th ACM Symposium on Theory of Computing*, 2008.

[31] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. *Twenty-first Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, 2005.

[32] O. Madani, D. J. Lizotte, and R. Greiner. Active model selection. In *UAI '04: Proc. 20th Conf. on Uncertainty in Artificial Intelligence*, pages 357–365, 2004.

[33] D. Mortensen. Job-search and labor market analysis. In O. Ashenfelter and R. Layard, editors, *Handbook of Labor Economics*, volume 2, pages 849–919. North Holland, Amsterdam, 1985.

[34] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[35] M. Rothschild. A two-armed bandit theory of market pricing. *J. Economic Theory*, 9:185–202, 1974.

[36] J. Schneider and A. Moore. Active learning in discrete input spaces. In 34th *Interface Symp.*, 2002.

[37] A. Silberstein, K. Munagala, and J. Yang. Energy-efficient extreme value estimation in sensor networks. In *SIGMOD*, 2006.

[38] V. Vazirani. *Approximation Algorithms*. Springer, 2001.

[39] G. B. Wetherill and K. D. Glazebrook. *Sequential Methods in Statistics (Monographs on Statistics and Applied Probability)*. Chapman & Hall, London, 1986.