Non-monotonic Resource Utilization in the Bandits with Knapsacks Problem

Raunak Kumar
Department of Computer Science
Cornell University
Ithaca, NY 14853
raunak@cs.cornell.edu

Robert D. Kleinberg
Department of Computer Science
Cornell University
Ithaca, NY 14853
rdk@cs.cornell.edu

Abstract

Bandits with knapsacks (BwK) [6] is an influential model of sequential decision-making under uncertainty that incorporates resource consumption constraints. In each round, the decision-maker observes an outcome consisting of a reward and a vector of nonnegative resource consumptions, and the budget of each resource is decremented by its consumption. In this paper we introduce a natural generalization of the stochastic BwK problem that allows non-monotonic resource utilization. In each round, the decision-maker observes an outcome consisting of a reward and a vector of resource *drifts* that can be positive, negative or zero, and the budget of each resource is incremented by its drift. Our main result is a Markov decision process (MDP) policy that has *constant* regret against a linear programming (LP) relaxation when the decision-maker *knows* the true outcome distributions. We build upon this to develop a learning algorithm that has *logarithmic* regret against the same LP relaxation when the decision-maker *does not know* the true outcome distributions. We also present a reduction from BwK to our model that shows our regret bound matches existing results [14].

1 Introduction

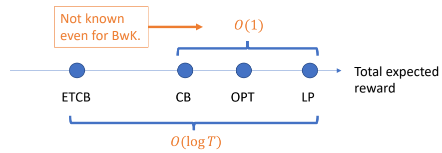
Multi-armed bandits are the quintessential model of sequential decision-making under uncertainty in which the decision-maker must trade-off between exploration and exploitation. They have been studied extensively and have numerous applications, such as clinical trials, ad placements, and dynamic pricing to name a few. We refer the reader to Bubeck and Cesa-Bianchi [7], Slivkins [17], Lattimore and Szepesvári [13] for an introduction to bandits. An important shortcoming of the basic stochastic bandits model is that it does not take into account resource consumption constraints that are present in many of the motivating applications. For example, in a dynamic pricing application the seller may be constrained by a limited inventory of items that can run out well before the end of the time horizon. The bandits with knapsacks (BwK) model [18, 19, 5, 6] remedies this by endowing the decision-maker with some initial budget for each of m resources. In each round, the outcome is a reward and a vector of nonnegative resource consumptions, and the budget of each resource is decremented by its consumption. The process ends when the budget of any resource becomes nonpositive. However, even this formulation fails to model that in many applications resources can get replenished or renewed over time. For example, in a dynamic pricing application a seller may receive shipments that increase their inventory level.

Contributions In this paper we introduce a natural generalization of BwK by allowing non-monotonic resource utilization. The decision-maker starts with some initial budget for each of m resources. In each round, the outcome is a reward and a vector of resource *drifts* that can be positive, negative or zero, and the budget of each resource is incremented by its drift. A negative drift has

Preprint. Under review.

the effect of decreasing the budget akin to consumption in BwK and a positive drift has the effect of increasing the budget. We consider two settings: (i) when the decision-maker *knows* the true outcome distributions and must design a Markov decision process (MDP) policy; and (ii) when the decision-maker *does not know* the true outcome distributions and must design a learning algorithm.

Our main contribution is an MDP policy, `ControlBudget(CB)`, that has *constant* regret against a linear programming (LP) relaxation. Such a result was not known even for BwK. We build upon this to develop a learning algorithm, `ExploreThenControlBudget(ETCB)`, that has *logarithmic* regret against the same LP relaxation. We also present a reduction from BwK to our model and show that our regret bound matches existing results.



Instead of merely sampling from the optimal probability distribution over arms, our policy samples from a perturbed distribution to ensure that the budget of each resource stays close to a decreasing sequence of thresholds. The sequence is chosen such that the expected leftover budget is a constant and proving this is a key step in the regret analysis. Our work combines aspects of related work on logarithmic regret for BwK [9, 14].

Related Work Multi-armed bandits have a rich history and logarithmic instance-dependent regret bounds have been known for a long time [12, 3]. Since then, there have been numerous papers extending the stochastic bandits model in a variety of ways [4, 16, 11, 6, 10, 2, 1].

To the best of our knowledge, there are three papers on logarithmic regret bounds for BwK. Flajolet and Jaillet [9] showed the first logarithmic regret bound for BwK. In each round, their algorithm finds the optimal basis for an optimistic version of the LP relaxation, and chooses arms from the resulting basis to ensure that the average resource consumption stays close to a pre-specified level. Even though their regret bound is logarithmic in T and inverse linear in the suboptimality gap, it is exponential in the number of resources. Li et al. [14] showed an improved logarithmic regret bound that is polynomial in the number of resources, but it scales inverse quadratically with the suboptimality gap and their definition of the gap is different from the one in Flajolet and Jaillet [9]. The main idea behind improving the dependence on the number of resources is to proceed in two phases: (i) identify the set of arms and binding resources in the optimal solution; (ii) in each round, solve an adaptive, optimistic version of the LP relaxation and sample an arm from the resulting probability distribution. Finally, Sankararaman and Slivkins [15] show a logarithmic regret bound for BwK against a *fixed-distribution* benchmark. However, the regret of this benchmark itself with the optimal MDP policy can be as large as $O(\sqrt{T})$ [9, 14].

2 Preliminaries

2.1 Model

Let T denote a finite time horizon, $\mathcal{X} = \{1, \ldots, k\}$ a set of k arms, $\mathcal{J} = \{1, \ldots, m\}$ denote a set of m resources, and $B_{0,j} = B$ denote the initial budget of resource j. In each round $t \in [T]$, if the budget of any resource is less than 1, then $\mathcal{X}_t = \{1\}$. Otherwise, $\mathcal{X}_t = \mathcal{X}$. The algorithm chooses an arm $x_t \in \mathcal{X}_t$ and observes an outcome $o_t = (r_t, d_{t,1}, \ldots, d_{t,m}) \in [0, 1] \times [-1, 1]^m$. The algorithm earns reward r_t and the budget of resource $j \in \mathcal{J}$ is incremented by drift $d_{t,j}$ as $B_{t,j} = B_{t-1,j} + d_{t,j}$.

Each arm $x \in \mathcal{X}$ has an outcome distribution over $[0, 1] \times [-1, 1]^m$ and o_t is drawn from the outcome distribution of the arm x_t. We use $\mu_x^o = (\mu_x^r, \mu_x^{d,1}, \ldots, \mu_x^{d,m})$ to denote the expected outcome vector of arm x consisting of the expected reward and the expected drifts for each of the m resources. We also use $\mu^{d,j} = (\mu_x^{d,j} : x \in \mathcal{X})$ to denote the vector of expected drifts for resource j. We assume that arm $x^0 = 1 \in \mathcal{X}$ is a null arm with three important properties: (i) its reward is zero a.s.; (ii) the drift for each resource is nonnegative a.s.; and (iii) the expected drift for each resource is positive. The second and third properties of the null arm plus the model's requirement that $x_t = 1$ if $\exists j$ s.t. $B_{t-1,j} < 1$ ensure that the budgets are nonnegative a.s. and can be safely increased from 0.

Our model is intended to capture applications featuring resource renewal, such as the following. In each round, each resource gets replenished by some random amount independent of the past and the chosen arm consumes some random amount of each resource. If the consumption is less than replenishment, the resource gets renewed. The random variable $d_{t,j}$ then models the net replenishment minus consumption. The full model presented above is more general because it allows both the consumption and replenishment to depend on the arm pulled.

We consider two settings in this paper.

MDP setting The decision-maker *knows* the true outcome distributions. In this setting the model implicitly defines an MDP, where the state is the budget vector, the actions are arms, and the transition probabilities are defined by the outcome distributions of the arms.

Learning setting The decision-maker *does now know* the true outcome distributions.

The goal is to design to an MDP policy for the first setting and a learning algorithm for the second, and bound their regret against an LP relaxation as defined in the next subsection.

2.2 Linear Programming Relaxation

Similar to Badanidiyuru et al. [6, Lemma 3.1], we consider the following LP relaxation that provides an upper bound on the total expected reward of any algorithm:

$$\mathsf{OPT}_{\mathsf{LP}} = \max_p \left\{ \sum_{x \in \mathcal{X}} p_x \mu_x^r : \sum_{x \in \mathcal{X}} p_x \mu_x^{d,j} \geq -B/T \; \forall j \in \mathcal{J}, \; \sum_{x \in \mathcal{X}} p_x = 1, \; p_x \geq 0 \; \forall x \in \mathcal{X} \right\}. \quad (1)$$

Lemma 2.1. *The total expected reward of any algorithm is at most* $T \cdot \mathsf{OPT}_{\mathsf{LP}}$.

The proof of this lemma, similar to those in existing works [2, 6], follows from the observations that (i) the variables $p = \{p_x : x \in \mathcal{X}\}$ can be interpreted as the probability of choosing arm x in a round; and (ii) if we set p_x equal to the expected number of times x is chosen by an algorithm divided by T, then it is a feasible solution for the LP.

Definition 2.1 (Regret)**.** The regret of an algorithm $\mathcal{A}$ is defined as $R_T(\mathcal{A}) = T \cdot \mathsf{OPT}_{\mathsf{LP}} - \mathsf{REW}(\mathcal{A})$, where $\mathsf{REW}(\mathcal{A})$ denotes the total expected reward of $\mathcal{A}$.

2.3 Assumptions

We assume that the initial budget of every resource is $B \leq T$. This assumption is without loss of generality because otherwise we can scale the drifts by dividing them by the smallest budget. This results in a smaller support set for the drift distribution that is still contained in $[-1, 1]$.

Our assumptions about the null arm x^0 are a major difference between our model and BwK. In BwK the budgets can only decrease and the process ends when the budget of any resource reaches 0. However, in our model the budgets can increase or decrease, and the process ends at the end of the time horizon. Our assumptions about the null arm allow us to increase the budget from 0 without making it negative. A side-effect of this is that in our model we can even assume that B is a small constant because we can always increase the budget by pulling the null arm, in contrast to existing literature on BwK that assume the initial budgets are large and often scale with the time horizon.

A standard assumption for achieving logarithmic regret in stochastic bandits is that the gap between the total expected reward of an optimal arm and that of a second-best arm is positive. There are a few different ways in which one could translate this to our model where the optimal solution is a mixture over arms. We make the following choice. We assume that there exists a unique set of arms X^* that form the support set of the LP solution and a unique set of resources J^* that correspond to binding constraints in the LP solution [14]. We define the gap of the problem instance in Definition 4.1 and our uniqueness assumption[1] implies that the gap is strictly positive.

We make a few separation assumptions parameterized by four positive constants that can be arbitrarily small. First, the smallest magnitude of the drifts, $\delta_{\text{drift}} = \min\{|\mu_x^{d,j}| : x \in \mathcal{X}, j \in \mathcal{J}\}$, satisfies $\delta_{\text{drift}} > 0$. Second, the smallest singular value of the LP constraint matrix, denoted by $\sigma_{\min}$, satisfies $0 < \sigma_{\min} < 1$. Third, the LP solution p^* satisfies $p_x^* \geq \delta_{\text{support}} > 0$ for all $x \in X^*$. Fourth,

[1]This assumption is essentially without loss of generality because the set of problem instances with multiple optimal solutions is a set of measure zero.

$\sum_{x \in X^*} p_x^* \mu_x^{d,j} \geq \delta_{\text{slack}} > 0$ for all resources $j \notin J^*$. The first assumption is necessary for logarithmic regret bounds because otherwise one can show that the regret of the optimal algorithm for the case of one resoure and one zero-drift arm is $\Theta(\sqrt{T})$ (Appendix A). The second and third assumptions are essentially the same as in existing literature on logarithmic regret bounds for BwK [9, 14]. The fourth assumption allows us to design algorithms that can increase the budgets of the non-binding resources away from 0, thereby reducing the number of times the algorithm has to pull the null arm. Otherwise, if they have zero drift, then, as stated above, the regret of the optimal algorithm for the case of one resource and one zero-drift arm is $\Theta(\sqrt{T})$ (Appendix A).

3 MDP Policy with Constant Regret

In this section we design an MDP policy, `ControlBudget` (Algorithm 2), with constant regret in terms of T for the setting when the learner knows the true outcome distributions and our model implicitly defines an MDP (Section 2.1). At a high level, `ControlBudget`, which shares similarities with Flajolet and Jaillet [9, Algorithm UCB-Simplex], plays arms to keep the budgets close to a decreasing sequence of thresholds. The choice of this sequence allows us to show that the expected leftover budgets and the expected number of null arm pulls are constants. This is a key step in proving the final regret bound. We start by considering the special case of one resource in Section 3.1 because it provides intuition for the general case of multiple resources in Section 3.2.

3.1 Special Case: One Resource

Since there is only one resource we drop the superscript j in this section. We say that an arm x is a positive (resp. negative) drift arm if $\mu_x^d > 0$ (resp. $\mu_x^d < 0$). The following lemma characterizes the possible solutions of the LP (Eq. (1)).

Lemma 3.1. *The solution of the LP relaxation (Eq.* (1)*) is supported on at most two arms. Furthermore, if $T \geq B/\delta_{drift}$, then the solution belongs to one of three categories: (i) supported on a single positive drift arm; (ii) supported on the null arm and a negative drift arm; (iii) supported on a positive drift arm and a negative drift arm.*

The proof of this lemma follows from properties of LPs and a case analysis of which constraints are tight. Our MDP policy, `ControlBudget` (Algorithm 1), deals with the three cases separately and satisfies the following regret bound.[2]

Algorithm 1: `ControlBudget` (for $m = 1$)

Input: time horizon T, initial budget B, set of arms $\mathcal{X}$, set of resources $\mathcal{J}$, constant $c > 0$.

Set $B_0 = B$.
if *LP solution is supported on positive drift arm x^p* **then**
  for $t = 1, 2, \ldots, T$ **do**
    If $B_{t-1} < 1$, pull x^0. Otherwise, pull x^p.
  end
else if *LP solution is supported on null arm x^0 and negative drift arm x^n* **then**
  for $t = 1, 2, \ldots, T$ **do**
    Define threshold $\tau_t = c \log(T - t)$.
    If $B_{t-1} < \max\{1, \tau_t\}$, pull x^0. Otherwise, pull x^n.
  end
else if *LP solution is supported on positive drift arm x^p and negative drift arm x^n* **then**
  for $t = 1, 2, \ldots, T$ **do**
    Define threshold $\tau_t = c \log(T - t)$.
    If $B_{t-1} < 1$, pull x^0. If $1 \leq B_{t-1} < \tau_t$, pull x^p. Otherwise, pull x^n.
  end

[2] In this theorem and the rest of the paper, we use $\tilde{C}$ to denote a constant that depends on problem parameters, including k, m, and the various separation constants mentioned in Section 2.3, but *does not depend on* T. We use this notation because the main focus of this work is how the regret scales as a function of T.

Theorem 3.1. *If $c \geq \frac{6}{\delta^2_{drift}}$, the MDP policy ControlBudget (Algorithm 1) satisfies*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C}, \tag{2}$$

where $\tilde{C} = O\left(\delta_{drift}^{-4} \ln\left(\left(1-\exp\left(-\frac{\delta^2_{drift}}{8}\right)\right)^{-1}\right) + \delta_{drift}^{-1}\left(1-\exp\left(\delta^2_{drift}\right)\right)^{-2}\right)$ is a constant.

We defer all proofs in this section to Appendix B, but we include a proof sketch of most results in the main paper following the statement. The proof of Theorem 3.1 follows from the following sequence of lemmas.

Lemma 3.2. *If the LP solution is supported on a positive drift arm x^p, then*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C}, \tag{3}$$

where $\tilde{C} = O\left(\delta_{drift}^{-3} \ln\left(\left(1-\exp\left(-\frac{\delta^2_{drift}}{8}\right)\right)^{-1}\right)\right)$ is a constant.

We can write the regret in terms of the norm of $\xi = (\xi_{x^p})$, where ξ_{x^p} is the expected difference between the number of times x^p is played by the LP and by `ControlBudget`. This is equal to the expected number of times the policy plays the null arm and, in turn, is equal to the expected number of rounds in which the budget is below 1. Since both x^0 and x^p have positive drift, this is a transient random walk that drifts away from 0. It is known that such a walk spends a constant number of rounds in any state in expectation.

Lemma 3.3. *If the LP solution is supported on the null arm x^0 and a negative drift arm x^n, then*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C} \cdot \mathbb{E}[B_T], \tag{4}$$

where $\tilde{C} = O(\delta_{drift}^{-1})$ is a constant.

We can write the regret in terms of the norm of $\xi = (\xi_{x^0}, \xi_{x^n})$, where ξ_x is the expected difference between the number of times x is played by the LP and by `ControlBudget`. Since both constraints (resource and sum-to-one) are tight, the lemma follows by writing $\xi = D^{-1}b$ and taking norms, where D is the LP constraint matrix and $b = (-\mathbb{E}[B_T], 0)$.

Lemma 3.4. *If the LP solution is supported on a positive drift arm x^p and a negative drift arm x^n, then*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C} \cdot \max\{\mathbb{E}[B_T], \mathbb{E}[N_{x^0}]\}, \tag{5}$$

where $\mathbb{E}[N_{x^0}]$ denotes the expected number null arm pulls and $\tilde{C} = O(\delta_{drift}^{-1})$ is a constant.

This lemma follows similarly to the previous one by writing regret in terms of the norm of $\xi = (\xi_{x^p}, \xi_{x^n})$ and writing $\xi = D^{-1}b$ for $b = (-\mathbb{E}[B_T], \mathbb{E}[N_{x^0}])$.

Therefore, proving that $R_T(\texttt{ControlBudget})$ is a constant in T requires proving that both the expected leftover budget and expected number of null arm pulls are constants. Intuitively, we could ensure $\mathbb{E}[B_T]$ is small by playing the negative drift arm whenever the budget is at least 1. However, there is constant probability of the budget decreasing below 1 and the expected number of null arm pulls becomes $O(T)$. `ControlBudget` solves the tension between the two objectives by carefully choosing a decreasing sequence of thresholds τ_t. The threshold is initially far from 0 to ensure low probability of pre-mature resource depletion, but decreases to 0 over time to ensure small expected leftover budget and decreases at a rate that ensures the expected number of null arm pulls is a constant.

Lemma 3.5. *If the LP solution is supported on a positive drift arm x^p and a negative drift arm x^n, and $c \geq \frac{6}{\delta^2_{drift}}$, then*

$$\mathbb{E}[N_{x^0}] \leq \tilde{C}, \tag{6}$$

where $\tilde{C} = O\left(\delta_{drift}^{-3} \ln\left(\left((1-\exp\left(-\frac{\delta^2_{drift}}{8}\right)\right)^{-1}\right)\right)$ is a constant.

If the budget is below the threshold, i.e., $B_{t-1} < \tau_t$ for some t, then `ControlBudget` pulls x^p until $B_s \geq \tau_{s+1}$ for some $s \geq t$. Since x^p has positive drift, the event that repeated pulls decrease the budget towards 0 is a low probability event. Using this, our choice of $\tau_t = c\log(T-t)$ for an appropriate constant c, and summing over all rounds shows that the expected number of rounds in which the budget is less than 1 is a constant in T.

Lemma 3.6. *If the LP solution is supported on two arms, and $c \geq \frac{6}{\delta^2_{drift}}$, then*

$$\mathbb{E}[B_T] \leq \tilde{C}, \tag{7}$$

where $\tilde{C} = \tilde{O}\left(\left(1 - \exp\left(\delta^2_{drift}\right)\right)^{-2} + \delta^{-2}_{drift}\right)$ is a constant.

If $B_{t-1} \geq \tau_t$, then `ControlBudget` pulls a negative drift arm x^n. We can upper bound the expected leftover budget by conditioning on q, the number of consecutive pulls of x^n at the end of the timeline. The main idea in completing the proof is that (i) if q is large, then it corresponds to a low probability event; and (ii) if q is small, then the budget in round $T - q$ was smaller than τ_q, which is a decreasing sequence in q, and there are few rounds left so the budget cannot increase by too much.

3.2 General Case: Multiple Resources

Now we use the ideas from Section 3.1 to tackle the case of $m > 1$ resources that is much more challenging. Generalizing Lemma 3.1, the solution of the LP relaxation (Eq. (1)) is supported on at most $\min\{k, m\}$ arms. Informally, our MDP policy, `ControlBudget` (Algorithm 2), samples an arm from a probability distribution that ensures drifts bounded away from 0 in the "correct directions": (i) a binding resource j has drift at least γ_t if $B_{t-1,j} < \tau_t$ and drift at most $-\gamma_t$ if $B_{t-1,j} \geq \tau_t$; and (ii) a non-binding resource j has drift at least $\frac{1}{2}\gamma_t$ if $B_{t-1,j} < \tau_t$. This allows us to show that the expected leftover budget for each binding resource and the expected number of null arm pulls are constants in terms of T.

Algorithm 2: `ControlBudget` (for general m)

Input: time horizon T, initial budget B, set of arms $\mathcal{X}$, set of resources $\mathcal{J}$, constant $c > 0$.

Set $B_{0,j} = B$ for all $j \in \mathcal{J}$.
Define threshold $\tau_t = c \log(T - t)$.
for $t = 1, 2, \ldots, T$ **do**
  if $\exists j \in \mathcal{J}$ *such that* $B_{t-1,j} < 1$ **then**
    Pull the null arm x^0.
  else
    Define $s_t \in \{\pm 1\}^{|X^*|-1} \times 0$ as follows. Let j denote the resource corresponding to row $i \in [|X^*| - 1]$ in the matrix D and vector b. Then, the ith entry of s_t is -1 if $B_{t-1,j} < \tau_t$ and $+1$ otherwise.
    Define γ_t to be the solution to the following constrained optimization problem:

$$\max_{\gamma \in [0,1]} \left\{\gamma : p = D^{-1}(b + \gamma s_t) \geq 0,\ p^T \mu^{d,j} \geq \frac{\gamma}{2}\ \forall j \in \mathcal{J} \setminus J^* \text{ if } B_{t-1,j} < \tau_t\right\}. \tag{8}$$

    Sample an arm from the probability distribution $p_t = D^{-1}(b + \gamma_t s_t)$.
  end
end

Theorem 3.2. *If $c \geq \frac{6}{\gamma^{*2}}$, the regret of* `ControlBudget` *(Algorithm 2) satisfies*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C}, \tag{9}$$

where γ^* (defined in Lemma 3.9) and $\tilde{C}$ are constants with $\tilde{C} = O\left(m\sigma^{-1}_{\min}\left(m(\gamma^*)^{-3}\ln\left(\left((1 - \exp\left(-\gamma^{*2}\right)\right)^{-1}\right) + \left(1 - \exp(\gamma^{*2})\right)^{-2}\right)\right)$.

We defer all proofs in this section to Appendix C, but we include a proof sketch of most results in the main paper following the statement. The proof of Theorem 3.2 follows from the following sequence of lemmas. The next two lemmas are generalizations of Lemmas 3.3 and 3.4 with essentially the same proofs. Recall J^* denotes the unique set of resources that correspond to binding constraints in the LP solution (Section 2.3).

Lemma 3.7. *If the LP solution includes the null arm x^0 in its support, then*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C} \cdot \left(\sum_{j \in J^*} \mathbb{E}[B_{T,j}]\right), \tag{10}$$

where $\tilde{C} = O(\sigma_{\min}^{-1})$ is a constant.

Lemma 3.8. *If the LP solution does not include the null arm x^0 in its support, then*

$$R_T(\texttt{ControlBudget}) \leq \tilde{C} \cdot \left(\sum_{j \in J^*} \mathbb{E}[B_{T,j}] + \mathbb{E}[N_{x^0}] \right), \tag{11}$$

where $\mathbb{E}[N_{x^0}]$ denotes the expected number of null arm pulls and $\tilde{C} = O(\sigma_{\min}^{-1})$ is a constant.

Lemmas 3.10 and 3.11 are generalizations of Lemmas 3.5 and 3.6 with similar proofs after taking a union bound over resources. But we first need Lemma 3.9 that lets us conclude there is drift of magnitude at least $\gamma^* > 0$ in the "correct directions" as stated earlier.

Lemma 3.9. *[9, Lemma 14] In each round t, $\gamma_t \geq \gamma^* = \frac{\sigma_{\min} \min\{\delta_{support}, \delta_{slack}\}}{4m}$.*

The proof of this lemma is identical to Flajolet and Jaillet [9, Lemma 14] but we provide a proof in the appendix for completeness.

Lemma 3.10. *If the LP solution does not include the null arm in its support, then*

$$\mathbb{E}[N_{x^0}] \leq \tilde{C}, \tag{12}$$

where $\tilde{C} = O\left(m(\gamma^*)^{-3} \ln \left(\left((1 - \exp\left(-\frac{\gamma^{*2}}{8} \right) \right)^{-1} \right) \right)$ is a constant.

Lemma 3.11. *If the LP solution is supported on more than one arm, then for all $j \in J^*$*

$$\mathbb{E}[B_{T,j}] \leq \tilde{C}, \tag{13}$$

where $\tilde{C} = \tilde{O}\left(\left(1 - \exp(\gamma^{*2})\right)^{-2} + (\gamma^*)^{-2} \right)$ is a constant.

A subtle but important point is that the regret analysis does not require `ControlBudget` to know the true expected drifts in order to find the probability vector p_t. It simply requires the algorithm to know X^*, J^*, and find any probability vector p_t that ensures drifts bounded away from 0 in the "correct directions" as stated earlier. We use this property in our learning algorithm, `ExploreThenControlBudget` (Algorithm 3), in the next section.

4 Learning Algorithm with Logarithmic Regret

In this section we design a learning algorithm, `ExploreThenControlBudget` (Algorithm 3), with logarithmic regret in terms of T for the setting when the learning does not know the true distributions. Our algorithm, which can be viewed as combining aspects of Li et al. [14, Algorithm 1] and Flajolet and Jaillet [9, Algorithm UCB-Simplex], proceeds in three phases. It uses phase one of Li et al. [14, Algorithm 1] to identify the set of optimal arms X^* and the set of binding constraints J^* by playing arms in a round-robin fashion, and using confidence intervals and properties of LPs. This is reminiscent of successive elimination [8], except that the algorithm tries to identify the optimal arms instead of eliminating suboptimal ones. In the second phase the algorithm continues playing the arms in X^* in a round-robin fashion to shrink the confidence radius further. In the third phase the algorithm plays a variant of the MDP policy `ControlBudget` (Algorithm 2) with a slighly different optimization problem for γ_t because it only has empirical esitmates of the drifts.

4.1 Additional Notation and Preliminaries

For all arms $x \in \mathcal{X}$ and rounds $t \geq k$, define the upper confidence bound (UCB) of the expected outcome vector μ_x^o as $\mathsf{UCB}_t(x) = \bar{o}_t(x) + \mathsf{rad}_t(x)$, where $\mathsf{rad}_t(x) = \sqrt{8 n_t(x)^{-1} \log T}$ denotes the confidence radius, $n_t(x)$ denotes the number of times x has been played before t, and $\bar{o}_t(x) = n_t(x)^{-1} \sum_t o_t \mathbb{1}[x_t = x]$ denotes the empirical mean outcome vector of x. The lower confidence bound (LCB) is defined similarly as $\mathsf{LCB}_t(x) = \bar{o}_t(x) - \mathsf{rad}_t(x)$.

For all arms $x \in \mathcal{X}$, let OPT_{-x} denote the value of the LP relaxation (Eq. (1)) with the additional constraint $p_x = 0$, and for all resources $j \in \mathcal{J}$, let OPT_{-j} denote the value when the objective has an extra $-\sum_x p_x \mu_x^{d,j} + B/T$ term [14]. Intuitively, these represent how important it is to play arm x or make the resource constraint for j a binding constraint. Define the UCB of OPT_{-x} to be the value of the LP when the expected outcome is replaced by its UCB, and denote this by $\mathsf{UCB}_t(\mathsf{OPT}_{-x})$. The LCB for OPT_{-x}, and UCB and LCB for OPT_{-j} and $\mathsf{OPT}_{\mathsf{LP}}$ are defined similarly.

Definition 4.1 (Gap [14])**.** The gap of the problem instance is defined as

$$\Delta = \min\left\{\min_{x\in X^*}\left\{\mathsf{OPT}_{\mathsf{LP}} - \mathsf{OPT}_{-x}\right\}, \min_{j\notin J^*}\left\{\mathsf{OPT}_{\mathsf{LP}} - \mathsf{OPT}_{-j}\right\}\right\}. \tag{14}$$

4.2 Learning Algorithm and Regret Analysis

Algorithm 3: `ExploreThenControlBudget`

Input: time horizon T, initial budget B, set of arms $\mathcal{X}$, set of resources $\mathcal{J}$, constant $c > 0$.

Set $B_{0,j} = B$ for all $j \in \mathcal{J}$.
Initialize $t = 1, X^* = \emptyset, J' = \emptyset$.
while $t < T - k$ *and* $|X^*| + |J'| < m$ **do**
  Play each arm in $\mathcal{X} \setminus \{x^0\}$ in a round-robin fashion. Play x^0 if $\exists j$ such that $B_{t-1,j} < 1$.
  For each $x \in \mathcal{X}$, if $\mathsf{UCB}_t(\mathsf{OPT}_{-x}) < \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}})$, then add x to X^*.
  For each $j \in \mathcal{J}$, if $\mathsf{UCB}_t(\mathsf{OPT}_{-j}) < \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}})$, then add j to J'.
end
Set $J^* = \mathcal{J} \setminus J'$.
while $t < T - |X^*|$ *and* $n_t(x) < \frac{32 \log T}{\gamma^{*2}}$ *for all* $x \in X^*$*, where* γ^* *is defined in Lemma 3.9* **do**
  Play each arm in X^* in a round-robin fashion.
end
while $t < T$ **do**
  if $\exists j \in \mathcal{J}$ *such that* $B_{t-1,j} < 1$ **then**
    Pull the null arm x^0.
  else
    Define s_t as in Algorithm 2.
    Choose $(\gamma_t, p_t) = \max_{\gamma\in[0,1]} \gamma$ such that there exists a probability vector p satisfying

$$\mathsf{LCB}_t(p^T\mu^{d,j}) \geq \frac{\gamma}{8}\ \forall j \in \mathcal{J} \setminus J^* \text{ if } B_{t-1,j} < \tau_t, \tag{15}$$

$$\mathsf{LCB}_t(p^T\mu^{d,j}) \geq \frac{\gamma}{8}\ \forall j \in J^* \text{ if } B_{t-1,j} < \tau_t \tag{16}$$

$$\mathsf{UCB}_t(p^T\mu^{d,j}) \leq -\frac{\gamma}{8}\ \forall j \in J^* \text{ if } B_{t-1,j} \geq \tau_t. \tag{17}$$

    Sample an arm from the probability distribution p_t.
  end
end

Theorem 4.1. *If $c \geq \frac{6}{\gamma^*}$, the regret of* `ExploreThenControlBudget` *(Algorithm 3) satisfies*

$$R_T(\texttt{ExploreThenControlBudget}) \leq \tilde{C} \cdot \log T, \tag{18}$$

where γ^* (defined in Lemma 3.9), $\tilde{C}'$ and $\tilde{C}$ are constants with $\tilde{C}'$ denoting the constant in Theorem 3.2 and

$$\tilde{C} = O\left(\frac{km^2}{\min\{\delta_{drift}^2, \sigma_{\min}^2\}\Delta^2} + k(\gamma^*)^{-2} + \tilde{C}'\right). \tag{19}$$

We defer all proofs in this section to Appendix D, but we include a proof sketch of most results in the main paper following the statement. We refer to the three while loops of `ExploreThenControlBudget` (Algorithm 3) as the three phases. The lemmas and corollaries following the theorem below show that the first phase consists of at most logarithmic number of rounds. It is easy to see that the second phase consists of at most logarithmic number of rounds. The third phase plays a variant of the MDP policy `ControlBudget` (Algorithm 2). There exists a feasible solution to the optimization problem in Line 17 that ensures drifts bounded away from 0 in the "correct directions" (Appendix D.3). Combining the analysis of the first two phases with Theorem 3.2 lets us conclude that `ExploreThenControlBudget` has logarithmic regret.

Definition 4.2 (Clean Event)**.** The clean event is the event such that for all $x \in \mathcal{X}$ and all $t \geq k$, (i) $\mu_x^0 \in [\mathsf{LCB}_t(x), \mathsf{UCB}_t(x)]$; and (ii) after the first n pulls of the null arm the sum of the drifts for each resource is at least w, where $w = \frac{4096 k m^2 \log T}{\delta_{\text{drift}}^2 \Delta^2}$ and $n = \frac{2w}{\mu_{x^0}^d}$.

Lemma 4.1. *The clean event occurs with probability at least* $1 - 5mT^{-2}$.

The lemma follows from Hoeffding's inequality. Since the complement of the clean event contributes $O(mT^{-1})$ to the regret, it suffices to bound the regret conditioned on the clean event.

Lemma 4.2. *If the clean event occurs, then* $\mathsf{UCB}_t(\mathsf{OPT}_{\mathsf{LP}}) - \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}}) \leq \frac{8m}{\sigma_{\min}} \mathsf{rad}_t$. *A similar statement is true for* OPT_{-x} *and* OPT_{-j} *for all* $x \in \mathcal{X}$ *and* $j \in \mathcal{J}$.

The proof follows from a perturbation analysis of the LP and uses the confidence radius to bound the perturbations in the rewards and drifts.

Corollary 4.1. *If the clean event occurs and* $n_t(x) > \frac{2048 m^2 \log T}{\sigma_{\min}^2 \Delta^2}$ *for all* $x \in \mathcal{X}$*, then*

$$\mathsf{UCB}_t(\mathsf{OPT}_{-x}) < \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}}) \text{ and } \mathsf{UCB}_t(\mathsf{OPT}_{-j}) < \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}}) \tag{20}$$

for all $x \in X^*$ *and* $j \in \mathcal{J} \setminus J^*$.

This follows from substituting the bound on $n_t(x)$ into the definition of $\mathsf{rad}_t(x)$ and applying Lemma 4.2. In the worst case, each pull of an arm can cause the budget to drop below 1, but the clean event implies that the first n pulls of x^0 have enough total drift to allow $n_t(x)$ pulls of each non-null arm in phase 1. This allows us to upper bound the duration of phase 1 as follows.

Corollary 4.2. *If the clean event occurs, then phase 1 of* `ExploreThenControlBudget` *has at most*

$$\tilde{C} \cdot \log T \tag{21}$$

rounds, where $\tilde{C} = O\left(\frac{km^2}{\min\{\delta_{\textit{drift}}^2, \sigma_{\min}^2\}\Delta^2}\right)$ *is a constant.*

4.3 Reduction from BwK

Theorem 4.2. *Suppose* $\frac{B}{T} \geq \delta_{\textit{drift}}$. *Consider a* BwK *instance such that (i) for each arm and resource, the expected consumption of that resource differs from* $\frac{B}{T}$ *by at least* $\delta_{\textit{drift}}$*; and (ii) all the other assumptions required by Theorem 4.1 (Section 2.3) are also satisfied. Then, there is an algorithm for* BwK *whose regret satisfies the same bound as in Theorem 4.1 with the same constant* $\tilde{C}$.

Due to space constraints, we present the reduction in Appendix D.4.

5 Conclusion

In this paper we introduced a natural generalization of BwK that allows non-monotonic resource utilization. We first considered the setting when the decision-maker knows the true distributions and presented an MDP policy with *constant* regret against an LP relaxation. Then we considered the setting when the decision-maker does not know the true distributions and presented a learning algorithm with *logarithmic* regret against the same LP relaxation. Finally, we also presented a reduction from BwK to our model and showed a regret bound that matches existing results [14].

An important direction for future research is to obtain optimal regret bounds. The regret bound for our algorithm scales as $O(\text{poly}(k)\text{poly}(m)\Delta^{-2})$, where k is the number of arms, m is the number of resources and Δ is the suboptimality parameter. A modification to our algorithm along the lines of Flajolet and Jaillet [9, algorithm UCB-Simplex] that considers each support set of the LP solution explicitly leads to a regret bound that scales as $O(\text{poly}(k)2^m\Delta^{-1})$. It is an open question, even for BwK, to obtain a regret bound that scales as $O(\text{poly}(k)\text{poly}(m)\Delta^{-1})$ or show that the trade-off between the dependence on the number of resources and the suboptimality parameter is unavoidable.

Another natural follow-up to our work is to develop further extensions, such as considering an infinte set of arms [11], studying adversarial observations [4, 10], or incorporating contextual information [16, 1] as has been the case elsewhere throughout the literature on bandits.

Acknowledgments and Disclosure of Funding

We thank Sloan Nietert and Ayush Sekhari for helpful discussions. This research was partially supported by the NSERC Postgraduate Scholarships-Doctoral Fellowship 545847-2020.

References

[1] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

[2] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[4] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

[6] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[7] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012.

[8] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

[9] Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800v4*, 2015.

[10] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219. IEEE, 2019.

[11] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

[12] Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[13] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[14] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, pages 6483–6492. PMLR, 2021.

[15] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34, 2021.

[16] Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

[17] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12 (1-2):1–286, 2019.

[18] Long Tran-Thanh, Archie Chapman, Enrique Munoz De Cote, Alex Rogers, and Nicholas R Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[19] Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1134–1140, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss potential improvements in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our paper studies a theoretical model and we do not believe there are any negative societal impacts of this paper.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state our assumptions in Section 2.3.
   (b) Did you include complete proofs of all theoretical results? [Yes] We include complete proofs in the appendix. But we include a proof sketch of most lemmas in the main paper following the lemma statement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Regret Bounds for One Arm, One Resource, and Zero Drift

In this section we will consider the case when $\mathcal{X} = \{x^0, x\}$, $\mathcal{J} = \{1\}$, and x has zero drift, i.e., $\mu_x^d = 0$. Since x is the only arm besides the null arm, we assume without loss of generality that its reward is equal to 1 deterministically. The optimal policy is to pull x^0 when $B_{t-1} < 1$ and x otherwise. We will show that the regret of this policy is $\Theta(\sqrt{T})$.

Theorem A.1. *The regret of the MDP policy is $O(\sqrt{T})$.*

Proof. The optimal solution of the LP relaxation (Eq. (1)) is $p_x = 1$ and $p_{x^0} = 0$. Since x^0 and x have reward equal to 0 and 1 deterministically, $\mathsf{OPT}_{\mathsf{LP}} = 1$.Therefore, the regret of the MDP policy is equal to the expected number rounds in which the budget is less than 1. That is,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right].$$

Since $\mathbb{E}[d_t] = 0$ when $B_{t-1} \geq 1$ and $\mathbb{E}[d_t] = \mu_{x^0}^d$ when $B_{t-1} < 1$, we can write

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] \\ &= \frac{1}{\mu_{x^0}^d}\mu_{x^0}^d \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] \\ &= \frac{1}{\mu_{x^0}^d}\mathbb{E}\left[\sum_{t=1}^{T} \mu_{x^0}^d \mathbb{1}[B_{t-1} < 1] + 0\mathbb{1}[B_{t-1} \geq 1]\right] \\ &= \frac{1}{\mu_{x^0}^d}\mathbb{E}\left[\sum_{t=1}^{T} d_t\right] \\ &= \frac{1}{\mu_{x^0}^d}\left(\mathbb{E}\left[B_T\right] - B\right). \end{aligned}$$

Since $B_0 = B$, the budget is updated as $B_t = B_{t-1} + d_t$ and $d_t \in [-1, 1]$, we have

$$\begin{aligned} \mathbb{E}\left[B_t^2 | B_{t-1}\right] &= \mathbb{E}\left[B_{t-1}^2 + 2B_{t-1}d_t + d_t^2 | B_{t-1}\right] \\ &= \mathbb{E}\left[B_{t-1}^2 | B_{t-1}\right] + \mathbb{E}\left[2B_{t-1}d_t | B_{t-1}\right] + \mathbb{E}\left[d_t^2 | B_{t-1}\right] \\ &\leq B_{t-1}^2 + \mathbb{E}\left[2B_{t-1}d_t | B_{t-1}\right] + 1^2 \\ &= B_{t-1}^2 + 2B_{t-1}\mu_{x^0}^d \mathbb{1}[B_{t-1} < 1] + 1 \\ &\leq B_{t-1}^2 + 2\mu_{x^0}^d + 1 \\ \Rightarrow \mathbb{E}\left[B_T^2\right] &= O(T). \end{aligned}$$

Using Jensen's inequality, we have

$$\mathbb{E}\left[B_T\right] \leq \sqrt{\mathbb{E}\left[B_T^2\right]} = O(\sqrt{T}).$$

This completes the proof. □

Theorem A.2. *If $\mathbb{E}\left[d_t^2 | B_{t-1}\right] \geq \sigma^2 > 0$, then the regret of the MDP policy is $\Omega(\sqrt{T})$.*

Proof. Using the proof of Theorem A.1, it suffices to provide a lower bound on $\mathbb{E}\left[B_T\right]$. Since the budget is updated as $B_t = B_{t-1} + d_t$, $\mathbb{E}\left[d_t | B_{t-1}\right] \geq 0$, and $\mathbb{E}\left[d_t^2 | B_{t-1}\right] \geq \sigma^2$, we have

$$\begin{aligned} \mathbb{E}\left[B_t^2 | B_{t-1}\right] &= \mathbb{E}\left[B_{t-1}^2 + 2B_{t-1}d_t + d_t^2 | B_{t-1}\right] \\ &= \mathbb{E}\left[B_{t-1}^2 | B_{t-1}\right] + \mathbb{E}\left[2B_{t-1}d_t | B_{t-1}\right] + \mathbb{E}\left[d_t^2 | B_{t-1}\right] \\ &\geq B_{t-1}^2 + \mathbb{E}\left[2B_{t-1}d_t | B_{t-1}\right] + \sigma^2 \\ &= B_{t-1}^2 + 2B_{t-1}\mathbb{E}\left[d_t | B_{t-1}\right] + \sigma^2 \\ &\geq B_{t-1}^2 + \sigma^2 \\ \Rightarrow \mathbb{E}\left[B_T^2\right] &\geq \Omega(T). \end{aligned}$$

The Cauchy-Schwarz inequality yields that

$$\mathbb{E}\left[\left(B_T^{1/2}\right)^2\right]^{1/2}\mathbb{E}\left[\left(B_T^{3/2}\right)^2\right]^{1/2} \geq \mathbb{E}\left[B_T^2\right] \geq \Omega(T).$$

Squaring both sides yields that

$$\mathbb{E}\left[B_T\right]\mathbb{E}\left[B_T^3\right] \geq \mathbb{E}\left[B_T^2\right]^2 \geq \Omega(T^2).$$

It suffices to show that $\mathbb{E}\left[B_T^3\right] = O(T^{3/2})$ because this will imply that $\mathbb{E}\left[B_T\right] = \Omega(T^{1/2})$. Since $d_t \in [-1,1]$, we have

$$\begin{aligned}
&\mathbb{E}\left[B_t^3|B_{t-1}\right] \\
&= \mathbb{E}\left[B_{t-1}^3 + 3B_{t-1}^2 d_t + 3B_{t-1}d_t^2 + d_t^3|B_{t-1}\right] \\
&= B_{t-1}^3 + 3B_{t-1}^2\mathbb{E}\left[d_t|B_{t-1}\right] + 3B_{t-1}\mathbb{E}\left[d_t^2|B_{t-1}\right] + \mathbb{E}\left[d_t^3|B_{t-1}\right] \\
&= B_{t-1}^3 + 3B_{t-1}^2\mu_{x^0}^d\mathbb{1}[B_{t-1} < 1] + 3B_{t-1}\mathbb{E}\left[d_t^2|B_{t-1}\right] + \mathbb{E}\left[d_t^3|B_{t-1}\right] \\
&\leq B_{t-1}^3 + 3\mu_{x^0}^d + 3B_{t-1} + 1.
\end{aligned}$$

Taking expectation on both sides yields

$$\begin{aligned}
\mathbb{E}\left[B_t^3\right] &\leq \mathbb{E}\left[B_{t-1}^3\right] + 3\mathbb{E}\left[B_{t-1}\right] + O(1) \\
&\leq \mathbb{E}\left[B_{t-1}^3\right] + O(\sqrt{t}),
\end{aligned}$$

where the last inequality follows from the proof of Theorem A.1 where we show that $\mathbb{E}\left[B_t\right] \leq O(\sqrt{t})$. Summing both sides over all rounds yields that

$$\mathbb{E}\left[B_T^3\right] = O(T^{3/2})$$

and this completes the proof. □

B Proofs for Section 3.1

B.1 Proof of Lemma 3.2

When the LP solution is supported on a positive drift arm x^p, $\mathsf{OPT}_{\mathsf{LP}} = 1$ because the LP plays it with probability 1. Therefore, the regret is equal to the expected number of times `ControlBudget` (Algorithm 1) pulls the null arm. This, in turn, is equal to the expected number of rounds in which the budget is less than 1.

Define

$$b_0 = 8\delta_{\text{drift}}^{-2}\ln\left(\frac{2}{1-\exp\left(-\frac{\delta_{\text{drift}}^2}{8}\right)}\right). \tag{22}$$

Then, we have that for all $b \geq b_0$,

$$\begin{aligned}
\sum_{k=b}^{\infty}\mathbf{Pr}\left[B_{s+k}\in[0,1)|B_s = b\right] &\leq \sum_{k=b}^{\infty}\exp\left(-\frac{\delta_{\text{drift}}^2 k}{8}\right) && (23)\\
&= \exp\left(-\frac{\delta_{\text{drift}}^2 b}{8}\right)\left(1-\exp\left(-\frac{\delta_{\text{drift}}^2}{8}\right)\right)^{-1}. && (24)
\end{aligned}$$

where the first inequality follows from Azuma-Hoeffding's inequality. By our choice of b_0, we have that

$$\sum_{k=b}^{\infty}\mathbf{Pr}\left[B_{s+k}\in[0,1)|B_s = b\right] \leq \frac{1}{2}. \tag{25}$$

In words, the probability that the budget ever drops below 1 once it exceeds b_0 is at most $\frac{1}{2}$. Now, consider the following recursive definition for two disjoint sequence of indices s_i and s'_i. Let $s_0 = \min\{t \geq 1 : B_{t-1} \in [0,1)\}$, and define

$$s'_i = \min\{t > s_i : B_{t-1} \geq b_0 \text{ or } t-1 = T\} \tag{26}$$

$$s_{i+1} = \min\{t > s'_i : B_{t-1} \in [0,1)\}. \tag{27}$$

In words, s_i' denotes the first round after s_i in which the budget is at least b_0 and s_{i+1} denotes the first round after s_i' in which the budget is less than 1. Note that Eq. (25) implies that

$$\mathbf{Pr}\left[s_i \text{ is defined } | s_{i-1}' \text{ is defined}\right] \leq \frac{1}{2}. \tag{28}$$

Therefore,

$$\mathbf{Pr}\left[s_i \text{ is defined}\right] \leq \prod_{j=1}^{i} \mathbf{Pr}\left[s_j \text{ is defined } | s_{j-1}' \text{ is defined}\right] \leq \frac{1}{2^i}. \tag{29}$$

Now, we can upper bound the expected number of rounds in which the budget is below 1 as

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] &= \sum_{i=0}^{T-1} \mathbf{Pr}\left[s_i \text{ is defined}\right] \mathbb{E}\left[\sum_{t=s_i}^{s_i'} \mathbb{1}[B_{t-1} < 1]\right] && (30)\\
&\leq \sum_{t=0}^{T-1} 2^{-i} \mathbb{E}\left[\sum_{t=s_i}^{s_i'} \mathbb{1}[B_{t-1} < 1]\right] && (31)\\
&\leq \sum_{t=0}^{T-1} 2^{-i} \mathbb{E}\left[s_i' - s_i\right] && (32)\\
&\leq \sum_{t=0}^{T-1} 2^{-i} \frac{1}{\delta_{\text{drift}}} \mathbb{E}\left[B_{s_i'} - B_{s_i}\right] && (33)\\
&\leq \sum_{t=0}^{T-1} 2^{-i} \frac{1}{\delta_{\text{drift}}} (b_0 + 1) && (34)\\
&\leq 2 \frac{b_0 + 1}{\delta_{\text{drift}}}, && (35)
\end{aligned}$$

where Eq. (33) follows because both the null arm and the positive drift arm have drift at least δ_{drift}. Therefore, we have that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] \leq \tilde{C}, \tag{36}$$

where

$$\tilde{C} = O\left(\delta_{\text{drift}}^{-3} \ln\left(\frac{2}{1 - \exp\left(-\frac{\delta_{\text{drift}}^2}{8}\right)}\right)\right). \tag{37}$$

B.2 Proof of Lemma 3.3

Let p^* denote the optimal solution to the LP relaxation and note that Tp_x^* denotes the expected number of times the LP plays arm x. Since the LP solution is supported on two arms, both the budget and sum-to-one constraints are tight. Therefore, we have

$$D(Tp^*) = b_{\mathsf{LP}}, \tag{38}$$

where

$$D = \begin{bmatrix} \mu_{x^0}^d & \mu_{x^n}^d \\ 1 & 1 \end{bmatrix},\ p^* = \begin{bmatrix} p_{x^0}^* \\ p_{x^n}^* \end{bmatrix},\ b_{\mathsf{LP}} = \begin{bmatrix} -B \\ T \end{bmatrix}. \tag{39}$$

Let N_x denote the number of times `ControlBudget` (Algorithm 1) plays arm x. Since it plays the null arm x^0 and the negative drift arm x^n, the sum-to-one constraint is tight. However, the budget constraint may not be tight because there may be leftover budget. Therefore, we have

$$DN = b_{\mathsf{LP}} - b, \tag{40}$$

where

$$N = \begin{bmatrix} \mathbb{E}[N_{x^0}] \\ \mathbb{E}[N_{x^n}] \end{bmatrix},\ b = \begin{bmatrix} -E[B_T] \\ 0 \end{bmatrix}. \tag{41}$$

Define

$$\xi = \begin{bmatrix} \xi_{x^0} \\ \xi_{x^n} \end{bmatrix} = \begin{bmatrix} Tp^*_{x^0} - \mathbb{E}[N_{x^0}] \\ Tp^*_{x^n} - \mathbb{E}[N_{x^n}] \end{bmatrix}. \tag{42}$$

Subtracting Eq. (40) from Eq. (38) we have $\xi = D^{-1}b$, where the LP constraint matrix D is invertible by our assumption that the drifts are nonzero. Finally, letting μ^r denote the vector of expected rewards, the regret can be expressed as

$$\begin{aligned} R_T(\texttt{ControlBudget}) &= \xi^T \mu^r &(43)\\ &\leq |\xi^T \mu^r| &(44)\\ &\leq \|\xi\|_1 \|\mu^r\|_\infty &(45)\\ &\leq \|D^{-1}\|_1 \|b\|_1 &(46)\\ &\leq C_{\delta_{\text{drift}}} \mathbb{E}[B_T], &(47) \end{aligned}$$

where $C_{\delta_{\text{drift}}} = O(\delta_{\text{drift}}^{-1})$ is a constant. This completes the proof.

B.3 Proof of Lemma 3.4

Let p^* denote the optimal solution to the LP relaxation and note that Tp^*_x denotes the expected number of times the LP plays arm x. Since the LP solution is supported on two arms, both the budget and sum-to-one constraints are tight. Therefore, we have

$$D(Tp^*) = b_{\mathsf{LP}}, \tag{48}$$

where

$$D = \begin{bmatrix} \mu^d_{x^p} & \mu^d_{x^n} \\ 1 & 1 \end{bmatrix}, \; p^* = \begin{bmatrix} p^*_{x^p} \\ p^*_{x^n} \end{bmatrix}, \; b_{\mathsf{LP}} = \begin{bmatrix} -B \\ T \end{bmatrix}. \tag{49}$$

Let N_x denote the number of times `ControlBudget` (Algorithm 1) plays arm x. Since it plays the null arm x^0 when the budget is less than 1 and may have leftover budget, neither the budget nor the sum-to-one constraint are tight. Therefore, we have

$$DN = b_{\mathsf{LP}} - b, \tag{50}$$

where

$$N = \begin{bmatrix} \mathbb{E}[N_{x^p}] \\ \mathbb{E}[N_{x^n}] \end{bmatrix}, \; b = \begin{bmatrix} -E[B_T] \\ \mathbb{E}[N_{x^0}] \end{bmatrix}. \tag{51}$$

Define

$$\xi = \begin{bmatrix} \xi_{x^p} \\ \xi_{x^n} \end{bmatrix} = \begin{bmatrix} Tp^*_{x^p} - \mathbb{E}[N_{x^p}] \\ Tp^*_{x^n} - \mathbb{E}[N_{x^n}] \end{bmatrix}. \tag{52}$$

Subtracting Eq. (50) from Eq. (48) we have $\xi = D^{-1}b$, where the LP constraint matrix D is invertible by our assumption that the drifts are nonzero. Finally, letting μ^r denote the vector of expected rewards, the regret can be expressed as

$$\begin{aligned} R_T(\texttt{ControlBudget}) &= \xi^T \mu^r &(53)\\ &\leq |\xi^T \mu^r| &(54)\\ &\leq \|\xi\|_1 \|\mu^r\|_\infty &(55)\\ &\leq \|D^{-1}\|_1 \|b\|_1 &(56)\\ &\leq C_{\delta_{\text{drift}}} \left(\mathbb{E}[B_T] + \mathbb{E}[N_{x^0}]\right), &(57) \end{aligned}$$

where $C_{\delta_{\text{drift}}} = O(\delta_{\text{drift}}^{-1})$ is a constant. This completes the proof.

B.4 Proof of Lemma 3.5

Divide the T rounds into two phases: $P_1 = \{1, \ldots, T - \exp(3/c)\}$ and $P_2 = \{1, \ldots, T\} \setminus P_1$. Note that P_2 consists of $\exp(3/c) = O(\exp(\delta_{\text{drift}})) = O(1)$ rounds, where the last equality follows because drifts are bounded by 1. Therefore, the expected number of null arm pulls in this phase is $O(1)$ and it suffices to bound the expected number of null arm pulls in P_1.

Consider the following recursive definition for three disjoint sequences of indices t_i, t_i' and t_i''. Let $t_0 = 0$, and define

$$t_i' = \min\{t > t_i : B_{t-1} \geq \tau_t \text{ or } t-1 = T\}, \tag{58}$$

$$t_i'' = \min\{t > t_i' : B_{t-1} < \tau_t\}, \tag{59}$$

$$t_{i+1} = \min\{t > t_i'' : B_t < 1\}. \tag{60}$$

We can bound the expected number of rounds in which the budget is less than 1 as

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] \tag{61}$$

$$= \sum_{i=0}^{T-1} \mathbf{Pr}\left[t_i \text{ exists }\right] \mathbb{E}\left[\sum_{t=t_i}^{t_i'} \mathbb{1}[B_{t-1} < 1]\right] \tag{62}$$

$$\leq \underbrace{\mathbb{E}\left[\sum_{t=t_0}^{t_0'-1} \mathbb{1}[B_{t-1} < 1]\right]}_{(a)} + \sum_{i=0}^{T-1} \mathbf{Pr}\left[t_{i+1} \text{ exists } | t_i', t_i'' \text{ exist}\right] \underbrace{\mathbb{E}\left[\sum_{t=t_i}^{t_i'-1} \mathbb{1}[B_{t-1} < 1]\right]}_{(a)}. \tag{63}$$

In rounds $\{t_i, \ldots, t_i' - 1\}$, the algorithm pulls the null and positive drift arms. The proof of Lemma 3.2 shows that the expected number of null arm pulls in these rounds is at most $\tilde{C}$, where $\tilde{C}$ is defined in Eq. (37). Therefore, we can bound the term (a) in the above inequality by $\tilde{C}$ and we have that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1} < 1]\right] \leq \tilde{C}\left(1 + \sum_{i=0}^{T-1} \mathbf{Pr}\left[t_{i+1} \text{ exists } | t_i', t_i'' \text{ exist}\right]\right). \tag{64}$$

If t_i' exists, then $B_{t_i'-1} \geq \tau_{t_i'}$. If t_i'' exists, then $\tau_{t_i''} - 1 \leq B_{t_i''-1} < \tau_{t_i''}$ because (i) t_i'' is the first round after t_i' in which the budget is below the threshold; and (ii) the drifts are bounded by 1, so it cannot be lower than $\tau_{t_i''} - 1$. The algorithm pulls the negative drift arm x^n in the rounds $\{t_i', \ldots, t_i'' - 1\}$ and the positive drift arm x^p in the rounds $\{t_i'', \ldots, t_{i+1} - 1\}$. Since the drifts are bounded by 1, it takes at least $\tau_{t_i''} - 2$ rounds for the budget to drop below 1 after repeated pulls of x^p. Using this and the observation that the budget dropping below 1 is contained in the event that the total drift in those rounds is nonpositive, we can bound (a) as

$$\mathbf{Pr}\left[t_{i+1} \text{ exists } | t_i'', t_i' \text{ exist}\right] \leq \sum_{q=t_i''+\tau_{t_i''}-2}^{T} \mathbf{Pr}\left[\sum_{t=t_i''+1}^{q} d_t \leq 0\right] \tag{65}$$

$$\leq \sum_{q=t_i''+\tau_{t_i''}-2}^{T} \exp\left(-\frac{1}{2}\delta_{\text{drift}}^2(\tau_{t_i''} - 2)\right) \tag{66}$$

$$\leq \sum_{q=t_i''+\tau_{t_i''}-2}^{T} \exp\left(-\frac{1}{2}\delta_{\text{drift}}^2\tau_{t_i''}\right) \tag{67}$$

$$= \sum_{q=t_i''+\tau_{t_i''}-2}^{T} \exp\left(-\frac{1}{2}\delta_{\text{drift}}^2 c\log(T - t_i'')\right) \tag{68}$$

$$\leq \sum_{q=t_i''+\tau_{t_i''}-2}^{T} (T - t_i'')^{-3}, \tag{69}$$

where the second inequality follows from the Azuma-Hoeffding inequality applied to the sequence of drifts sampled from x^p and the last inequality follows because $c \geq \frac{6}{\delta_{\text{drift}}^2}$. The summation is over at most $T - t_i''$ terms because there are at most $T - t_i''$ rounds left after round t_i''. Therefore, we have that

$$\mathbf{Pr}\left[t_{i+1} \text{ exists } | t_i'', t_i' \text{ exist}\right] \leq (T - t_i'')^{-2}. \tag{70}$$

Substituting this in Eq. (64), we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[B_{t-1}<1]\right] \le \tilde{C}\left(1+\sum_{i=0}^{T-1}\mathbf{Pr}\left[t_{i+1}\text{ exists }|t_i', t_i''\text{ exist}\right]\right) \tag{71}$$

$$\le \tilde{C}\left(1+\sum_{i=0}^{T-1}(T-t_i'')^{-2}\right) \tag{72}$$

$$\le \tilde{C}\left(1+\sum_{i=0}^{\infty}(T-t_i'')^{-2}\right) \tag{73}$$

$$\le \tilde{C}\left(1+\frac{\pi^2}{6}\right). \tag{74}$$

This completes the proof.

B.5 Proof of Lemma 3.6

Let E_q denote the event that the negative drift arm x^n is pulled consecutively in exactly the last q rounds, i.e., $x_t = x^n$ for all $t \ge T-q+1$ and $x_t \in \{x^0, x^p\}$ for $t = T-q$ (if $q \neq T$). Note that the events $(E_q : q = 0, \ldots, T)$ are disjoint. Let S_q denote the event that the total drift in the last q pulls of x^n is greater than $\frac{1}{2}\mu^d_{x^n} q$, i.e., $\sum_{t\ge T-q+1} d_t > \frac{1}{2}\mu^d_{x^n} q$. We can upper bound the expected leftover budget by conditioning on these events as follows.

$$\mathbb{E}[B_T] = \sum_{q=0}^{T}\mathbf{Pr}[E_q]\,\mathbb{E}[B_T|E_q] \tag{75}$$

$$\le \sum_{q=0}^{T}\mathbb{E}[B_T|E_q] \tag{76}$$

$$= \sum_{q=0}^{T}\underbrace{\mathbb{E}[B_T|E_q,S_q]}_{(a)}\ \underbrace{\mathbf{Pr}[S_q|E_q]}_{(b)} + \underbrace{\mathbb{E}[B_T|E_q,S_q^c]}_{(c)}\ \underbrace{\mathbf{Pr}[S_q^c|E_q]}_{(d)}. \tag{77}$$

If $q = 0$, then the expected leftover budget is trivially at most a constant. We can bound the four terms for $q \ge 1$ as follows:

(a) We have
$$\mathbb{E}[B_T|E_q,S_q] \le c\log q + q \tag{78}$$
because (i) `ControlBudget` (Algorithm 1) pulls x^0 or x^p in round $T-q$ if $B_{T-q-1} < \tau_{T-q} = c\log q$; and (ii) conditioned on the event S_q, the total drift in the last q rounds can be at most q as the drifts are bounded by 1.

(b) We have
$$\mathbf{Pr}[S_q|E_q] \le \exp\left(-\frac{1}{16}(\mu^d_{x^n})^2 q\right) \tag{79}$$
because (i) the sequence of drifts observed from q pulls of the negative drift arm x^n is a supermartingale difference sequence; and (ii) by the Azuma-Hoeffding inequality, the probability the sum S_q is greater than half its expected value is at most $\exp\left(-\frac{1}{16}(\mu^d_{x^n})^2 q\right)$.

(c) We have
$$\mathbb{E}[B_T|E_q,S_q^c] \le \left(c\log q + \frac{1}{2}\mu^d_{x^n} q\right) \tag{80}$$
because (i) `ControlBudget` (Algorithm 1) pulls x^0 or x^p in round $T-q$ if $B_{T-q-1} < \tau_{T-q} = c\log q$; and (ii) conditioned on the event S_q^c, the total drift in the last q rounds can be at most $\frac{1}{2}\mu^d_{x^n} q$.

(d) We have

$$\mathbf{Pr}[S_q^c|E_q] \leq 1 \tag{81}$$

trivially.

Therefore,

$$\mathbb{E}[B_T] \leq \sum_{q=0}^{T} \underbrace{(c\log q + q)\exp\left(-\frac{1}{16}(\mu_{x^n}^d)^2 q\right)}_{(e)} + \underbrace{\left(c\log q + \frac{1}{2}\mu_{x^n}^d q\right)}_{(f)}. \tag{82}$$

This summation is a constant in terms of T:

1. Term (e) is a constant because $c\log q < q$ for q large enough and $\sum_{q=1}^{\infty} q\exp(-aq)$ converges to $\exp(a)(1-\exp(a))^{-2}$.
2. Term (f) is a constant because this term is negative for q large enough as $\mu_{x^n}^d < 0$ and is maximized at $q = \frac{2c}{|\mu_{x^n}^d|}$.

Finally, we can bound the expected leftover budget as

$$\mathbb{E}[B_T] \leq \tilde{C} = \tilde{O}\left(\left(1-\exp\left(\frac{\delta_{\text{drift}}^2}{16}\right)\right)^{-2} + \frac{1}{\delta_{\text{drift}}^2}\right), \tag{83}$$

where the last equality follows when $c \geq \frac{6}{\delta_{\text{drift}}^2}$. This completes the proof.

C Proofs for Section 3.2

C.1 Proof of Lemma 3.9

It suffices to show that $\gamma = \frac{\sigma_{\min}\min\{\delta_{\text{support}},\delta_{\text{slack}}\}}{4m}$ is a feasible solution the Eq. (8).

First, we show that $p = D^{-1}(b+\gamma s_t) \geq 0$. For each $x \in X$,

$$\begin{aligned}
e_x^T D^{-1}(b+\gamma s_t) &= e_x^T D^{-1} b + \gamma e_x^T D^{-1} s_t \\
&= p_x^* + \gamma e_x^T D^{-1} s_t \\
&\geq \delta_{\text{support}} - \gamma \|D^{-1} s_t\|_2 \\
&\geq \delta_{\text{support}} - \gamma \frac{1}{\sigma_{\min}}\sqrt{m} \\
&\geq 0.
\end{aligned}$$

Second, we show that for any non-binding resource j, $d_j^T D^{-1}(b+\gamma s_t) \geq \frac{\delta_{\text{slack}}}{2}$:

$$\begin{aligned}
d_j^T D^{-1}(b+\gamma s_t) &= \sum_{x\in X} d_j(x,\mu)p_x^* + \gamma d_j^T D^{-1} s_t \\
&\geq \delta_{\text{slack}} - \gamma |d_j^T D^{-1} s_t| \\
&\geq \delta_{\text{slack}} - \gamma \|d_j^T\|_2 \|D^{-1}\|_2 \|s_t\|_2 \\
&\geq \delta_{\text{slack}} - \gamma \frac{1}{\sigma_{\min}} m \\
&\geq \frac{\delta_{\text{slack}}}{2} \\
&\geq \frac{\gamma}{2},
\end{aligned}$$

where the last inequality follows because $\sigma_{\min}, \delta_{\text{slack}}, \delta_{\text{support}} < 1$.

C.2 Proof of Lemma 3.10

Divide the T rounds into two phases: $P_1 = \{1, \ldots, T - \exp(3/c)$ and $P_2 = \{1, \ldots, T\} \setminus P_1$. Note that P_2 consists of $\exp(3/c) = O(\exp(\gamma^*)) = O(1)$ rounds, where the last equality follows because γ^* is bounded by 1. Therefore, the expected number of null arm pulls in this phase is $O(1)$ and it suffices to bound the expected number of null arm pulls in P_1.

We can write the expected number of rounds in which there exists a resource whose budget is less than 1 as

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j\in\mathcal{J}} \mathbb{1}[B_{t-1,j} < 1]\right] = \sum_{j\in\mathcal{J}} \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[B_{t-1,j} < 1]\right]}_{(a)}. \tag{84}$$

We can bound term (a) above the same way as in the proof of Lemma 3.5 (Appendix B.4) with δ_{drift} replaced by γ^*.. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j\in\mathcal{J}} \mathbb{1}[B_{t-1,j} < 1]\right] \le m\tilde{C}\left(1 + \frac{\pi^2}{6}\right), \tag{85}$$

where $\tilde{C}$ is defined in Eq. (37).

C.3 Proof of Lemma 3.11

Consider an arbitrary resource $j \in J^*$. Recall the vector s_t defined in `ControlBudget` (Algorithm 2). If i denote the row corresponding to resource j, then the ith entry of s_t, denoted by $s_t(i)$, is -1 if $B_{t-1,j} < \tau_t$ and $+1$ otherwise.

Let E_q denote the event that the $s_t(i)$ is equal to -1 consecutively in exactly the last q rounds, i.e., $s_t(i) = -1$ for all $t \ge T - q + 1$ and $s_t(i) = +1$ for $t = T - q$ (if $q \ne T$). Note that the events $(E_q : q = 0, \ldots, T)$ are disjoint. Let S_q denote the event that the total drift for j in the last q rounds is greater than $\frac{1}{2}(-\gamma^*)q$, i.e., $\sum_{t\ge T-q+1} d_t > \frac{1}{2}(-\gamma^*)q$. We can upper bound the expected leftover budget of resource j by conditioning on these events as follows.

$$\begin{aligned}
\mathbb{E}[B_{T,j}] &= \sum_{q=0}^{T} \mathbf{Pr}[E_q]\, \mathbb{E}[B_{T,j}|E_q] && (86)\\
&\le \sum_{q=0}^{T} \mathbb{E}[B_{T,j}|E_q] && (87)\\
&= \sum_{q=0}^{T} \underbrace{\mathbb{E}[B_{T,j}|E_q, S_q]}_{(a)}\, \underbrace{\mathbf{Pr}[S_q|E_q]}_{(b)} + \underbrace{\mathbb{E}[B_{T,j}|E_q, S_q^c]}_{(c)}\, \underbrace{\mathbf{Pr}[S_q^c|E_q]}_{(d)}. && (88)
\end{aligned}$$

If $q = 0$, then the expected leftover budget is trivially at most a constant. We can bound the four terms for $q \ge 1$ as follows:

(a) We have
$$\mathbb{E}[B_{T,j}|E_q, S_q] \le c\log q + q \tag{89}$$
because (i) `ControlBudget` (Algorithm 2) sets $s_t(i) = +1$ in round $T - q$ if $B_{T-q-1,j} < \tau_{T-q} = c\log q$; and (ii) conditioned on the event S_q, the total drift in the last q rounds can be at most q as the drifts are bounded by 1.

(b) We have
$$\mathbf{Pr}[S_q|E_q] \le \exp\left(-\frac{1}{16}(\gamma^*)^2 q\right) \tag{90}$$
because (i) the sequence of drifts observed in rounds $t \ge T - q + 1$ is a supermartingale difference sequence with $\mathbb{E}[d_{s,j}|d_{T-q+1,j}, \ldots, d_{s-1,j}] \le -\gamma^*$; and (ii) by the Azuma-Hoeffding inequality, the probability the sum S_q is greater than half its expected value is at most $\exp\left(-\frac{1}{16}(\gamma^*)^2 q\right)$.

(c) We have

$$\mathbb{E}[B_{T,j}|E_q, S_q^c] \le \left(c\log q + \frac{1}{2}(-\gamma^*)q\right) \tag{91}$$

because (i) `ControlBudget` (Algorithm 2) sets $s_t(i) = +1$ in round $T-q$ if $B_{T-q-1,j} < \tau_{T-q} = c\log q$; and (ii) conditioned on the event S_q^c, the total drift in the last q rounds can be at most $\frac{1}{2}(-\gamma^*)q$.

(d) We have

$$\mathbf{Pr}[S_q^c|E_q] \le 1 \tag{92}$$

trivially.

Therefore,

$$\mathbb{E}[B_{T,j}] \le \sum_{q=0}^{T} \underbrace{(c\log q + q)\exp\left(-\frac{1}{16}(\gamma^*)^2 q\right)}_{(e)} + \underbrace{\left(c\log q + \frac{1}{2}(-\gamma^*)q\right)}_{(f)}. \tag{93}$$

This summation is a constant in terms of T:

1. Term (e) is a constant because $c\log q < q$ for q large enough and $\sum_{q=1}^{\infty} q\exp(-aq)$ converges to $\exp(a)(1-\exp(a))^{-2}$.
2. Term (f) is a constant because this term is negative for q large enough and is maximized at $q = \frac{2c}{\gamma^*}$.

Finally, we can bound the expected leftover budget as

$$\mathbb{E}[B_{T,j}] \le \tilde{C} = \tilde{O}\left(\left(1-\exp\left(\frac{\gamma^{*2}}{16}\right)\right)^{-2} + \frac{1}{\gamma^{*2}}\right), \tag{94}$$

where the last equality follows when $c \ge \frac{6}{\gamma^{*2}}$. This completes the proof.

D Proofs for Section 4

D.1 Proof of Lemma 4.1

It suffices to show that the complement of the clean event occurs with probability at most $5mT^{-2}$.

For (i) in the definition of the clean event (Definition 4.2), by taking a union bound over the components of the outcome vector and using Hoeffding's inequality, we have

$$\begin{aligned}
\mu_x^o \notin [\mathsf{LCB}_t(x), \mathsf{UCB}_t(x)] &\le 2(m+1)\exp\left(-2n_t(x)\mathsf{rad}_t(x)^2\right) && (95)\\
&\le 4m\exp\left(-2n_t(x)\frac{8\log T}{n_t(x)}\right) && (96)\\
&\le 4mT^{-2}. && (97)
\end{aligned}$$

For (ii) in the definition of the clean event (Definition 4.2), a similar approach works. Let $S_{n,j}$ denote the sum of the drifts for resource $j \in \mathcal{J}$ after n pulls of the null arm x^0. By the union bound and

Hoeffding's inequality,

$$\begin{aligned}
\mathbf{Pr}\left[\exists j \in \mathcal{J} \text{ s.t. } S_{n,j} < w\right] &\leq m \exp\left(-\frac{1}{4} w \mu^d_{x^0}\right) && (98)\\
&\leq m \exp\left(-\frac{1}{4} w \delta_{\text{drift}}\right) && (99)\\
&\leq m \exp\left(-\frac{1}{4} \frac{1024 k m^2 \log T}{\delta^2_{\text{drift}} \sigma^2_{\min}} \delta_{\text{drift}}\right) && (100)\\
&= m \exp\left(-\frac{256 k m^2 \log T}{\delta_{\text{drift}} \sigma^2_{\min}}\right) && (101)\\
&\leq m \exp\left(-256 k m^2 \log T\right) && (102)\\
&\leq m \exp\left(-256 \log T\right) && (103)\\
&\leq m T^{-2}, && (104)
\end{aligned}$$

where Eq. (102) follows because $\delta_{\text{drift}} \in (0, 1]$ and $\sigma_{\min} \in (0, 1)$. This shows that the probability of the complement of the clean event is at most $5mT^{-2}$ and completes the proof.

D.2 Proof of Lemma 4.2

We will prove the lemma for $\mathsf{OPT}_{\mathsf{LP}}$ because the other cases are similar. Simplifying and overloading notation for this proof, we denote the probability simplex over k dimensions as Δ_k, and the vector of expected rewards, the matrix of expected drifts and the right-hand side of the budget constraints as

$$r = \begin{bmatrix} \mu^r_1 \\ \vdots \\ \mu^r_k \end{bmatrix}, \; D = \begin{bmatrix} \mu^{d,1}_1 & \dots & \mu^{d,1}_k \\ & \ddots & \\ \mu^{d,m}_1 & \dots & \mu^{d,m}_k \end{bmatrix}, \; b = -\frac{B}{T}\mathbf{1}. \tag{105}$$

We will use $\bar{r}$ and $\bar{D}$ to denote the empirical versions of the rewards and drifts. We can write

$$\begin{aligned}
\mathsf{OPT}_{\mathsf{LP}} &= \max_{p \in \Delta_k} r^T p && \text{s.t. } Dp \geq b,\\
\mathsf{UCB}_t(\mathsf{OPT}_{\mathsf{LP}}) &= \max_{q \in \Delta_k} (\bar{r} + \mathsf{rad}_t)^T q && \text{s.t. } (\bar{D} + \mathsf{rad}_t) q \geq b\\
&\leq \max_{q \in \Delta_k} (r + 2\mathsf{rad}_t)^T q && \text{s.t. } (D + 2\mathsf{rad}_t) q \geq b\\
&\leq 2\mathsf{rad}_t + \max_{q \in \Delta_k} r^T q && \text{s.t. } Dq \geq b - 2\mathsf{rad}_t,
\end{aligned}$$

where the second-last inequality follows because we are conditioning on the clean event. Therefore, using D' and b' to denote the submatrix and subvector corresponding to the binding constraints, we have

$$\begin{aligned}
\mathsf{UCB}_t(\mathsf{OPT}_{\mathsf{LP}}) - \mathsf{OPT}_{\mathsf{LP}} &\leq 2\mathsf{rad}_t + |r^T p - r^T q|\\
&\leq 2\mathsf{rad}_t + |r^T (D')^{-1} b' - r^T (D')^{-1} (b' - 2\mathsf{rad}_t)|\\
&\leq 2\mathsf{rad}_t + \|r\|_2 \|(D')^{-1}\|_2 \|2\mathsf{rad}_t\|_2\\
&\leq 2\mathsf{rad}_t + 2m\mathsf{rad}_t \frac{1}{\sigma_{\min}}\\
&\leq \frac{4m}{\sigma_{\min}} \mathsf{rad}_t,
\end{aligned}$$

where last inequality follows because $\sigma_{\min} < 1 \leq m$. Since the LCB is defined by subtracting rad_t from the empirical means, we obtain the same upper bound on $\mathsf{OPT}_{\mathsf{LP}} - \mathsf{LCB}_t(\mathsf{OPT}_{\mathsf{LP}})$ and using the triangle inequailty completes the proof.

D.3 Proof of Theorem 4.1

Since the complement of the clean event occurs with probability at most $O(mT^{-2})$ and contributes $O(T)$ to the regret, it suffices to bound the regret conditioned on the clean event. So, condition on the clean event for the rest of the proof. Phase one contributes at most

$$O\left(\frac{km^2}{\min\{\delta_{\text{drift}}^2, \sigma_{\min}^2\}\Delta^2}\right) \cdot \log T \tag{106}$$

to the regret by Corollary 4.2. Phase two contributes at most

$$O\left(\frac{k}{\gamma^{*2}}\right) \log T \tag{107}$$

to the regret.

Observe that after phase two, $\texttt{rad}_t(x) \leq \frac{\gamma^{*2}}{2}$ for all $x \in X^*$. Combining this with Eqs. (8) and (15) to (17), we have that $(\gamma^*, D^{-1}(b + \gamma^* s_t))$ is a feasible solution to the optimization problem solved by `ExploreThenControlBudget` (Algorithm 3). Therefore, (γ_t, p_t) ensure that there is drift of magnitude at least $\frac{\gamma^*}{8}$ in the "correct directions". As noted in the end of Section 3.2, the regret analysis of `ControlBudget` (Algorithm 2) requires the algorithm to know X^*, J^*, and find a probability vector p_t that ensures drifts bounded away from zero in the "correct directions". Therefore, by Theorem 3.2, phase three contributes at most $\tilde{C}'$ to the regret, where $\tilde{C}$ is the constant in Theorem 3.2. Combining the contribution from the three phases, we have that

$$R_T(\texttt{ExploreThenControlBudget}) \leq \tilde{C} \cdot \log T, \tag{108}$$

where γ^* (defined in Lemma 3.9) and $\tilde{C}$ are constants with

$$\tilde{C} = O\left(\frac{km^2}{\min\{\delta_{\text{drift}}^2, \sigma_{\min}^2\}\Delta^2} + k(\gamma^*)^{-2} + \tilde{C}'\right). \tag{109}$$

D.4 Proof of Theorem 4.2

Note that BwK is not automatically a special case of our model because of our assumption that the null arm has strictly positive drift for every resource. In this section we present a reduction from BwK with $\frac{B}{T}$ bounded away from 0 to our model. We show that our results imply a logarithmic regret bound for BwK under certain assumptions.

Reduction Assume we are given an instance of BwK with $\frac{B}{T} \geq \delta_{\text{drift}} > 0$. (Existing results on logarithmic regret for BwK also assume the ratio of the initial budget to the time horizon is bounded away from 0 [14].) We will reduce the given BwK instance to a problem in our model. The reduction initializes an instance of `ExploreThenControlBudget` (Algorithm 3) running in a simulated environment with the same set of arms as in the given BwK instance, plus an additional null arm whose drift is equal to δ_{drift} deterministically for each resource. The reduction will maintain two time counters: t_a is the actual number of time steps that have elapsed in the BwK prlblem, and t_s is the number of time steps that have elapsed in the simulation environment in which Algorithm 3 is running. Likewise, there are two vectors that track the remaining budget: B_a is the remaining budget in the actual BwK problem our reduction is solving, while B_s is the remaining budget in the simulation environment. These two budget vectors will always be related by the equation

$$B_s = B_a - T\delta_{\text{drift}}\mathbf{1} + t_s\delta_{\text{drift}}\mathbf{1}. \tag{110}$$

In particular, the initial budget of each resource is initialized (at simulated time $t_s = 0$) to $B - T\delta_{\text{drift}}$.

Each step of the reduction works as follows. We call Algorithm 3 to simulate one time step in the simulated environment. If Algorithm 3 recommends to pull a non-null arm x, we pull arm x, increment both of the time counters (t_a and t_s), and update the vector of remaining resource amounts, B_a, according to the resources consumed by arm x. If Algorithm 3 recommends to pull the null arm, we do not pull any arm, and we leave t_a and B_a unchanged; however, we still increment the simulated time counter t_s. Finally, regardless of whether a null or non-null arm was pulled, we update B_s to satisfy Eq. (110).

Correctness Since the reduction pulls the same sequence of non-null arms as Algorithm 3 until the BwK stopping condition is met and the additional pulls of the null arm in the simulation environment yield zero reward, the total reward in the actual BwK problem equals the total reward earned in the simulation environment at the time when the BwK stopping condition is met and the reduction ceases running. Since Algorithm 3 maintains the invariant that B_s is a nonnegative vector, Eq. (110) ensures that B_a will also remain nonnegative as long as $t_s \geq T$ must hold. Theorem 4.1 ensures that the total expected reward earned in the simulation environment and hence, also in the BwK problem itself, is bounded below by $T \cdot \mathsf{OPT}_{\mathsf{LP}} - \tilde{C} \cdot \log T$, where $\tilde{C}$ is the constant in Theorem 4.1 and $\mathsf{OPT}_{\mathsf{LP}}$ denotes the optimal value of the LP relaxation (Eq. (1)) for the simulation environment.

We would like to show that this implies the regret of the reduction (with respect to the LP relaxation of BwK) is bounded by $\tilde{C} \cdot \log T$. To do so, we must show that the LP relaxations of the original BwK problem and the simulation environment have the same optimal value. Let μ_x^r and $\mu_x^{d,j}$ denote the expected reward and expected drifts in the actual BwK problem with arm set $\mathcal{X}$, and let $\hat{\mu}_x^r$ and $\hat{\mu}_x^{d,j}$ denote the expected reward and drifts in the simulation environment with arm set $\mathcal{X}^+ = \mathcal{X} \cup \{x^0\}$. The two LP formulations are as follows.

$$
\begin{array}{llll}
\max_p & \sum_{x \in \mathcal{X}} p_x \mu_x^r & & \\
\text{s.t.} & \sum_{x \in \mathcal{X}} p_x \mu_x^{d,j} & \geq -\frac{B}{T} & \forall j \in \mathcal{J}, \\
& \sum_{x \in \mathcal{X}} p_x & \leq 1 & \\
& p_x & \geq 0 & \forall x \in \mathcal{X}.
\end{array}
\qquad
\begin{array}{llll}
\max_p & \sum_{x \in \mathcal{X}} p_x \hat{\mu}_x^r & & \\
\text{s.t.} & \sum_{x \in \mathcal{X}} p_x \hat{\mu}_x^{d,j} & \geq -\frac{B}{T} - \delta_{\text{drift}} & \forall j \in \mathcal{J}, \\
& \sum_{x \in \mathcal{X}^+} p_x & = 1 & \\
& p_x & \geq 0 & \forall x \in \mathcal{X}^+.
\end{array}
$$

The differences between the two LP formulations lie in substituting $\hat{\mu}$ for μ, substituting $\mathcal{X}^+$ for $\mathcal{X}$, and transforming the inequality constraint $\sum_{x \in \mathcal{X}} p_x \leq 1$ into an equality constraint $\sum_{x \in \mathcal{X}^+} p_x = 1$. We know that $\mu_x^r = \hat{\mu}_x^r$ for every $x \in \mathcal{X}$ and $\hat{\mu}_{x^0}^r = 0$. Furthermore, $\hat{\mu}_x^{d,j}$ denotes the expected drift of resource j in the simulation environment when arm x is pulled. This can be written as the sum of two terms: drift $\mu_x^{d,j}$ is the expectation of the (non-positive) quantity added to the jth component of budget vector B_a when pulling arm x in the actual BwK environment; in addition to this non-positive drift, there is a deterministic positive drift of δ_{drift} due to incrementing the simulation time counter t_s and recomputing B_s using Eq. (110). Hence, $\hat{\mu}_x^{d,j} = \mu_x^{d,j} + \delta_{\text{drift}}$ for all $x \in \mathcal{X}$ and $j \in \mathcal{J}$. Furthermore, $\hat{\mu}_{x^0}^{d,j} = \delta_{\text{drift}}$. Hence, for any vector $\vec{p}$ representing a probability distribution on $\mathcal{X}^+$, we have

$$\sum_{x \in \mathcal{X}^+} p_x \hat{\mu}_x^{d,j} = \left(\sum_{x \in \mathcal{X}} p_x \hat{\mu}_x^{d,j} \right) + \delta_{\text{drift}}. \tag{111}$$

Accordingly, a vector $\vec{p}$ satisfies the constraints of the BwK LP relaxation above if and only if the probability vector on $\mathcal{X}^+$ obtained from $\vec{p}$ by setting $p_{x^0} = 1 - \sum_{x \in \mathcal{X}} p_x$ satisfies the constraints of the second LP relaxation above. This defines a one-to-one correspondence between the sets of vectors feasible for the two LP formulations. Furthermore, this one-to-one correspondence preserves the value of the objective function because $\hat{\mu}_x^r = \mu_x^r$ for $x \in \mathcal{X}$ and $\hat{\mu}_{x^0}^r = 0$. Thus, the optimal value of the two linear programs is the same. This completes the proof.