Budget-Constrained Bandits over General Cost and Reward Distributions

Supplementary Material

Semih Cayci
ECE, The Ohio State University

Atilla Eryilmaz
ECE, The Ohio State University

R. Srikant
CSL and ECE, UIUC

Abstract

We consider a budget-constrained bandit problem where each arm pull incurs a random cost, and yields a random reward in return. The objective is to maximize the total expected reward under a budget constraint on the total cost. The model is general in the sense that it allows correlated and potentially heavy-tailed cost-reward pairs that can take on negative values as required by many applications. We show that if moments of order $(2+\gamma)$ for some $\gamma > 0$ exist for all cost-reward pairs, $O(\log B)$ regret is achievable for a budget $B > 0$. In order to achieve tight regret bounds, we propose algorithms that exploit the correlation between the cost and reward of each arm by extracting the common information via linear minimum mean-square error estimation. We prove a regret lower bound for this problem, and show that the proposed algorithms achieve tight problem-dependent regret bounds, which are optimal up to a universal constant factor in the case of jointly Gaussian cost and reward pairs.

1 Introduction

Multi-armed bandit problem (MAB) has been the prominent model for the exploration-and-exploitation dilemma since its introduction (Robbins, 1952; Lai and Robbins, 1985; Berry and Fristedt, 1985). Due to the universality of the dilemma, bandit algorithms have found a broad area of applications from medical trials and dynamic pricing to ad allocation. As a common feature of all MAB instances, each action depletes a cost from a limited budget, and a random reward is obtained in return. In such a setting, the aim of the decision maker is to balance the exploration and exploitation at every step so as to maximize the cumulative reward until depleting the budget. In the classical MAB setting, each action is assumed to consume a known deterministic amount of resource, i.e., one time-slot. However, in many problems of interest, different tasks consume different and random amount of resources, which can be unbounded and potentially correlated with the reward. The applications of this extended setting include routing in communications and task scheduling in computing systems, where the controller sequentially makes a selection among multiple arms (alternative paths or task types) so as to maximize the total reward (i.e., throughput) within a given time budget. In these applications, the cost (i.e., completion time) and reward of each arm pull can be potentially correlated and heavy-tailed (Harchol-Balter, 2000; Jelenković and Tan, 2013).

In this paper, we investigate the unique dynamics of this extended budget-constrained bandit setting with general cost and reward distributions. Unlike the classical stochastic MAB problem, each action incurs a random cost and yields a random reward in our model. Under a budget constraint B, the objective of the controller is to maximize the expected cumulative reward until the total cost exceeds the budget. As we will see, the correlation and variability of the cost-reward pairs can have a substantial impact on the performance in this bandit setting, which we incorporate in the design of learning algorithms for near-optimal performance. Many of our results are obtained for a very general setting where the cost and reward can be correlated and heavy-tailed, but sharper results are presented for some interesting special cases.

1.1 Main Contributions

The main objective in this paper is to design efficient algorithms that achieve provably tight regret bounds

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

in an extended setting of correlated and potentially heavy-tailed cost and reward. Our main contributions are as follows:

1. **Exploiting the correlation:** One of the key contributions in this work is to use a linear minimum mean square (LMMSE) estimator to extract and exploit the correlation between the cost and reward of an arm (see Section 4.2). Furthermore, we incorporate the effect of variability in cost-reward pairs through variance. Consequently, we achieve provably tight problem-dependent regret bounds in an extended setting of unbounded cost and reward.

2. **Extension to unbounded cost and reward:** We develop novel design and analysis methods for the setting of unbounded and potentially heavy-tailed cost and reward pairs, and show that $O\big(\log(B)\big)$ regret is achievable if moments of order $2+\gamma$ exist for some $\gamma > 0$ for all cost and reward pairs (see Section 4.3).

3. **Regret lower bounds:** We establish a regret lower bound for the budget-constrained bandit problem (see Section 5). By using this result, we obtain explicit regret lower bounds for jointly Gaussian cost-reward distributions. Consequently, we prove that the algorithms we propose in this paper achieve tight regret bounds, which are optimal up to a constant factor in the case of jointly Gaussian cost and reward.

1.2 Related Work

The classical stochastic multi-armed bandit problem, which is a specific case of the model we study in this paper, has been extensively studied in the literature. For detailed discussion on the basic model, we refer to (Bubeck et al., 2012; Berry and Fristedt, 1985).

The budget-constrained MAB problem and its variants were investigated in a variety of papers. In (Tran-Thanh et al., 2012) and (Combes et al., 2015), budget-constrained multi-armed bandit problem is investigated where each arm pull incurs an arm-dependent and deterministic cost. In (Guha and Munagala, 2009), the budgeted-bandit problem with deterministic costs is investigated from a Bayesian perspective, and constant-factor approximation algorithms are proposed. In (György et al., 2007), the continuous-time extension of the MAB problem with side information is investigated, which is an early example for the budget-constrained bandit problem. In (Badanidiyuru et al., 2013; Agrawal and Devanur, 2014), the bandit problem under multiple budget constraints is examined, and problem-independent regret bounds of order $\tilde{O}(\sqrt{B})$ are obtained. Bandits with knapsacks have been extended to other bandit settings (Agrawal and Devanur, 2016; Badanidiyuru et al., 2014; Sankararaman and Slivkins, 2017; Ding et al., 2013). In (Xia et al., 2015, 2016), the budget-constrained MAB problem is explored in a similar setting to ours. In these works, the cost and reward of each arm are supported in $[0, 1]$, and the correlation between them is not exploited. In (Cayci et al., 2019), the authors consider a variation of the budget-constrained bandit problem where the controller has the option to interrupt an ongoing cycle for a faster alternative. The interruption mechanism brings significantly different dynamics to the problem that is investigated in this paper.

Bandits with heavy-tailed reward distributions are considered in (Liu and Zhao, 2011; Bubeck et al., 2013). These papers are still in the scope of the classical MAB setting: the budget is consumed deterministically at rate 1 by each action, so the dynamics of the random resource consumption with heterogeneous statistics are not included in the model.

2 System Setup

In this paper, we consider a bandit problem with K arms. The set of arms is denoted by $\mathbb{K} = \{1, 2, \ldots, K\}$. Each arm $k \in \mathbb{K}$ is described by a two-dimensional random process $\{(X_{n,k}, R_{n,k}) : n \geq 1\}$ that is independent from other arms. If arm k is chosen at n-th epoch, it incurs a cost of $X_{n,k}$ and yields a reward of $R_{n,k}$, where both are learned via a bandit feedback only after the decision is made. The controller has a cost budget $B > 0$, and tries to maximize the expected cumulative reward it receives by sampling the arms wisely under this budget constraint.

The pair $(X_{n,k}, R_{n,k})$ is assumed to be independent and identically distributed over n, but the cost $X_{n,k}$ and reward $R_{n,k}$ can be positively correlated. We allow $X_{n,k}$ to take on negative values, but the drift is assumed to be positive, i.e., there exists $\mu_* > 0$ such that $\mathbb{E}[X_{n,k}] \geq \mu_* > 0$ for all k.

Let π be an algorithm that yields a sequence of arm pulls $\{I_n^\pi \in \mathbb{K} : n \geq 1\}$. Under π, the history until epoch n is the following filtration:

$$\mathcal{F}_n^\pi = \sigma(\{(X_{j,k}, R_{j,k}) : I_j^\pi = k, 1 \leq j \leq n\}), \qquad (1)$$

where $\sigma(X)$ denotes the sigma-field of a random variable X. We call an algorithm π admissible if π is non-anticipating, i.e., $\{I_n^\pi = k\} \in \mathcal{F}_{n-1}^\pi$ for all k, n. The set of all admissible policies is denoted as Π.

The total cost incurred in n epochs under an admissible policy $\pi \in \Pi$ is a controlled random walk which is defined as $S_n^\pi = \sum_{i=1}^n X_{i,I_i^\pi}$. The arm pulling process under an algorithm π continues until the budget B is depleted. We assume that the reward corresponding to the final epoch during which the budget is depleted is gathered by the controller. Thus, the total number of pulls under π is defined as follows:

$$N_\pi(B) = \inf\Big\{n : S_n^\pi > B\Big\}. \tag{2}$$

Note that the total number of pulls $N_\pi(B)$ is a stopping time adapted to the filtration $\{(\mathcal{F}_t^\pi) : t \geq 0\}$. With these definitions, the cumulative reward under a policy π can be written as follows:

$$\texttt{REW}_\pi(B) = \sum_{i=1}^{N_\pi(B)} R_{i,I_i^\pi}. \tag{3}$$

The objective in this paper is to design algorithms that achieve maximum $\mathbb{E}[\texttt{REW}_\pi(B)]$, or equivalently minimum regret, which is defined as follows:

$$Reg_\pi(B) = \mathbb{E}[\texttt{REW}_{\pi^{\texttt{opt}}}(B)] - \mathbb{E}[\texttt{REW}_\pi(B)], \tag{4}$$

where $\pi^{\texttt{opt}}(B)$ denotes the optimal policy:

$$\pi^{\texttt{opt}}(B) \in \arg\max_{\pi' \in \Pi} \mathbb{E}[\texttt{REW}_{\pi'}(B)],$$

for any $B > 0$.

In the following section, we investigate the optimal policy that maximizes the expected cumulative reward when all arm distributions are known, and provide low-complexity approximations that have desirable performance characteristics.

3 Approximations of the Oracle

The optimization problem described in Section 2 is a variant of the unbounded knapsack problem, and it is known that similar stochastic control problems are PSPACE-hard (Badanidiyuru et al., 2013; Papadimitriou and Tsitsiklis, 1999). In order to find a tractable benchmark, we will consider approximation algorithms with provably good performance in this section.

The main quantity of interest will be the reward rate, which is defined as follows:

$$r_k = \frac{\mathbb{E}[R_{1,k}]}{\mathbb{E}[X_{1,k}]},\ k \in \mathbb{K}. \tag{5}$$

Intuitively, if arm k is chosen persistently until the budget $B > 0$ is depleted, the cumulative reward becomes $r_k B + o(B)$ as $B \to \infty$. The additive $o(B)$ term is $O(1)$ if $\mathbb{E}[(X_{1,k}^+)^2] < \infty$ by Lorden's inequality (Asmussen, 2008). Hence, pulling the arm with the highest reward rate is a logical choice.

In the following, we prove that the optimality gap is $O(1)$ under mild moment conditions, which covers the case of heavy-tailed cost-reward pairs.

Definition 1 (Optimal Static Algorithm)**.** *Let k^* be the arm with the highest reward rate:*

$$k^* \in \arg\max_{k \in \mathbb{K}} r_k.$$

The optimal static policy, denoted by π^*, pulls k^* until the budget is depleted: $I_n^{\pi^*} = k^*$ for all $n \leq N_{\pi^*}(B)$.

The main result of this section is the following proposition, which implies that π^* is a plausible approximation algorithm for $\pi^{\texttt{opt}}(B)$ for all $B > 0$ under mild moment conditions.

Assumption 1. *There exists $\gamma > 0$ such that $\mathbb{E}[(X_{1,k}^+)^{2+\gamma}] < \infty$ for all $k \in \mathbb{K}$.*

Proposition 1 (Optimality Gap for π^*)**.** *Under Assumption 1, there exists a constant*

$$G^\star = G^\star\Big(\min_k \mathbb{E}[X_{1,k}], \max_k Var(X_{1,k})\Big) < \infty,$$

independent of B such that the following holds:

$$\max_{\pi \in \Pi}\ \mathbb{E}[\texttt{REW}\pi(B)] - \mathbb{E}[\texttt{REW}\pi^*(B)] \leq G^\star, \tag{6}$$

for any $B > 0$. Consequently, π^* is asymptotically optimal as $B \to \infty$.

Proof. The proof of Proposition 1 is based on tools from stochastic control, and is given in Appendix A. □

Proposition 1 implies that the optimality gap of the optimal static policy is a constant with respect to the budget B, which depends on the first- and second-order moments of the cost. This extends the result presented in (Xia et al., 2016) for bounded and strictly positive costs to unbounded costs with positive drift that can take on negative values. Also, for small B values, there can be dynamic policies that outperform this simple static policy (Dean et al., 2004). However, the optimality gap is still $O(1)$ for these dynamic policies, therefore we consider π^* for its simplicity and efficiency.

Now that we have an accurate approximation for the oracle, we propose the first and basic algorithms that assume the knowledge of second-order moments.

4 Algorithms for Known Second-Order Moments

In this section, we will assume that the second-order moments of all cost-reward pairs are known by the decision maker. First, in Section 4.2, we will consider the case $(X_{n,k}, R_{n,k})$ are jointly Gaussian, and propose a learning algorithm that achieves tight regret bound on

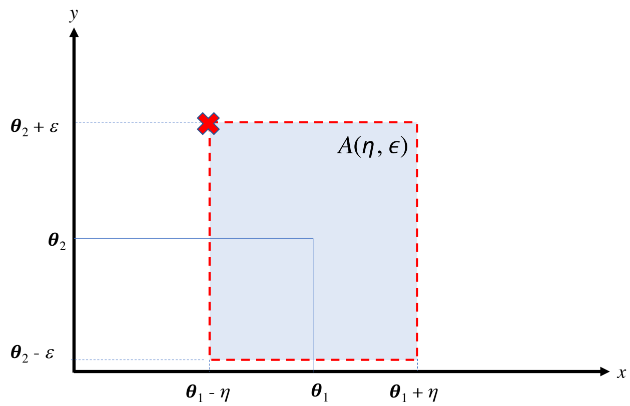

Figure 1: If $(\hat{\theta}_1, \hat{\theta}_2)$ is in the high-probability set $A(\eta, \epsilon)$, then the maximum deviation of $\hat{r} = \frac{\hat{\theta}_2}{\hat{\theta}_1}$ from r is $\frac{\lambda(\epsilon + r\eta)}{\theta_1}$, and it is achieved at the marked corner.

the order of $O(\log(B))$ by using the correlation information. Then, in Section 4.3, we will study the general case where the cost and reward can be unbounded and potentially heavy-tailed, and propose algorithms that achieve the same regret bounds (up to a constant) as the sub-Gaussian case.

The following proposition provides a basis for the algorithm design and analysis throughout the paper.

4.1 Preliminaries: Rate Estimation

Let $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ be a pair of unknown constants for which $r = \frac{\theta_2}{\theta_1}$ is to be estimated. The following proposition yields a useful device to obtain concentration results for r from concentration results for θ_1 and θ_2 for this estimation procedure.

Proposition 2 (Rate Estimation)**.** *Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be estimators for $\theta_1 > 0, \theta_2 \geq 0$, respectively. If*

$$\eta \in \big(0, \frac{\theta_1(\lambda - 1)}{\lambda}\big), \tag{7}$$

for some $\lambda > 1$, then we have the following result:

$$\mathbb{P}\Big(|r - \frac{\hat{\theta}_2}{\hat{\theta}_1}| > \frac{\lambda(\epsilon + r\eta)}{\theta_1}\Big) \leq \mathbb{P}(|\hat{\theta}_1 - \theta_1| > \eta) + \mathbb{P}(|\hat{\theta}_2 - \theta_2| > \epsilon).$$

Therefore, if $\hat{\theta}_1$ and $\hat{\theta}_2$ both achieve exponential convergence rate, then $\frac{\hat{\theta}_2}{\hat{\theta}_1}$ converges to r exponentially fast. The intuition behind the proposition is illustrated in Figure 4.1.

Remark 1 (Stability of the rate estimator)**.** The condition $\eta < \theta_1$, i.e., sufficient concentration of the estimator around the true parameter θ_1, is crucial for Proposition 2. Note that if the variability of the mean estimator is high and thus $A(\eta, \epsilon)$ intersects with the y-axis, then the above bound is useless as $\hat{r}$ can have arbitrarily large deviations from r.

In the following, we propose algorithms under the assumption that the second-order moments for each arm k is known by the controller.

4.2 Sub-Gaussian Case: Algorithm `UCB-B1`

The main idea behind `UCB-B1` is to use an upper confidence bound for the reward rate r_k. Let $T_k(n)$ be the number of pulls for arm k in the first n stages and $\widehat{r}_{k,n} = \frac{\max\{0, \widehat{\mathbb{E}}_n[R_k]\}}{\max\{b, \widehat{\mathbb{E}}_n[X_k]\}}$ where

$$\widehat{\mathbb{E}}_n[X_k] = \frac{1}{T_k(n)} \sum_{i=1}^{n} \mathbb{I}\{I_i = k\} X_{i,k},$$

$$\widehat{\mathbb{E}}_n[R_k] = \frac{1}{T_k(n)} \sum_{i=1}^{n} \mathbb{I}\{I_i = k\} R_{i,k},$$

and $b \leq \mathbb{E}[X_{1,k}]/2$ for all k. Instead of estimating $\mathbb{E}[X_{1,k}]$ and $\mathbb{E}[R_{1,k}]$ separately from the samples of $(X_{n,k}, R_{n,k})$, the correlation between $X_{n,k}$ and $R_{n,k}$ can be exploited to tighten the upper confidence bound for r_k. This is achieved by estimating $R_{n,k}$ by a linear estimator $\omega X_{n,k}$ so as to minimize $Var(R_{n,k} - \omega X_{n,k})$. Let

$$V(X_{1,k}, R_{1,k}) = \min_{\omega \in \mathbb{R}} \; Var(R_{1,k} - \omega X_{1,k}). \tag{8}$$

If $Var(X_{n,k}) > 0$, we have:

$$\begin{aligned} \omega_k &= \arg\min_{\omega \in \mathbb{R}} Var(R_{1,k} - \omega X_{1,k}), \\ &= \frac{Cov(X_{1,k}, R_{1,k})}{Var(X_{1,k})}, \end{aligned} \tag{9}$$

by the orthogonality principle (Poor, 2013), and the optimal value of the objective is given by:

$$V(X_{1,k}, R_{1,k}) = Var(R_{1,k}) - \omega_k^2 Var(X_{1,k}).$$

If $Var(X_{n,k}) = 0$, we have $V(X_{1,k}, R_{1,k}) = Var(R_{1,k})$. This implies that ω_k and V can be computed from the second-order moments of $(X_{n,k}, R_{n,k})$, which are assumed to be given in this section. For simplicity, we assume $\omega_k \leq r_k$ for all k throughout the paper.

For non-negative (M_X, M_R, L) that will be specified later, let

$$\epsilon_{k,n}^{\mathtt{B}} = \frac{2\alpha M_R \log(n)}{3T_k(n)} + \sqrt{L\alpha \frac{V(X_{1,k}, R_{1,k}) \log(n)}{T_k(n)}},$$

$$\eta_{k,n}^{\mathtt{B}} = \frac{2\alpha M_X \log(n)}{3T_k(n)} + \sqrt{L\alpha \frac{Var(X_{1,k}) \log(n)}{T_k(n)}}.$$

Then, if $S_n^\pi < B$, i.e., there is a remaining budget, then the UCB-B1 Algorithm pulls an arm at stage $n+1$ according to:

$$I_{n+1} \in \arg\max_k \left\{\widehat{r}_{k,n} + \widehat{c}_{k,n}^{\mathtt{B1}}\right\},$$

where

$$\widehat{c}_{k,n}^{\mathtt{B}} = 1.4\frac{\epsilon_{k,n}^{\mathtt{B}} + (\widehat{r}_{k,n} - \omega_k)\eta_{k,n}^{\mathtt{B}}}{(\widehat{\mathbb{E}}_n[X_k])^+}$$

if the stability condition (7) holds for $\eta = \eta_{k,n}^{\mathtt{B}}$ and $\lambda = 1.28$, and $\widehat{c}_{k,n}^{\mathtt{B}} = \infty$ otherwise.

The regret performance of UCB-B1 is presented in the following theorem.

Theorem 1 (Regret Upper Bound for UCB-B1)**.** *Let* $\Delta_k = r^* - r_k$*,* $\lambda = 1.28$*,*

$$\sigma_k^2 = V(X_{1,k}, R_{1,k}) + (r^* - \omega_k)^2 Var(X_{1,k}), \qquad (10)$$

for all $k \in \mathbb{K}$ *and recall that* $\mu_* = \min_k \mathbb{E}[X_{1,k}]$*.*

1. ***Bounded Cost and Reward:*** *If* $|X_{1,k}| \leq M_X$*,* $|R_{1,k}| \leq M_R$ *a.s.,* $\alpha > 2$ *and* $L = 2$*, then the regret under* UCB-B1 *is upper bounded as:*

   $$Reg_{\pi^{\mathtt{B1}}}(B) \leq \alpha \sum_{k:\Delta_k>0} \log\Big(\frac{2B}{\mu_*}\Big) C_k^{\mathtt{B1}} + O(1),$$

   for some constant $\zeta > 1$ *where* $M_k = M_R + r_k M_X$ *and*

   $$C_k^{\mathtt{B1}} = \frac{42\sigma_k^2}{\Delta_k \mathbb{E}[X_{1,k}]} + 42M_k + 21M_X\Delta_k,$$

   for all k*.*

2. ***Jointly Gaussian Cost and Reward:*** *Let* $(X_{n,k}, R_{n,k})$ *be jointly Gaussian with known second-order moments. Then,* UCB-B1 *with* $\alpha > 2$*,* $M_X = M_R = 0$ *and* $L = \frac{1}{2}$ *yields the following regret bound:*

   $$Reg_{\pi^{\mathtt{B1}}}(B) \leq \alpha \sum_{k:\Delta_k>0} \log\Big(\frac{2B}{\mu_*}\Big)\frac{11\sigma_k^2}{\Delta_k\mathbb{E}[X_{1,k}]} + O(1),$$

   where σ_k *is defined in* (10)*.*

Proof. The detailed proof, which will provide basis for the analysis of other algorithms proposed in this work, can be found in Appendix C. Note that the total reward is a controlled and stopped random walk with potentially unbounded support. Thus, the regret analysis requires new methods from the theory of martingales and stopped random walks. As such, we follow a proof strategy based on establishing a high-probability upper bound for $N_\pi(B)$, which can be found in Appendix B. □

4.3 Heavy-Tailed Case: Algorithm UCB-M1

In this subsection, we design a general algorithm that achieves the regret in the sub-Gaussian case (up to a constant) under the mild moment condition that $\mathbb{E}[(X_{1,k}^+)^{2+\gamma}] < \infty$ for all k.

The empirical mean estimator played a central role in the design of the UCB-B1 Algorithm for sub-Gaussian distributions, which is proved to achieve $O(\log(B))$ regret. However, if we consider heavy-tailed distributions, the empirical mean estimator fails to achieve exponential convergence rate due to the frequent outliers (Bubeck et al., 2013). The median-based estimators, introduced in (Nemirovsky and Yudin, 1983) provide an elegant method to boost the convergence speed in mean estimation. The idea of boosting the confidence of weak independent estimators by taking the median was extended to general point estimation problems (beyond the mean estimation) in (Minsker et al., 2015). In the following, we will use a variation of this method in the design of median-based rate estimators.

Consider arm $k \in \mathbb{K}$ at stage n. For

$$m = \lfloor 3.5\alpha\log(n)\rfloor + 1,$$

we partition the observed samples $\{(X_{i,k}, R_{i,k}) : I_i = k,\ 1 \leq i \leq n\}$ into index sets $G_1, G_2, \ldots, G_m$ of size $\lfloor T_k(n)/m\rfloor$ each. Then, for each $j \in \{1, 2, \ldots, m\}$, let $\tilde{r}_{k,G_j} = \frac{\max\{\widehat{\mathbb{E}}_{G_j}[R_k], 0\}}{\max\{\widehat{\mathbb{E}}_{G_j}[X_k], b\}}$ where $b \leq \mathbb{E}[X_{1,k}]/2$, and

$$\widehat{\mathbb{E}}_{G_j}[X_k] = \sum_{i\in G_j}\frac{X_{i,k}}{|G_j|}, \qquad \widehat{\mathbb{E}}_{G_j}[R_k] = \sum_{i\in G_j}\frac{R_{i,k}}{|G_j|}.$$

The median-based rate estimator for arm k at stage n is thus

$$\overline{r}_{k,n} = \underset{1\leq j\leq m}{\mathrm{median}}\ \tilde{r}_{k,G_j}.$$

The deviations in the cost and reward are as follows:

$$\epsilon_{k,n}^{\mathtt{M}} = 11\sqrt{\alpha\frac{V(X_{1,k}, R_{1,k})\log(n)}{T_k(n)}},$$

$$\eta_{k,n}^{\mathtt{M}} = 11\sqrt{\alpha\frac{Var(X_{1,k})\log(n)}{T_k(n)}}.$$

Therefore, the decision at stage $(n+1)$ under UCB-M1 is as follows:

$$I_{n+1} \in \arg\max_k \left\{\overline{r}_{k,n} + \widehat{c}_{k,n}^{\mathtt{M}}\right\} \qquad (11)$$

where

$$\widehat{c}_{k,n}^{\mathtt{M}} = \frac{2\sqrt{2}(\epsilon_{k,n}^{\mathtt{M}} + (\overline{r}_{k,n} - \omega_k)\eta_{k,n}^{\mathtt{M}}}{\left(\underset{1\leq j\leq m}{\mathrm{median}}\ \widehat{\mathbb{E}}_{G_j}[X_k]\right)^+},$$

if the condition (7) is satisfied for $\eta = \underset{1\leq j\leq m}{\text{median}}\, \widehat{\mathbb{E}}_{G_j}[X_k]$ and $\lambda = 1.28$.

For UCB-M1, we have the following regret upper bound.

Theorem 2 (Regret Upper Bound for UCB-M1)**.** *If the following moment conditions hold:*

- $\mathbb{E}[(X_{1,k}^+)^{2+\gamma}] < \infty$*, for all* k*,*
- $Var(R_{1,k}) < \infty$*, for all* k*,*

then the regret under UCB-M1 *satisfies the following upper bound:*

$$Reg_{\pi^{\text{M1}}}(B) \leq \alpha \sum_{k:\Delta_k>0} \log\Big(\frac{2B}{\mu_*}\Big)\frac{C\sigma_k^2}{\Delta_k\mathbb{E}[X_{1,k}]} + O(1), \tag{12}$$

where σ_k *is as defined in* (10) *and* $C > 0$ *is a constant.*

Proof. The proof uses tools from the theory of martingales and stopped random walks, and can be found in Appendix B and Appendix C. □

Remark 2. We have the following observations from Theorem 1 and 2:

- If $Var(X_{1,k}) \downarrow 0$ and $\mathbb{E}[X_{1,k}] = 1$, the regret upper bounds match with the existing regret bounds for the stochastic bandit problem.
- Note that for positively correlated $X_{n,k}$ and $R_{n,k}$, one can ignore the correlation and use an upper confidence bound based on the separate estimation of $X_{n,k}$ and $R_{n,k}$. From Theorem 1, it can be observed that this scheme leads to a loss of $O\big(\sum_k Cov(X_{1,k}, R_{1,k})\big)$. Moreover, as it will be seen in the next section, this is nearly the best way of exploiting the correlation in the case of jointly Gaussian cost and reward pairs.
- The UCB-M1 Algorithm achieves the same regret upper bound as the UCB-B1 Algorithm up to a constant with much less moment assumptions: while UCB-B1 requires sub-Gaussianity, UCB-M1 requires only existence of moments of order $(2+\gamma)$ for some $\gamma > 0$ for the costs, and second-order moments for the rewards. However, the constant that multiplies the $O(\log B)$ term is much higher in UCB-M1 than UCB-B1, which can be viewed as the cost of generality.
- If the cost is deterministic, i.e., $Var(X_{1,k}) = 0$, then the regret is monotonically decreasing in Δ_k as $O\Big(\frac{\log B}{\Delta_k}\Big)$ for each arm k. However, for random costs, since $r^* = r_k + \Delta_k$, the regret bounds have an additive term scaling linearly in Δ_k as $O\Big(\log\Big(\frac{2B}{\mu_*}\Big)\sum_k \frac{Var(X_{1,k})}{\mathbb{E}[X_{1,k}]}\Delta_k\Big)$, which might seem strange at first since the separability of a sub-optimal arm k increases with its corresponding Δ_k. This is a unique phenomenon observed in the case of stochastic costs: recall from Remark 1 that the rate estimator is unstable when the confidence interval for the estimation of $\mathbb{E}[X_{1,k}]$ is large, and thus it incurs $\mathbb{E}[X_{1,k}]\Delta_k$ regret per pull since rate estimation is unreliable. As it will be seen in Corollary 1, the same term appears with the same coefficient in the regret lower bound for jointly Gaussian cost-reward pairs, which implies that it is inevitable at least in that case.

5 Regret Lower Bound for Admissible Policies

In this section, we will propose regret lower bounds for the budget-constrained bandit problem based on (Lai and Robbins, 1985). In the specific case of jointly Gaussian cost-reward pairs, we can determine a lower bound explicitly, which provides useful insight about the impact of variability and correlation on the regret.

In order to establish a regret lower bound, assume that the joint distribution of $\{(X_{n,k}, R_{n,k}) : n \geq 1\}$ is parametrized by $\theta_k \in \Theta_k$ for some parameter space Θ_k, i.e., $(X_{n,k}, R_{n,k}) \sim P_{\theta_k}$. For any $k \in \mathbb{K}$ and $\theta \in \Theta_k$, let $r_k(\theta) = \frac{\mathbb{E}_\theta[R_{1,k}]}{\mathbb{E}_\theta[X_{1,k}]}$ be the reward rate (i.e., reward per unit cost). Furthermore, for a given bandit instance $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, let $r^* = \max_k r_k(\theta_k)$ be the optimal reward rate, and $\Delta_k = r^* - r_k(\theta_k)$. For admissible policies, we have the following regret lower bound, which is an extension of Lai-Robbins style regret lower bounds for the stochastic bandit problem (Lai and Robbins, 1985; Burnetas and Katehakis, 1996).

Theorem 3 (Regret Lower Bound)**.** *Suppose that* $\mathbb{E}[(X_{1,k})^{2+\gamma}] < \infty$ *for some* $\gamma > 0$ *and* $Var(R_{1,k}) < \infty$ *hold for all* k*. Assume that the following conditions are satisfied by* $P_{k,\theta}$ *for any* k*:*

1. *If* $r_k(\theta_1) > r_k(\theta_2)$*, then* $D(P_{k,\theta_2}||P_{k,\theta_1}) < \infty$*,*
2. *(Denseness)* $r_k(\Theta_k) = \{r_k(\theta) : \theta \in \Theta_k\}$ *is dense,*
3. *(Continuity)* $\theta \mapsto D(P_{k,\theta_k}||P_{k,\theta})$ *is a continuous mapping.*

For a given bandit instance $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$*, if* $\pi \in \Pi$ *is a policy such that* $\mathbb{E}[T_k^\pi(n)] = o(n^\alpha)$ *for any* $\alpha > 0$ *and* k *such that* $r_k(\theta_k) < r^*$*, then we have the following lower bound:*

$$\liminf_{B\to\infty} \frac{Reg_\pi(B)}{\log(B)} \geq \frac{1}{2}\sum_{k:\Delta_k>0}\frac{\mathbb{E}[X_{1,k}]\Delta_k}{D_k^\star}, \tag{13}$$

where $D_k^\star$ is the solution to the following optimization problem:

$$D_k^\star = \min_{\theta\in\Theta_k} D(P_{k,\theta_k}||P_{k,\theta}) \text{ subject to } r_k(\theta) \geq r^*.$$

Proof. The proof can be found in Appendix E. □

The regret lower bound has an explicit form if the cost and reward distributions of each arm is jointly Gaussian with a known covariance matrix.

Corollary 1 (Jointly Gaussian Cost and Reward)**.** *Let $(X_{n,k}, R_{n,k})$ be jointly Gaussian:*

$$(X_{n,k}, R_{n,k}) \sim \mathcal{N}(\mu_k, \Sigma_k),$$

for all $k \in \mathbb{K}$ where $\mu_k = \big(\mathbb{E}[X_{n,k}], \mathbb{E}[R_{n,k}]\big)$ and

$$\Sigma_k = \begin{pmatrix} Var(X_{n,k}) & Cov(X_{n,k}, R_{n,k}) \\ Cov(X_{n,k}, R_{n,k}) & Var(R_{n,k}) \end{pmatrix}.$$

If Σ_k is known and μ_k is unknown by the controller for all $k \in \mathbb{K}$, we have the following regret lower bound for the Gaussian case:

$$\liminf_{B\to\infty} \frac{Reg_\pi(B)}{\log(B)} \geq \sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\mathbb{E}[X_{1,k}]\Delta_k}, \tag{14}$$

where σ_k^2 is defined in (10)*.*

Proof. For known Σ_k, we have $D_k^\star = \frac{(\mathbb{E}[X_{1,k}]\Delta_k)^2}{2\sigma_k^2}$ for $\theta_k = \mu_k$ and $\Theta_k = \mathbb{R}_+^2$. Using this in Theorem 3 yields the result. □

Remark 3 (Optimality of UCB-B1 and UCB-M1)**.** *Comparing* (2) *and* (12) *with* (14)*, we can deduce that* UCB-B1 *and* UCB-M1 *achieve optimal regret up to a universal constant for the case of jointly Gaussian cost and reward pairs with known covariance matrix.*

6 Algorithms for Unknown Second-Order Moments

In Section 4, we proposed algorithms under the assumption that the second-order moments are known for each arm k. However, in practice, these second-order moments are unknown, and therefore to be estimated from the samples collected via bandit feedback. In this section, we will propose algorithms that use these second-order moment estimates to achieve tight regret bounds.

The general strategy in the development of the algorithms in this section is to use empirical estimates for the second-order moments that appear in UCB-B1 as a surrogate.

6.1 Bounded and Uncorrelated Cost and Reward: UCB-B2

For clarity, we first consider the case $X_{n,k}$ and $R_{n,k}$ are uncorrelated for all k and $X_{n,k} \in [0, M_X]$ and $R_{n,k} \in [0, M_R]$ almost surely for known $M_X, M_R > 0$. In this case, we will propose an algorithm based on a variant of the empirical Bernstein inequality, which was introduced in (Audibert et al., 2009).

For any k, let the variance estimate $\widehat{V}_{k,n}(X_k)$ be defined as follows:

$$\widehat{V}_{k,n}(X_k) = \frac{1}{T_k(n)} \sum_{i=1}^n \mathbb{I}\{I_i = k\}\big(X_{i,k} - \widehat{\mathbb{E}}_n[X_{1,k}]\big)^2,$$

where $\widehat{\mathbb{E}}_n[X_k]$ is the empirical mean of the observations up to epoch n.

The bias terms in UCB-B2 are defined as follows:

$$\epsilon_{k,n}^{\texttt{B2}} = \sqrt{\frac{2\widehat{V}_{k,n}(R_k)\log(n^\alpha)}{T_k(n)}} + \frac{3M_R\log(n^\alpha)}{T_k(n)},$$
$$\eta_{k,n}^{\texttt{B2}} = \sqrt{\frac{2\widehat{V}_{k,n}(X_k)\log(n^\alpha)}{T_k(n)}} + \frac{3M_X\log(n^\alpha)}{T_k(n)}.$$

Let $\widehat{r}_{k,n}$ be the empirical reward rate estimator in Section 4.2, and

$$\widehat{c}_{k,n}^{\texttt{B2}} = 1.4\frac{\epsilon_{k,n}^{\texttt{B2}} + \widehat{r}_{k,n}\eta_{k,n}^{\texttt{B2}}}{\big(\widehat{\mathbb{E}}_n[X_k]\big)^+}, \tag{15}$$

if the condition (7) is satisfied with $\lambda = 1.28$ ($\widehat{c}_{k,n}^{\texttt{B2}} = \infty$ otherwise). Then, at stage $n+1$, the following decision is made under UCB-B2:

$$I_{n+1} \in \arg\max_k \Big\{\widehat{r}_{k,n} + \widehat{c}_{k,n}^{\texttt{B2}}\Big\}.$$

The lack of knowledge for the second-order statistics loosen the upper confidence bound for the rate estimator, which in turn increases the regret. In the following, we provide the regret upper bounds for UCB-B2 to gain insight about the impact of using variance estimates on the performance of the algorithm.

Theorem 4 (Regret Upper Bound for UCB-B2)**.** *Let σ_k and M_k be as defined in Theorem 1. Then, we have the following upper bound for the regret under* UCB-B2*:*

$$Reg_{\pi^{\texttt{B2}}}(B) \leq \alpha \sum_{k:\Delta_k>0} \log\Big(\frac{2B}{\mu_*}\Big)(C_k^{\texttt{B1}} + \delta C_k) + O(1), \tag{16}$$

where

$$\delta C_k = 21\Big(\frac{M_X^4\Delta_k\mu_k}{Var^2(X_{1,k})} + \frac{Var(X_{1,k})\Delta_k}{\mu_k}\Big). \tag{17}$$

for $\mu_k = \mathbb{E}[X_{1,k}]$.

The proof of Theorem 4 involves the analysis of sample variance estimates, and can be found in Appendix F.

Remark 4 (Impact of Unknown Variances)**.** The additional terms are caused by the stability of the rate estimator: since we use a variance estimate in the upper confidence bound of $X_{n,k}$, the rate estimator suffers from a longer period of instability, which increases the regret coefficient proportional to Δ_k.

6.2 Learning the Correlation: `UCB-B2C`

Finally we consider the case $(X_{n,k}, R_{n,k})$ are bounded and correlated, but the second-order moments are unknown. In the absence of correlation, our goal was to estimate $Var(R_{1,k})$ and $Var(X_{1,k})$ from the samples of $(X_{n,k}, R_{n,k})$. When there is a correlation, we have an optimization problem: we need to establish confidence bounds for the LMMSE estimator ω_k defined in (9) as well as the minimum variance $Var(R_{1,k} - \omega_k X_{1,k})$ by using the samples of $(X_{n,k}, R_{n,k})$ observed via bandit feedback. We take a loss minimization approach in the statistical learning setting to estimate these quantities.

For any $k \in \mathbb{K}$, let the empirical LMMSE estimator be defined as follows:

$$\widehat{\omega}_{k,n} = \arg\min_{\omega' \in \mathbb{R}} \ \widehat{L}_{k,n}(\omega)$$

where the empirical loss function is the following:

$$\widehat{L}_{k,n}(\omega) = \sum_{i=1}^{n} \frac{\mathbb{I}\{I_i = k\}}{T_k(n)} \Big(R_i - \widehat{\mathbb{E}}_n[R] - \omega\big(X_i - \widehat{\mathbb{E}}_n[X]\big)\Big)^2.$$

It can be shown that $\widehat{\omega}_{k,n} \to \omega_k$ if $T_k(n) \to \infty$ as $n \to \infty$, and moreover the convergence rate is exponential and tight concentration bounds for $\widehat{\omega}_{k,n}$ and $\widehat{L}_{k,n}(\widehat{\omega}_{k,n})$ can be established. Let $M_Z = M_R + \overline{\omega} M_X$ where $\overline{\omega} > \max_k \ \omega_k$ is a given parameter, and let

$$\nu_{k,n}(\omega_k) = \frac{1.36 M_X M_Z}{Var(X_{1,k})} \sqrt{\frac{\log n^{\alpha}}{T_k(n)}}, \qquad (18)$$

$$\nu_{k,n}(L_k) = M_Z^2 \sqrt{\frac{2 \log n^{\alpha}}{T_k(n)}}. \qquad (19)$$

Then, it can be shown that $-\widehat{\omega}_{k,n} + \nu_{k,n}(\omega_k)$ and $\widehat{L}_{k,n}(\widehat{\omega}_{k,n}) + \nu_{k,n}(\omega_k)$ are high-probability upper bounds for $-\omega_k$ and $\min_\omega \ Var(R_{1,k} - \omega X_{1,k})$, respectively, for large enough $T_k(n)$.

The bias terms in `UCB-B2C` are defined as follows:

$$\epsilon_{k,n}^{\texttt{B2C}} = \sqrt{\frac{2\widehat{L}_{k,n}(\widehat{\omega}_{k,n}) \log(n^{\alpha})}{T_k(n)}} + \frac{3M_Z \log(n^{\alpha})}{T_k(n)},$$

$$\eta_{k,n}^{\texttt{B2C}} = \sqrt{\frac{2\widehat{V}_{k,n}(X_k) \log(n^{\alpha})}{T_k(n)}} + \frac{3M_X \log(n^{\alpha})}{T_k(n)}.$$

Then, at stage $n+1$, the following decision is made under `UCB-B2C`:

$$I_{n+1} \in \arg\max_k \ \Big\{\widehat{r}_{k,n} + \widehat{c}_{k,n}^{\texttt{B2C}}\Big\},$$

where

$$\widehat{c}_{k,n}^{\texttt{B2C}} = 1.4 \frac{\epsilon_{k,n}^{\texttt{B2C}} + (\widehat{r}_{k,n} - \widehat{\omega}_{k,n})^+ \eta_{k,n}^{\texttt{B2C}}}{\big(\widehat{\mathbb{E}}_n[X_k]\big)^+},$$

if the stability condition (7) is satisfied with $\lambda = 1.28$, and $\widehat{c}_{k,n}^{\texttt{B2C}} = \infty$ otherwise.

In the following, we investigate the impact of using second-order moment estimates on the regret of `UCB-B2C`. The proof can be found in Appendix G.

Theorem 5 (Regret Upper Bound for `UCB-B2C`)**.** *Let $C_k^{\texttt{B1}}$ be defined as in Theorem 1. Then, we have the following upper bound for the regret under* `UCB-B2`*:*

$$Reg_{\pi^{\texttt{B2C}}}(B) \leq \alpha \sum_{k:\Delta_k > 0} \log\Big(\frac{2B}{\mu_*}\Big)(C_k^{\texttt{B1}} + \delta C_k') + O(1),$$

where

$$\delta C_k' = \delta C_k + 42\Big(\frac{M_Z M_X}{\sqrt{Var(X_{1,k})}} + \frac{M_X^4 \Delta_k \mu_k}{Var^2(X_{1,k})}\Big). \qquad (20)$$

for $\mu_k = \mathbb{E}[X_{1,k}]$ and δC_k defined in (17)*.*

Note that the regret of `UCB-B2C` converges to the regret of `UCB-B2`, and they both approach to the performance of the `UCB-B1` Algorithm as $\Delta_k \downarrow 0$.

7 Conclusions

In this paper, we considered a very general setting for the budgeted bandit problem where each action incurs a potentially correlated and heavy-tailed cost-reward pair. We proved that positive expected cost and existence of moments of order $2+\gamma$ for some $\gamma > 0$ suffice for $O(\log B)$ regret for a given budget $B > 0$. For known second-order moments, we proposed two algorithms named `UCB-B1` and `UCB-M1` that exploit the correlation between cost and reward by using an LMMSE estimator. By proposing a regret lower bound, we proved that `UCB-B1` and `UCB-M1` achieve order optimality, and moreover they achieve optimal regret up to a universal constant for the specific case of jointly Gaussian cost and reward pairs, which underlines the significance of second-order moments and correlation in the regret performance. For the case of bounded cost and reward with unknown second-order moments, we proposed learning algorithms `UCB-B2` and `UCB-B2C` that estimate variances as well as LMMSE estimator to approach the performance of `UCB-B1`. We investigated the effect of using these estimates as surrogates in the absence of second-order moments, and showed that they approach the performance of `UCB-B1` in certain cases.

Acknowledgements

This research was supported in part by the NSF grants: CNS-NeTS-1514260, CNS-NeTS-1717045, CMMI-SMOR-1562065, CNS-ICN-WEN-1719371, NSF CCF 1934986, NSF NeTS 1718203, and CNS-SpecEES-1824337; ONR grants: ONR N00014-19-1-2621, ONR N00014-19-1-2566; the DTRA grant HDTRA1-18-1-0050; ARO W911NF-19-1-0379.

References

S. Agrawal and N. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*, pages 3450–3458, 2016.

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

S. Asmussen. *Applied probability and queues*, volume 51. Springer Science & Business Media, 2008.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134, 2014.

D. A. Berry and B. Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87, 1985.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

D. L. Burkholder. Distribution function inequalities for martingales. *the Annals of Probability*, pages 19–42, 1973.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

S. Cayci, A. Eryilmaz, and R. Srikant. Learning to control renewal processes with bandit feedback. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):43, 2019.

R. Combes, C. Jiang, and R. Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):245–257, 2015.

B. C. Dean, M. X. Goemans, and J. Vondrdk. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 208–217. IEEE, 2004.

W. Ding, T. Qin, X.-D. Zhang, and T.-Y. Liu. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

S. Guha and K. Munagala. Multi-armed bandits with metric switching costs. In *International Colloquium on Automata, Languages, and Programming*, pages 496–507. Springer, 2009.

A. Gut. *Stopped random walks*. Springer, 2009.

A. György, L. Kocsis, I. Szabó, and C. Szepesvári. Continuous time associative bandit problems. In *IJCAI*, pages 830–835, 2007.

M. Harchol-Balter. Task assignment with unknown duration. In *Proceedings 20th IEEE International Conference on Distributed Computing Systems*, pages 214–224. IEEE, 2000.

P. R. Jelenković and J. Tan. Characterizing heavy-tailed distributions induced by retransmissions. *Advances in Applied Probability*, 45(1):106–138, 2013.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

S. Lalley and G. Lorden. A control problem arising in the sequential design of experiments. *Annals of probability*, 14(1):136–172, 1986.

K. Liu and Q. Zhao. Multi-armed bandit problems with heavy-tailed reward distributions. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 485–492. IEEE, 2011.

S. Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

H. V. Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

K. A. Sankararaman and A. Slivkins. Combinatorial semi-bandits with knapsacks. *arXiv preprint arXiv:1705.08110*, 2017.

D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Y. Xia, H. Li, T. Qin, N. Yu, and T.-Y. Liu. Thompson sampling for budgeted multi-armed bandits. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Y. Xia, W. Ding, X.-D. Zhang, N. Yu, and T. Qin. Budgeted bandit problems with continuous random costs. In *Asian conference on machine learning*, pages 317–332, 2016.

A Proof of Proposition 1

Proof. The proof consists of two parts.

1. In the first part, we find an upper bound for $\mathbb{E}[\texttt{REW}_{\pi^{\texttt{opt}}(B)}(B)]$. In order to achieve this goal, we consider an arbitrary admissible algorithm $\pi \in \Pi$. Since π is admissible, we have the following relationship:

$$\mathbb{E}[R_{n,I_n^\pi}|\mathcal{F}_{n-1}^\pi] = r_{I_n}\mathbb{E}[X_{n,I_n^\pi}|\mathcal{F}_{n-1}^\pi]. \tag{21}$$

Let $W_t^\pi = \max\limits_{1\leq i\leq t} S_i^\pi$ for any $t > 0$. Then, inspired by the proof of Wald's equation (see Siegmund (2013); Xia et al. (2015)), we have the following inequality for the expected cumulative reward under π:

$$\begin{aligned}
\mathbb{E}[\texttt{REW}_\pi(B)] &= \mathbb{E}\Big[\sum_{i=1}^\infty \mathbb{I}\{W_{i-1}^\pi \leq B\}R_{i,I_i^\pi}\Big], \\
&= \mathbb{E}\Big[\sum_{i=1}^\infty \mathbb{E}\big[R_{i,I_i^\pi}|\mathcal{F}_{i-1}^\pi\big]\mathbb{I}\{W_{i-1}^\pi \leq B\}\Big], &(22)\\
&= \mathbb{E}\Big[\sum_{i=1}^\infty r_{I_i^\pi}\mathbb{E}\big[X_{i,I_i^\pi}|\mathcal{F}_{i-1}^\pi\big]\mathbb{I}\{W_{i-1}^\pi \leq B\}\Big], &(23)\\
&\leq r^*\mathbb{E}\Big[\sum_{i=1}^{N_\pi(B)} X_{i,I_i^\pi}\Big] = r^*\mathbb{E}\Big[S_{N_\pi(B)}^\pi\Big], &(24)
\end{aligned}$$

where (22) follows since π is admissible and $W_{i-1}^\pi \in \mathcal{F}_{i-1}$, and (23) follows from the relation (21) and the fact that $r_{I_i} \leq r^*$ with probability 1.

Note that $S_{N_\pi(B)}^\pi$ is a controlled random walk whose increments X_{i,I_i^π} are dependent. Therefore, classical second-order moment results in renewal theory, such as Lorden's inequality (Asmussen, 2008), are not directly applicable to provide an upper bound for $\mathbb{E}[S_{N_\pi(B)}^\pi]$. Instead, the following result for the first passage times of submartingales yields a tight upper bound for $\mathbb{E}[S_{N_\pi(B)}^\pi]$.

Proposition 3 (Lalley and Lorden (1986))**.** *Consider a stochastic process $\{(U_n) : n \geq 1\}$ with $\mathbb{E}[U_n] > 0$ adapted to the filtration $\mathcal{F}_n$. Let $S_n = \sum_{i=1}^n U_i$ with $S_0 = 0$ and $N(a) = \inf\{n : S_n > a\}$ be the first passage time of the random walk.*

Assume that there exists constants $\mu_*, \mu^*, \sigma^2 > 0$ such that

$$0 < \mu_* \leq \mathbb{E}[U_n|\mathcal{F}_{n-1}] \leq \mu^* < \infty,$$

and

$$Var(U_n|\mathcal{F}_{n-1}) \leq \sigma^2 < \infty,$$

with probability 1 for all $n \geq 1$. If there exists $\gamma > 0$ such that $\mathbb{E}[(U_n^+)^{2+\gamma}] < \infty$, then there exists a constant $G = G(\mu_*, \mu^*, \sigma^2)$ such that the following holds:

$$\mathbb{E}[S_{N(a)}] - a \leq G,$$

for any $a > 0$.

Note that we have

$$0 < \min_{k\in[K]} \mathbb{E}[X_{1,k}] \leq \mathbb{E}[X_{i,I_i^\pi}|\mathcal{F}_{i-1}] \leq \max_{k\in[K]} \mathbb{E}[X_{1,k}] < \infty,$$

and

$$Var(X_{i,I_i^\pi}|\mathcal{F}_{i-1}) \leq \max_{k\in[K]} Var(X_{1,k}) < \infty,$$

with probability 1 for all $i \geq 1$. Thus, under Assumption 1, Proposition 3 implies that there exists a constant $G > 0$ such that the following holds:

$$\mathbb{E}[S_{N_\pi(B)}^\pi] \leq B + G, \tag{25}$$

for all $B > 0$. Hence, (24) and (25) together imply the following upper bound:

$$\mathbb{E}[\mathtt{REW}_\pi(B)] \leq r^*(B+G), \tag{26}$$

for all $B > 0$ and any admissible policy $\pi \in \Pi$. Since the inequality (26) holds for any admissible $\pi \in \Pi$, we have the following result:

$$\mathbb{E}[\mathtt{REW}_{\pi^{opt}(B)}(B)] \leq r^*(B+G),\ \forall B > 0. \tag{27}$$

2. In the second part of the proof, we will find a lower bound for $\mathbb{E}[\mathtt{REW}_{\pi^*}(B)]$. Since π^* is a static policy and $N_{\pi^*}(B)$ is a stopping time, Wald's equation implies the following result Siegmund (2013):

$$\mathbb{E}[\mathtt{REW}_{\pi^*}(B)] = \mathbb{E}\big[R_{1,k^*}\big]\mathbb{E}\big[N_{\pi^*}(B)\big]. \tag{28}$$

For random walks with positive drift, the following inequality holds for any $B > 0$ Asmussen (2008); Gut (2009):

$$\mathbb{E}[N_{\pi^*}(B)] \geq \frac{B}{\mathbb{E}[X_{1,k^*}]}. \tag{29}$$

(28) and (29) together imply the following:

$$\mathbb{E}[\mathtt{REW}_{\pi^*}(B)] \geq r^*B,\ \forall B > 0. \tag{30}$$

Inequalities in (27) and (30) together imply that the optimality gap is bounded by a constant $G^\star = r^*G$ for all $B > 0$. □

Proposition 1 has a striking implication: the optimality gap is still bounded for unbounded and correlated cost and reward pairs, and this result requires only a mild moment assumption that $\mathbb{E}[(X_{1,k}^+)^{2+\gamma}],\ k \in [K]$ exists for some $\gamma > 0$. Therefore, the simple policy π^* serves as a plausible substitute for $\pi^{\mathtt{opt}}(B)$, which is NP-hard, for learning purposes.

B A Useful Upper Bound for Regret

The number of trials $N_\pi(B)$ under an admissible policy π is a random stopping time, which makes the regret computations difficult. The following proposition, which extends the strategy in (Xia et al., 2016) to the case of unbounded and potentially heavy-tailed cost-reward pairs that can take on negative values, provides a useful tool for regret computations.

Proposition 4 (Regret Upper Bounds for Admissible Policies)**.** *Suppose that*

$$\max_k\ \mathbb{E}[|X_{1,k} - \mathbb{E}[X_{1,k}]|^p] = u_{max} < \infty,$$

for some $p > 2$. Let $T_k(n)$ be the number of pulls for arm k in n trials, and $\mu_* = \min_k \mathbb{E}[X_{1,k}]$. The following upper bound holds for any admissible policy $\pi \in \Pi$ and $B > \mu_*/2$:

$$Reg_\pi(B) \leq \sum_k \mathbb{E}\Big[T_k\Big(\frac{2B}{\mu_*}\Big)\Big]\Delta_k\mathbb{E}[X_{1,k}] + \frac{\big(\frac{2p^2}{p-1}\big)^p u_{max}}{(2B-\mu_*)^{\frac{p}{2}}\mu_*^{\frac{p}{2}}(\frac{p}{2}-1)}\sum_k \Delta_k\mathbb{E}[X_{1,k}] + G^\star, \tag{31}$$

where $G^\star = G^\star(\mu_*, \sigma^2_{max})$ is a constant.

The proof of Proposition 4 relies on a variant of Chebyshev inequality for controlled random walks. Note that $2B/\mu_*$ is a high-probability upper bound for the total number of pulls $N_\pi(B)$, and $\Delta_k\mathbb{E}[X_{1,k}]$ is the average regret per pull for a suboptimal arm k. Proposition 4 implies that the expected regret after $2B/\mu^*$ pulls is $O(1)$.

Proof of Proposition 4. Take an arbitrary admissible policy $\pi \in \Pi$. The regret can be decomposed as follows:

$$Reg_\pi(B) = \underbrace{\mathbb{E}[\mathtt{REW}_{\pi^{opt}(B)}(B)] - \mathbb{E}[\mathtt{REW}_{\pi^*}(B)]}_{(a)} + \underbrace{\mathbb{E}[\mathtt{REW}_{\pi^*}(B)] - \mathbb{E}[\mathtt{REW}_\pi(B)]}_{(b)}. \tag{32}$$

Note that (a) in (31) is the optimality gap for π^*, which is upper bounded by a constant $G^\star = r^*G$ by Proposition 1. In the following, we provide an upper bound for (b) in (31).

First, note that the cumulative reward under π^* is upper bounded as follows:

$$\begin{aligned}\mathbb{E}[\texttt{REW}_{\pi^*}(B)] &= \mathbb{E}[N_{\pi^*}(B)]\cdot\mathbb{E}[R_{1,k^*}],\\ &\leq Br^* + r^*\frac{\mathbb{E}[X_{1,k^*}^2]}{\mathbb{E}[X_{1,k^*}]} = Br^* + c,\end{aligned}\tag{33}$$

where the first line follows from Wald's equation and the second line is a consequence of Lorden's inequality Asmussen (2008). Since $B \leq \sum_{i=1}^{N_\pi(B)} X_{i,I_i^\pi}$ under π, we can further upper bound $\mathbb{E}[\texttt{REW}_{\pi^*}(B)]$ as follows:

$$\begin{aligned}\mathbb{E}[\texttt{REW}_{\pi^*}(B)] &\leq \mathbb{E}\Big[\sum_{i=1}^{N_\pi(B)} r^* X_{i,I_i^\pi}\Big] + r^*\frac{\mathbb{E}[X_{1,k^*}^2]}{\mathbb{E}[X_{1,k^*}]},\\ &= \mathbb{E}\Big[\sum_k\sum_{i=1}^{\infty}\mathbb{I}\{W_{i-1}^\pi \leq B\}\mathbb{I}\{I_i^\pi = k\}r^*\mathbb{E}[X_{i,k}]\Big] + c.\end{aligned}\tag{34}$$

where

$$W_n^\pi = \max\{S_1^\pi, S_2^\pi, \ldots, S_n^\pi\}.$$

Similar to the proof of Proposition 1, we have the following equation for $\mathbb{E}[\texttt{REW}_\pi(B)]$:

$$\begin{aligned}\mathbb{E}[\texttt{REW}_\pi(B)] &= \mathbb{E}\Big[\sum_{i=1}^{N_\pi(B)} R_{i,I_i^\pi}\Big],\\ &= \mathbb{E}\Big[\sum_k\sum_{i=1}^{\infty}\mathbb{I}\{W_{i-1}^\pi \leq B\}\mathbb{I}\{I_i^\pi = k\}r_k\mathbb{E}[X_{i,k}]\Big]\end{aligned}\tag{35}$$

From (34) and (35), we have the following upper bound for (b) in (31):

$$\mathbb{E}[\texttt{REW}_{\pi^*}(B)] - \mathbb{E}[\texttt{REW}_\pi(B)] \leq \mathbb{E}\Big[\sum_k\sum_{i=1}^{\infty}\mathbb{I}\{W_{i-1}^\pi \leq B\}\mathbb{I}\{I_i^\pi = k\}\Delta_k\mathbb{E}[X_{i,k}]\Big] + c.\tag{36}$$

For any integer $n_0 > 1$, the RHS of (36) can be upper bounded as follows:

$$\begin{aligned}\mathbb{E}[\texttt{REW}_{\pi^*}(B)] - \mathbb{E}[\texttt{REW}_\pi(B)] &\leq \mathbb{E}\Big[\sum_{i=1}^{n_0}\sum_k\mathbb{I}\{I_i^\pi = k\}\Delta_k\mathbb{E}[X_{i,k}]\Big]\\ &\quad + \mathbb{E}\Big[\sum_{i>n_0}\mathbb{I}\{W_{i-1}^\pi \leq B\}\sum_k\Delta_k\mathbb{E}[X_{i,k}]\Big] + c,\\ &= \sum_k\mathbb{E}[T_k^\pi(n_0)]\Delta_k\mathbb{E}[X_{1,k}]\\ &\quad + \big(\sum_k\Delta_k\mathbb{E}[X_{i,k}]\big)\sum_{i>n_0}\mathbb{P}\Big(W_{i-1}^\pi \leq B\Big) + c.\end{aligned}\tag{37}$$

The following martingale-based concentration inequality will be crucial in finding a tight upper bound for the crossing probability of the controlled process W_n^π in (37).

Lemma 1 (Chebyshev Inequality for Submartingales). *Let $\{Z_n : n \geq 0\}$ be a stochastic process adapted to the filtration $\mathcal{F}_n$ such that there exists a pair (μ, u) satisfying*

$$\begin{aligned}\mathbb{E}[Z_n|\mathcal{F}_{n-1}] &\geq \mu > 0,\\ \mathbb{E}\Big[\big|Z_n - \mathbb{E}[Z_n|\mathcal{F}_{n-1}]\big|^p|\mathcal{F}_{n-1}\Big] &\leq u < \infty,\end{aligned}\tag{38}$$

almost surely for all $n \geq 1$ *for* $p > 2$*. Let* $S_n = \sum_{i=1}^{n} Z_i$ *and* $W_n = \max_{1 \leq i \leq n} S_i$*. For a given* $B > 0$*, let* $n_0 = \lceil \frac{2B}{\mu} \rceil$*. Then we have the following inequality:*

$$\mathbb{P}(W_{n_0+j} \leq B) \leq \frac{\left(\frac{2p^2}{p-1}\right)^p u}{\mu^p (n_0+j)^{p/2}}. \tag{39}$$

for all $j \geq 0$*.*

Under an admissible policy π, the increments X_{i,I_i^π} of the controlled random walk S_n^π satisfy $\mathbb{E}[X_{i,I_i^\pi}|\mathcal{F}_{i-1}] \geq \mu_*$ and $\mathbb{E}\Big[\Big|X_{i,I_i^\pi} - \mathbb{E}[X_{i,I_i^\pi}|\mathcal{F}_{i-1}]\Big|^p\Big|\mathcal{F}_{i-1}\Big] \leq u_{max}$ almost surely for all i. Therefore, the conditions in (38) are satisfied, and we have:

$$\mathbb{P}(W_{n_0+j}^\pi \leq B) \leq \frac{\left(\frac{2p^2}{p-1}\right)^p u_{max}}{(2B-\mu_*)^{p/2} \mu_*^{p/2} (n_0+j)^{p/2}}. \tag{40}$$

for $n_0 = 2B/\mu_*, k \geq 1$ and $j \geq 0$. Thus, for $B > \mu_*/2$,

$$\begin{aligned}\sum_{i>n_0} \mathbb{P}(W_{i-1}^\pi \leq B) &= \sum_{j=0}^{\infty} \mathbb{P}(W_{n_0+j}^\pi \leq B), \\ &\leq \frac{\left(\frac{2p^2}{p-1}\right)^p u_{max}}{(2B-\mu_*)^{p/2} \mu_*^{p/2} (p/2-1)}.\end{aligned} \tag{41}$$

Substituting $n_0 = \frac{2B}{\mu_*}$ and (41) into (37) completes the proof. □

B.1 Proof of Lemma 1

Let $Y_i = Z_i - \mathbb{E}[Z_i|\mathcal{F}_{i-1}]$ and $M_n = \sum_{i=1} Y_i$, and note that M_n is a martingale. By the assumption (38), $\mu \leq \mathbb{E}[Z_i|\mathcal{F}_{i-1}]$ holds almost surely for all $i \geq 1$. Therefore, the following relation holds:

$$\big\{W_n \leq B\big\} \subset \big\{S_n \leq B\big\} \subset \big\{M_n \leq B - n\mu\big\}. \tag{42}$$

Let $n_0 = \frac{2B}{\mu}$. Then, for any $j \geq 0$, we have the following inequality:

$$\begin{aligned}\mathbb{P}(W_{n_0+j} \leq B) &\leq \mathbb{P}(M_{n_0+j} \leq -\frac{\mu}{2}(n_0+j)), \\ &\leq \mathbb{P}\Big(\max_{1 \leq i \leq n_0+j} |M_i| > \frac{\mu}{2}(n_0+j)\Big), \\ &\leq \frac{2^p \mathbb{E}\Big[\big(\max_{1 \leq i \leq n_0+j} |M_i|\big)^p\Big]}{\mu^p (n_0+j)^p}.\end{aligned}$$

Then, by L^p maximum inequality for martingales (Theorem 4.4.4 in (Durrett, 2019)), we have:

$$\mathbb{E}\Big[\big(\max_{1 \leq i \leq n_0+j} |M_i|\big)^p\Big] \leq \Big(\frac{p}{p-1}\Big)^p \mathbb{E}[|M_{n_0+j}|^p]. \tag{43}$$

For the martingale M_n with increments $\{Y_n : n \geq 1\}$, let $Q_n = Y_1^2 + Y_2^2 \ldots + Y_n^2$ be the quadratic variation process. It is interesting to note that M_n and $\sqrt{Q_n}$ increase at the same rate in terms of $\mathcal{L}_p$-norm (Burkholder, 1973):

$$c_p \mathbb{E}[|Q_n|^{\frac{p}{2}}] \leq \mathbb{E}[|M_n|^p] \leq C_p \mathbb{E}[|Q_n|^{\frac{p}{2}}], \tag{44}$$

where $C_p \leq p^p$ and $c_p = 1/C_p$. By Hölder's inequality, we have the following result for all $i > 0$:

$$\mathbb{E}[|M_n|^p] \leq C_p n^{\frac{p}{2}-1} \mathbb{E}[\sum_{i=1}^{n} |Y_i|^p],$$

for all $n > 0$. Given (38), the following holds:

$$\mathbb{E}[|Y_i|^p] = \mathbb{E}\big[\mathbb{E}[|Y_i|^p \big| \mathcal{F}_{i-1}]\big], \tag{45}$$
$$\leq u, \tag{46}$$

for any $i \geq 1$. Therefore, we have:

$$\mathbb{P}(W_{n_0+j} \leq B) \leq \frac{\left(\frac{2p^2}{p-1}\right)^p u}{\mu^p (n_0+j)^{p/2}}. \tag{47}$$

C Proof of Theorem 1

Proof. The regret decomposition in Proposition 4 will be used for the proof. Note that we need to find the expected number of pulls, $\mathbb{E}[T_k(n)]$, for each arm k with $r_k < r^*$. The following proposition yields an upper bound for $\mathbb{E}[T_k(n)]$ for any $n > 0$.

Lemma 2. *Let $\Delta_k = r^* - r_k$ be the reward rate discrepancy and*

$$\sigma_k^2 = \begin{cases} Var(R_{1,k}) - \omega_k^2 Var(X_{1,k}) + (r^* - \omega_k)^2 Var(X_{1,k}), & Var(X_{1,k}) \neq 0, \\ Var(R_{1,k}), & Var(X_{1,k}) = 0, \end{cases} \tag{48}$$

for all $k \in \mathbb{K}$, and recall that $\mu_* = \min_k \ \mathbb{E}[X_{1,k}]$. Then we have the following upper bounds for $\mathbb{E}[T_k(n)]$, the expected number of pulls for arm k in n stages.

1. ***Bounded Cost and Reward:*** *If $\Delta_k > 0$ and $|X_{1,k}| \leq M_X$, $|R_{1,k}| \leq M_R$ a.s., then we have the following upper bound under* `UCB-B1` *with $\alpha > 2$ and $L = 2$:*

$$\mathbb{E}[T_k(n)] \leq 42 \log(n^\alpha)\Big(\frac{\sigma_k^2}{\Delta_k^2(\mathbb{E}[X_{1,k}])^2} + \frac{M_k}{\Delta_k \mathbb{E}[X_{1,k}]} + \frac{M_X}{\mathbb{E}[X_{1,k}]}\Big) + 12\frac{\alpha}{\alpha-2}, \tag{49}$$

   where $M_k = M_R + r_k M_X$.

2. ***Jointly Gaussian Cost and Reward:*** *Let $(X_{n,k}, R_{n,k})$ be jointly Gaussian with covariance matrix Σ_k for all k. Then,* `UCB-B1` *with $\alpha > 2$, $M_X = M_R = 0$ and $L = \frac{1}{2}$ yields the following:*

$$\mathbb{E}[T_k(n)] \leq 11 \log(n^\alpha)\frac{\sigma_k^2}{\Delta_k^2(\mathbb{E}[X_{1,k}])^2} + 12\frac{\alpha}{\alpha-2}. \tag{50}$$

The proof then follows from substituting $\mathbb{E}[T_k(n)]$ in (49) (or (50) for the Gaussian case) into (31). □

In the rest of this section, we prove Lemma 2.

C.1 Proof of Lemma 2

Consider a suboptimal arm k with $\Delta_k > 0$ and a given $n > 0$. For any $t < n$, let

$$\hat{c}_{k,t} = \frac{\lambda}{2-\lambda}\frac{\epsilon_{k,n}^{\mathtt{B}} + (\widehat{r}_{k,n} - \omega_k)\eta_{k,n}^{\mathtt{B}}}{\big(\widehat{\mathbb{E}}_n[X_k]\big)^+},$$

and

$$c_{k,t} = \frac{\lambda}{\mathbb{E}[X_{1,k}]}\Big(\frac{2M_k \log(n^\alpha)}{3T_k(t)} + \sqrt{\frac{L\log(n^\alpha)\sigma^2}{T_k(t)}}\Big), \tag{51}$$

where $\sigma^2 = \sqrt{V(X_{1,k}, R_{1,k})} + (r_k - \omega_k)\sqrt{Var(X_{1,k})}$ and $\lambda = 1.28$.

We have the following claim based on (Audibert et al., 2009).

Claim 1. *Given $n > 0$, for any $t < n$, if $I_{t+1} = k$ holds, at least one of the following must be true:*

- $E_1 = \{\hat{r}_{k^*,t} + \widehat{c}_{k^*,t} \leq r^*\}$,
- $E_2 = \{\hat{r}_{k,t} > r_k + \widehat{c}_{k,t}\}$,
- $E_3 = \{T_k(t) \leq L\Big(\frac{2\lambda^2}{2-\lambda}\Big)^2\Big(\frac{2\sigma_k^2}{\big(\Delta_k\mathbb{E}[X_{1,k}]\big)^2} + \frac{M_r}{\Delta_k\mathbb{E}[X_{1,k}]}\Big)\log(n^\alpha)\}$,
- $E_4 = \{T_k(t) \leq L\big(\frac{\lambda}{\lambda-1}\big)^2\Big(\frac{Var(X_{1,k})}{\big(\mathbb{E}[X_{1,k}]\big)^2} + \frac{M_X}{\mathbb{E}[X_{1,k}]}\Big)\log(n^\alpha)\}$,

Proof. For notational convenience, let $s = T_k(t)$ and $\ell = \log(n^\alpha)$. Suppose to the contrary that neither holds. Then, we have:

$$E_4^c \subset \{\frac{2M_X\ell}{3s} + \sqrt{\frac{LVar(X_{1,k})\ell}{s}} \leq \mathbb{E}[X_{1,k}]\frac{(\lambda-1)}{\lambda}\}, \tag{52}$$

which implies that the rate estimator is stable, thus the concentration inequality in Proposition 2 holds. In order to see (52), let $x = \frac{\lambda}{\lambda-1}$, $\mu_k = \mathbb{E}[X_{1,k}]$ and

$$u = Lx^2\Big(\frac{Var(X_{1,k})}{\big(\mathbb{E}[X_{1,k}]\big)^2} + \frac{M_X}{\mathbb{E}[X_{1,k}]}\Big)\ell. \tag{53}$$

Then, for any $s \geq u$, we have the following:

$$\frac{2M_X\mu_k^2}{6x^2\big(M_X\mu_k + Var(X_{1,k})} + \frac{1}{x}\sqrt{\frac{Var(X_{1,k})\mu_k^2}{Var(X_{1,k}) + M_X\mu_k}} \leq \frac{\mu_k}{x},$$

since $x > 1$ and $\frac{1-\beta}{3x} + \sqrt{\beta} \leq 1$ for $\beta = \frac{Var(X_{1,k})}{Var(X_{1,k})+M_X\mu_k} \in [0,1]$.

Second, for large t, we have the following relation:

$$E_4^c \cap E_3^c \subset \{\widehat{c}_{k,t} \leq \frac{\Delta_k}{2}\}. \tag{54}$$

with high probability. In order to prove (54), note that the following holds:

$$c_{k,t} \leq \widehat{c}_{k,t} \leq \frac{\lambda}{2-\lambda}c_{k,t}, \tag{55}$$

with high probability under the event E_4^c. Let

$$v = L\Big(\frac{2\lambda^2}{2-\lambda}\Big)^2\Big(\frac{2\sigma_k^2}{\Delta_k^2\mu_k^2} + \frac{M_r}{\Delta_k\mu_k}\Big)\ell, \tag{56}$$

and note that $\sigma^2 \leq 2\sigma_k^2$ by Cauchy-Schwarz inequality. Then, by (55), for any $s \geq v$, we have:

$$\begin{aligned}\widehat{c}_{k,t} &\leq \frac{\Delta_k}{2}\Big(\frac{M_r\Delta_k\mu_k}{12\lambda\big(2\sigma_k^2 + M_r\Delta_k\mu_k\big)} + \sqrt{\frac{2\sigma_k^2}{2\sigma_k^2 + M_r\Delta_k\mu_k}}\Big),\\ &\leq \frac{\Delta_k}{2},\end{aligned}$$

where the last line holds since $\frac{1-\beta}{12\lambda} + \sqrt{\beta} \leq 1$ for $\lambda > 1$ and $\beta = \frac{2\sigma_k^2}{2\sigma_k^2+M_r\Delta_k\mu_k} \in [0,1]$. Since the concentration inequality holds and $E_4^c \cap E_3^c \subset \{\widehat{c}_{k,t} \leq \Delta_k/2\}$, we have:

$$\bigcap_{i=1}^4 E_i^c \subset \Big\{\widehat{r}_{k,t} + \widehat{c}_{k,t} \leq \widehat{r}_{k^*,t} + \widehat{c}_{k^*,t}\Big\},$$

which implies that $I_{t+1} = k^* \neq k$. $\square$

In order to bound $\mathbb{P}(E_1 \cup E_2)$, let $Z_{n,k} = R_{n,k} - \omega_k X_{n,k}$ and

$$\epsilon_{k,t} = \frac{2M_Z\ell}{3s} + \sqrt{L\frac{V(X_{1,k}, R_{1,k})\ell}{s}},$$
$$\eta_{k,t} = \frac{2M_X\ell}{3s} + \sqrt{L\frac{Var(X_{1,k})\ell}{s}},$$

where $M_Z = M_R + \omega_k M_Z$. Then, the following inequality based on Proposition 2 will be used:

$$\mathbb{P}(|\widehat{r}_{k,t} - r_k| > c_{k,t}) = \mathbb{P}\big(\Big|\frac{\widehat{\mathbb{E}}_t[Z_k]}{\widehat{\mathbb{E}}_t[X_k]} - \frac{\mathbb{E}[Z_k]}{\mathbb{E}[X_k]}\Big| > c_{k,t}\big),$$
$$\leq \mathbb{P}\Big(\Big|\widehat{\mathbb{E}}_t[Z_k] - \mathbb{E}[Z_k]\Big| > \epsilon_{k,t}\Big) + \mathbb{P}\Big(\Big|\widehat{\mathbb{E}}_t[X_k] - \mathbb{E}[X_k]\Big| > \eta_{k,t}\Big).$$

Note that for sub-Gaussian cost and reward pairs, $M_X = M_R = 0$ and $L = 1/2$ yields Hoeffding's inequality. For the specific case of bounded cost and reward pairs with bounds M_X and M_R, respectively, $L = 2$ leads to Bernstein's inequality. Using this concentration inequality with (55), we have the following:

$$|\widehat{r}_{k,t} - r_k| > \widehat{c}_{k,t},$$

with high probability. These, along with the union bound, imply the following:

$$\mathbb{P}\big(E_1 \cup E_2\big) \leq \frac{12}{t^{\alpha-1}}.$$

By using this result and Claim 1, we obtain the following inequality:

$$\mathbb{E}[T_k(n)] \leq u + v + \sum_{t=1}^{\infty}\frac{12}{t^{\alpha-1}},$$

where u and v are defined in (53) and (56), respectively. Choosing $\lambda = 1.28$ and substituting $\mathbb{E}[T_k(n)]$ into Proposition 4 proves the result.

D Proof of Theorem 2

For any k, if $X_{n,k}$ or $R_{n,k}$ has heavy tails, then the empirical rate estimator is weak in the sense that the convergence rate is polynomial rather than exponential (Bubeck et al., 2013). In the following, we propose a median-based rate estimator, and prove that it is robust in the sense that an exponential convergence rate is achieved even if the cost and reward are heavy-tailed. The correlation between $X_{1,k}$ and $R_{1,k}$ is exploited for improved coefficients.

Proposition 5 (Median-based rate estimation)**.** *For any given $\delta \in (0,1)$, let*

$$m = \lceil 3.5\log(\delta^{-1})\rceil + 1,$$

and $G_1, G_2, \ldots, G_m$ be a partition of $[s]$ where $|G_j| = \lfloor \frac{s}{m} \rfloor$ for each j. Define $\widehat{\mathbb{E}}_{G_j}[X_k]$ (and $\widehat{\mathbb{E}}_{G_j}[R_k]$) be the sample mean of $X_{n,k}$ (and $R_{n,k}$) in partition G_j, and $\tilde{r}_{j,k} = \frac{\widehat{\mathbb{E}}_{G_j}[R_k]}{\widehat{\mathbb{E}}_{G_j}[X_k]}$ for each j. Given $\lambda > 1$, if

$$s \geq 135\Big(\frac{\lambda}{\lambda-1}\Big)^2 Var(X_{1,k})\log(1.4\delta^{-1}), \tag{57}$$

then the following inequality holds:

$$\mathbb{P}\Big(\big|\overline{r}_{s,k} - r_k\big| > \frac{22\lambda}{\mathbb{E}[X_{1,k}]}\sqrt{\frac{\sigma_k^2\log(\delta^{-1})}{s}}\Big) \leq 1.4\delta,$$

where $\overline{r}_{s,k} = \underset{1\leq i\leq m}{median}\ \tilde{r}_{j,k}$ and σ_k is defined in (10) .

Proof. Given $\lambda > 1$, for any $j \in [m]$ and $p \in (0, \frac{1}{2})$, if

$$\sqrt{\frac{4mVar(X_{1,k})}{sp}} \leq \frac{\mathbb{E}[X_{1,k}](\lambda - 1)}{\lambda},$$

we have the following:

$$\mathbb{P}(|\tilde{r}_{j,k} - r_k| > \frac{\lambda}{\mathbb{E}[X_{1,k}]}\sqrt{\frac{8m\sigma_k^2}{sp}}) \leq p,$$

by Chebyshev's inequality and Proposition 2. Therefore, by Theorem 3.1 in (Minsker et al., 2015), we have:

$$\mathbb{P}\Big(|\bar{r}_{s,k} - r_k| > \frac{1-\beta}{\sqrt{1-2\beta}}\frac{\lambda}{\mathbb{E}[X_{1,k}]}\sqrt{\frac{8m\sigma_k^2}{sp}}\Big) \leq e^{-m\psi(\beta;p)},$$

for $\beta \in (p, \frac{1}{2})$ and

$$\psi(\beta;p) = \beta \log\Big(\frac{\beta}{p}\Big) + (1-\beta)\log\Big(\frac{1-\beta}{1-p}\Big).$$

For a given $\delta \in (0,1)$, the values $m = \lfloor 3.5\log(\delta^{-1})\rfloor + 1$, $\beta = 8/17$ and $p = 0.1$ yield the result. □

The proof of Theorem 2 is based on the regret decomposition in Appendix B and the following lemma.

Lemma 3. *For any $\lambda > 1$ and $\alpha > 2$, we have:*

$$\mathbb{E}[T_k(n)] \leq \log(n^\alpha)\Big(\frac{484\lambda^2\sigma_k^2}{\Delta_k^2(\mathbb{E}[X_{1,k}])^2} + \frac{135(\frac{\lambda}{\lambda-1})^2Var(X_{1,k})}{(\mathbb{E}[X_{1,k}])^2}\Big) + 48\frac{\alpha}{\alpha-2}, \tag{58}$$

for any k that satisfies $r_k < r^*$.

Lemma 3 is proved in an identical way to Lemma 2 by using the concentration inequality proposed in Proposition 5.

E Proof of Theorem 3

Proof. The regret under any admissible policy can be lower bounded as follows:

Lemma 4. *For any $B > 0$, let*

$$\phi_\pi(B) = \sum_k \mathbb{E}[\mathbb{I}\{I_{N_\pi(B)} = k\}]\mathbb{E}[X_{N_\pi(B),k}],$$

be the average cost in the last epoch under an admissible policy π, $\mu_+ = \max_k \mathbb{E}[X_{1,k}^+]$ and $\mu_* = \min_k \mathbb{E}[X_{1,k}]$. Then, the regret under π is lower bounded as follows:

$$Reg_\pi(B) \geq \sum_k \Delta_k \mathbb{E}[X_{1,k}]\mathbb{E}[T_k(\lceil\sqrt{2B/\mu_*}\rceil)] - \frac{\mu_+}{\mu_*}(1 + \frac{1}{\sqrt{2B}})\sum_k \Delta_k\mathbb{E}[X_{1,k}] - \phi_\pi(B). \tag{59}$$

Then, under the conditions stated in Theorem 3, the following result provides an asymptotic lower bound for $\mathbb{E}[T_k(n)]$ for any k with $r_k < r^*$.

Lemma 5. *If $\pi \in \Pi$ is a policy such that $\mathbb{E}[T_k^\pi(n)] = o(n^\alpha)$ for any $\alpha > 0$ and k such that $r_k(\theta_k) < r^*$, then we have the following lower bound:*

$$\liminf_{n\to\infty} \frac{\mathbb{E}[T_k(n)]}{\log(n)} \geq \frac{1}{D_k^\star}, \tag{60}$$

where $D_k^\star$ is the solution to the following optimization problem:

$$D_k^\star = \min_{\theta\in\Theta_k} D(P_{k,\theta_k}||P_{k,\theta}) \text{ subject to } r_k(\theta) \geq r^*.$$

Lemma 5 can be proved by a straightforward adaptation of Theorem 1 in (Burnetas and Katehakis, 1996).

If the moment condition $\mathbb{E}[(X_{1,k})^{2+\gamma}] < \infty$ holds for all k, then the term $\phi_\pi(B) = O(1)$ as $B \to \infty$ by Lorden's inequality (Asmussen, 2008). Therefore, using (59) and (60), we obtain the result. □

E.1 Proof of Lemma 4

Take any admissible policy π and $B > 0$. We have the following inequalities:

$$\begin{aligned} Reg_\pi(B) &= \mathbb{E}[\texttt{REW}_{\pi^{\texttt{opt}}(B)}(B)] - \mathbb{E}[\texttt{REW}_\pi(B)], \\ &\geq \mathbb{E}[\texttt{REW}_{\pi^*}(B)] - \mathbb{E}[\texttt{REW}_\pi(B)], \end{aligned}$$

since $\mathbb{E}[\texttt{REW}_{\pi^{\texttt{opt}}(B)}(B)] \geq \mathbb{E}[\texttt{REW}_{\pi^*}(B)]$ by definition. Then, by using a similar decomposition as (34), we have the following:

$$Reg_\pi(B) \geq \mathbb{E}[\sum_{t=1}^{\infty}\sum_k \Delta_k \mathbb{E}[X_{1,k}]\mathbb{I}\{W_{t-1} \leq B\}\mathbb{I}\{I_t = k\}] - r^*\phi_\pi(B), \tag{61}$$

$$\geq \mathbb{E}[\sum_{t=1}^{n_0}\sum_k \Delta_k \mathbb{E}[X_{1,k}]\mathbb{I}\{W_{t-1} \leq B\}\mathbb{I}\{I_t = k\}] - r^*\phi_\pi(B) \tag{62}$$

for any $n_0 > 0$, where $W_t^\pi = \max\limits_{1\leq i\leq t} S_i^\pi$. Since $\mathbb{I}\{W_{t-1}^\pi \leq B\} = 1 - \mathbb{I}\{W_{t-1}^\pi > B\}$, we have:

$$Reg_\pi(B) \geq \sum_k \mathbb{E}[T_k(n_0)]\Delta_k\mathbb{E}[X_{1,k}] - (\sum_k \Delta_k \mathbb{E}[X_{1,k}])\sum_{t=1}^{n_0}\mathbb{P}(W_{t-1}^\pi > B) - r^*\phi_\pi(B). \tag{63}$$

We have the following result:

$$\begin{aligned} \mathbb{P}(W_t^\pi > B) &\leq \mathbb{P}(\max_{1\leq i\leq t}(S_i^\pi)^+ > B), \\ &\leq \frac{\mathbb{E}[(S_t^\pi)^+]}{B}, \\ &\leq \frac{\mathbb{E}[\sum_{i=1}^t X_{i,I_i}^+]}{B} \leq \frac{t\mu_+}{B}, \end{aligned} \tag{64}$$

where the second inequality follows from Doob's martingale inequality (Durrett, 2019), and the last inequality is true since $\mu_+ \geq X_{i,I_i}^+$ with probability 1 for all i. Substituting (64) into (63), and setting $n_0 = \sqrt{2B/\mu_*}$ yields the result.

F Proof of Theorem 4

In the design of UCB-B2, empirical variance estimates are used, which require a modified analysis compared to UCB-B1.

Lemma 6. *If $\Delta_k > 0$ and $|X_{1,k}| \leq M_X$, $|R_{1,k}| \leq M_R$ a.s., then we have the following upper bound under* UCB-B2 *with $\alpha > s$:*

$$\begin{aligned} \mathbb{E}[T_k(n)] \leq 21\log(n^\alpha)\Big(\frac{M_X^4}{Var^2(X_{1,k})} + \frac{2M_X}{\mathbb{E}[X_{1,k}]} + \frac{3Var(X_{1,k})}{\mathbb{E}^2[X_{1,k}]}\Big) \\ + 42\log(n^\alpha)\Big(\frac{\sigma_k^2}{\Delta_k^2(\mathbb{E}[X_{1,k}])^2} + \frac{M_k}{\Delta_k\mathbb{E}[X_{1,k}]}\Big) + 48\frac{\alpha}{\alpha-2}, \end{aligned} \tag{65}$$

where $\sigma_k = Var(R_{1,k}) - \omega_k^2 Var(X_{1,k})$ and $M_k = M_R + r_k M_X$.

Proof. The proof follows along the same lines as Theorem 1 and the proof of Theorem 3 in (Audibert et al., 2009). For any k, let the variance estimate $\widehat{V}_{k,n}(X_k)$ be defined as follows:

$$\widehat{V}_{k,n}(X_k) = \frac{1}{T_k(n)}\sum_{i=1}^n \mathbb{I}\{I_i = k\}\big(X_{i,k} - \widehat{\mathbb{E}}_n[X_{1,k}]\big)^2,$$

where $\widehat{\mathbb{E}}_n[X_k]$ is the empirical mean of the observations up to epoch n. Also, let $\nu_{k,n}$ be defined for $X_k \in [0, M_X]$ as follows:

$$\nu_{k,n}(X_k) = M_X^2\Big(\frac{7\log(n^\alpha)}{6T_k(n)} + \sqrt{\frac{\log(n^\alpha)}{2T_k(n)}}\Big),\ \alpha > 2.$$

Then, it can be shown by using Bernstein's inequality that $\widehat{V}_{k,n}(X_k)+\nu_{k,n}(X_k)$ is an upper bound for $Var(X_{1,k})$ with high probability. Using this result, we obtain the sample size required for the stability of the rate estimator by using identical steps as Theorem 1. $\square$

G Proof of Theorem 5

The proof of Theorem 4 follows the same steps as Theorem 5, with the difference that the correlation between $X_{n,k}$ and $R_{n,k}$ are estimated in the latter. In order to observe the effect of using LMMSE estimates to exploit correlation, we first present concentration bounds for ω_k and $\min_\omega\ Var(R_{1,k} - \omega X_{1,k})$.

G.1 Preliminaries

Throughout this subsection, we consider a generic iid stochastic process (X_n, R_n) with $X_n \in [0, M_X]$ and $R_n \in [0, M_R]$. For this process, let $\omega_* = \arg\min_\omega\ L(\omega)$ where

$$L(\omega) = Var(R_1 - \omega X_1),$$

and $\widehat{\omega}_s = \arg\min_\omega\ \widehat{L}_s(\omega)$ where

$$\widehat{L}_s(\omega) = \frac{1}{s}\sum_{i=1}^{s}\Big(R_i - \widehat{\mathbb{E}}_s[R] - \omega(X_i - \widehat{\mathbb{E}}_s[X])\Big)^2.$$

Note that $\omega_* = \frac{Cov(X_1,R_1)}{Var(X_1)}$ and $\widehat{\omega}_s = \frac{\widehat{Cov}_s(X,R)}{\widehat{Var}_s(X)}$ where

$$\widehat{Cov}_s(X,R) = \frac{1}{s}\sum_{i=1}^{s}(R_i - \widehat{\mathbb{E}}_s[R])(X_i - \widehat{\mathbb{E}}_s[X]),$$

is the empirical covariance and $\widehat{Var}_s(X) = \widehat{Cov}_s(X,X)$. In the following, we propose concentration inequalities for ω_* and $L(\omega_*)$.

Proposition 6 (Concentration of LMMSE Estimator)**.** *Let $M_Z \geq M_R + \omega_* M_X$ and $\lambda = 1 + \frac{1}{2\sqrt{2}}$. Then, for any $\delta \in (0,1)$, if*

$$s \geq \frac{63M_X^4\log(\delta^{-1})}{Var^2(X_1)}, \tag{66}$$

then the following inequalities hold simultaneously:

$$\mathbb{P}(|\omega_* - \widehat{\omega}_s| > \frac{\lambda M_Z M_X}{Var(X_1)}\sqrt{\frac{\log(\delta^{-1})}{s}}) \leq 12\delta,$$
$$\mathbb{P}(|L(\omega_*) - \widehat{L}_s(\widehat{\omega}_s)| > M_Z^2\sqrt{\frac{2\log(\delta^{-1})}{s}}) \leq 18\delta.$$

Proof. For the first inequality, recall that $\omega_* = \frac{Cov(X_1,R_1)}{Var(X_1)}$ and $\widehat{\omega}_s$ is the ratio of empirical estimates for $Cov(X_1,R_1)$ and $Var(X_1)$. Therefore, we can use Proposition 2 for the proof. Note that (66) is the stability condition for the estimator $\widehat{\omega}_s$. Since $s \geq \frac{1}{2}\log(\delta^{-1})$, Hoeffding's inequality yields the following result for the empirical covariance:

$$\mathbb{P}(|\widehat{Cov}_s(X_1,R_1) - Cov(X_1,R_1)| > M_X M_R\sqrt{\frac{\log(\delta^{-1})}{s}}) \leq 6\delta. \tag{67}$$

Using this twice for $\widehat{Cov}_s(X_1,R_1)$ and $\widehat{Var}_s(X_1)$, we obtain the first inequality.

For the second inequality, first we make the following decomposition:

$$|\widehat{L}_s(\widehat{\omega}_s) - L(\omega_*)| = |\widehat{L}_s(\omega_*) - L(\omega_*)| + |\widehat{L}_s(\widehat{\omega}_s) - \widehat{L}_s(\omega_*)|. \tag{68}$$

For the first term on the RHS of (68), we have the following result:

$$|\widehat{L}_s(\omega_*) - L(\omega_*)| \leq M_Z^2 \sqrt{\frac{\log(\delta^{-1})}{s}},$$

by applying Hoeffding's inequality for the variance (67) to the decomposition:

$$Var(R_1 - \omega X_1) = Var(R_1) + \omega^2 Var(X_1) - 2Cov(X_1, R_1),$$

and its empirical counterpart. For the second term on the RHS of (68), note that the following identity holds by the orthogonality principle:

$$\widehat{L}_s(\omega) = \widehat{L}_s(\widehat{\omega}_s) + |\omega - \widehat{\omega}_s|^2 \widehat{Var}_s(X_1), \tag{69}$$

for any $\omega \in \mathbb{R}$. Therefore, by union bound, we have the following result:

$$\mathbb{P}\Big(|L_s(\omega_*) - \widehat{L}_s(\widehat{\omega}_s)| > M_Z^2\Big(\sqrt{\frac{\log(\delta^{-1})}{s}} + \frac{3\lambda^2 M_X^2 \log(\delta^{-1})}{2Var(X_1)s}\Big)\Big) \leq 18\delta,$$

from the concentration result for $|\omega_* - \widehat{\omega}_s|$ and (67) with $M_X^2\sqrt{\frac{\log(\delta^{-1})}{s}} \leq \frac{Var(X_1)}{2}$ by (66). Since s is assumed to be sufficiently large by (66), we have:

$$\sqrt{\frac{\log(\delta^{-1})}{s}} > \frac{3\lambda^2 M_X^2 \log(\delta^{-1})}{2Var(X_1)s},$$

which concludes the proof. $\square$

G.2 Proof of Theorem 5

The proof follows a similar steps as the proof of Theorem 4 (see Appendix F). The main difference is the use of LMMSE estimator as a surrogate for $V(X_{1,k}, R_{1,k})$. By using Proposition 6, one can show the following:

$$\mathbb{E}[T_k(n)] \leq 21\log(n^\alpha)\Big(\frac{3M_X^4}{Var^2(X_{1,k})} + \frac{2M_X}{\mathbb{E}[X_{1,k}]} + \frac{3Var(X_{1,k})}{\mathbb{E}^2[X_{1,k}]}\Big) + 42\log(n^\alpha)\Big(\frac{\sigma_k^2}{\Delta_k^2(\mathbb{E}[X_{1,k}])^2} + \frac{M_k + M}{\Delta_k \mathbb{E}[X_{1,k}]}\Big) + 64\frac{\alpha}{\alpha - 2},$$

where $M = \frac{M_X M_Z}{\sqrt{Var(X_{1,k})}}$, $\sigma_k = Var(R_{1,k}) - \omega_k^2 Var(X_{1,k})$ and $M_k = M_R + r_k M_X$.