Multi-armed Bandits with Cost Subsidy

Deeksha Sinha MIT/Facebook
Karthik Abinav Sankararaman Facebook
Abbas Kazerouni Facebook
Vashist Avadhanula Facebook

Abstract

In this paper, we consider a novel variant of the multi-armed bandit (MAB) problem, *MAB with cost subsidy*, which models many real-life applications where the learning agent has to pay to select an arm and is concerned about optimizing cumulative costs and rewards. We present two applications, *intelligent SMS routing problem* and *ad audience optimization problem* faced by several businesses (especially online platforms) and show how our problem uniquely captures key features of these applications. We show that naive generalizations of existing MAB algorithms like Upper Confidence Bound and Thompson Sampling do not perform well for this problem. We then establish fundamental lower bound of $\Omega(K^{1/3}T^{2/3})$ on the performance of any online learning algorithm for this problem, highlighting the hardness of our problem in comparison to the classical MAB problem (where T is the time horizon and K is the number of arms). We also present a simple variant of *explore-then-commit* and establish near-optimal regret bounds for this algorithm. Lastly, we perform extensive numerical simulations to understand the behavior of a suite of algorithms for various instances and recommend a practical guide to employ different algorithms.

1 Introduction

In the traditional (stochastic) MAB problem (Robbins (1952)), the learning agent has access to a set of K actions (arms) with unknown but fixed reward distributions and has to repeatedly select an arm to maximize the cumulative reward. Here, the challenge is designing a policy that balances the tension between acquiring information about arms with little historical observations and exploiting the most rewarding arm based on existing information. The aforementioned exploration-exploitation trade-off has been extensively studied leading to a number of simple but extremely effective algorithms like Upper Confidence Bound (Auer et al. (2002a)) and Thompson Sampling (Thompson (1933); Agrawal and Goyal (2017a)), which have been further generalized and applied in a wide range of application domains including online advertising (Langford and Zhang (2008); Cesa-Bianchi et al. (2014), Oliver and Li (2011)), recommendation systems (Li et al. (2015, 2011); Agrawal et al. (2016)), social networks and crowd sourcing (Anandkumar et al. (2011); Sankararaman et al. (2019), Slivkins and Vaughan (2014)); see Bubeck and Cesa-Bianchi (2012) and Slivkins (2019) for a detailed review. However, most of these approaches cannot be generalized to settings involving multiple metrics (for example reward and cost) when the underlying trade-offs between these metrics are not known a priori.

In many real-world applications of MAB, some of which we will elaborate below, it is common for the agent to incur costs to play an arm, with *high performing arms* costing more. Though, one can model this in the traditional MAB framework by considering cost subtracted from the reward as the modified objective, such a modification is not always meaningful, particularly in settings where the reward and cost associated with an arm represent different quantities (for example click rate and cost of an ad). In such problems, it is natural for the learning agent to *optimize* for both the metrics, typically trying to avoid incurring exorbitant costs for a marginal increase in cumulative reward. Motivated by the aforementioned scenario, in this paper, we consider a variant of the MAB problem, where the agent is not only concerned about balancing the exploration-exploitation trade-offs to maximize the cumulative reward but also balance the trade-offs associated with multiple objectives that are intrinsic to several practical applications. More specifically, in this work, we study a stylized problem, where to man-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

age costs, the agent is willing to tolerate a small loss from the *highest reward*, measured as the reward that could be obtained by the traditional MAB problem in absence of costs. We refer to this problem as MAB problem with a cost subsidy (see Sec 1.1 for exact problem formulation), where the subsidy refers to the amount of reward the learning agent is willing to forgo to improve costs. Before we explain our problem and technical contributions in detail, we will elaborate on the applications that motivate this problem.

Intelligent SMS Routing. Many businesses such as banks, delivery services, airlines, hotels, and various online platforms send SMSes (text messages) to their users for a variety of reasons including two-factor authentication, order confirmations, appointment reminders, transaction alerts, and as a direct marketing line (see Twilio and Uber (2020)). These text messages referred to as Application-to-Person (A2P) messages constitute a significant portion of all text messages sent through cellular networks today. In fact, A2P messages are forecasted to be a $86.3 billion business by 2025 (MarketWatch (2020)).

To deliver these messages, businesses typically enlist the support of telecom aggregators, who have private agreements with mobile operators. Each aggregator offers a unique combination of quality, as measured by the fraction of text messages successfully delivered by them and price per message. Surprisingly, it is common for delivery rates of text messages to not be very high (see Canlas et al. (2010); Meng et al. (2007); Zerfos et al. (2006); Osunade and Nurudeen for QoS analysis in different geographies) and for aggregator's quality to fluctuate with time due to various reasons ranging from network outage to traffic congestion. Therefore, the platform's problem of balancing the tension between inferring aggregator's quality through exploration and exploiting the current *best performing aggregator* to maximize the number of messages delivered to users leads to a standard MAB formulation. However, given the large volume of messages that need to be dispatched, an MAB based solution that focuses exclusively on quality of the aggregator could result in exorbitant spending for the business. A survey of businesses shows that the number of text messages they are willing to send will have a significant drop if the cost per SMS is increased by a few cents per SMS (Ovum (2017)). Moreover, in many situations platforms have back up communication channels such as email based authentication or notifications via in-app/website features, which though not as effective as a text message in terms of read rate, can be used if guaranteeing the text message delivery proves to be very costly. Therefore, it is natural for businesses to prefer an aggregator with lower costs as long as their quality is comparable to the aggregator with the best quality.

Ad-audience Optimization. We now describe another real-world application in the context of online advertisements. Many advertisers (especially small-to-medium scale businesses) have increasingly embraced the notion of *auto-targeting* where they let the advertising platform identify a *high-quality* audience group (*e.g.,* Koningstein (2006); Amazon (2019); Facebook (2016); Google (2014)). To enable this, the platform explores the audience-space to identify *cheaper* opportunities that also give high click-through-rate (ctr) and conversion rate. Here, it is possible for different audience groups to have different yields i.e. quality (CTR/conversion rate) for a specific ad. However, it may require vastly different bids to reach different audiences due to auction overlap with other ad campaigns with smaller audience targeting. Thus, the algorithm is faced with a similar trade-off; as long as a particular audience-group gives a high-yield, the goal is to find the cheapest one.

We now present a novel formulation of a multi-armed bandit problem that captures key features of these applications, where our goal is to develop a cost sensitive MAB algorithm that balances both the exploration-exploitation trade-offs as well as the tension between conflicting metrics in a multi-objective setting.

1.1 Problem formulation

To formally state our problem, given an instance I, in every round $t \in [T]$ the agent chooses an arm $i \in [K]$ and realizes a reward r_t, sampled independently from a fixed, but unknown distribution $\mathcal{F}_i$ with mean μ_i (or μ_i^I) and incurs a cost c_i (or c_i^I), which is known a priori. Here, in order to manage costs, we allow the agent to be agnostic between arms, whose expected reward is greater than $1-\alpha$ fraction of the highest expected reward, for a fixed and known value of α, which we refer to as the *subsidy factor*. The agent's objective is to learn and pull the cheapest arm among these *high* quality arms as frequently as possible.

More specifically, let m_* denote the arm with highest expected mean, *i.e.,* $m_* = \text{argmax}_{i\in[K]}\ \mu_i$, and $\mathcal{C}_*$ be the set of arms whose expected reward is within $1-\alpha$ factor of the highest expected reward, *i.e.,* $\mathcal{C}_* = \{i \in [K] \mid \mu_i \geq (1-\alpha)\mu_{m_*}\}$. We refer to the quantity $(1-\alpha)\mu_{m_*}$ as the *smallest tolerated reward*. Without loss of generality, we assume the reward distribution has support $[0,1]$. The goal of the agent is to design a policy (algorithm) π that will learn the cheapest arm whose expected reward is at least as large as the smallest tolerated reward. In other words, the agent needs to learn the identity and simultaneously maximize the number of plays of arm $i_* = \text{argmin}_{i\in\mathcal{C}_*}\ c_i$.

Since in the SMS application, the reward is the quality of the chosen aggregator, we will use the terms reward and quality interchangeably.

To measure the performance of any policy π, we propose two notions of regret - quality and cost regret, with the agent's goal being minimizing both of them:

$$\begin{aligned}&\mathsf{Quality\_Reg}_\pi(T,\alpha,\boldsymbol{\mu},\boldsymbol{c})\\&\quad=\mathbb{E}\left[\sum_{t=1}^{T}\max\{(1-\alpha)\mu_{m_*}-\mu_{\pi_t},0\}\right],\\&\mathsf{Cost\_Reg}_\pi(T,\alpha,\boldsymbol{\mu},\boldsymbol{c})\\&\quad=\mathbb{E}\left[\sum_{t=1}^{T}\max\{c_{\pi_t}-c_{i_*},0\}\right],\end{aligned}\tag{1}$$

where $\boldsymbol{c} = (c_1, \cdots c_K), \boldsymbol{\mu} = (\mu_1, \cdots \mu_K)$ and the expectation is over the randomness in policy π. Equivalently, the cost and quality regret of policy π on an instance I of the problem is denoted as $\mathsf{Quality\_Reg}_\pi(T, \alpha, I)$ and $\mathsf{Cost\_Reg}_\pi(T, \alpha, I)$ where the instance is defined by the distributions of the reward and cost of each arm. The objective then is to design a policy that simultaneously minimizes both the cost and quality regret for all possible choices of $\boldsymbol{\mu}$ and $\boldsymbol{c}$ (equivalently all instances I).

Choice of objective function. Note that a parametrized linear combination of reward and cost metrics, i.e. $\mu - \lambda c$ for an appropriately chosen λ is a popular approach to balance cost-reward trade-off. From an application stand-point there are two important considerations which favor using two the above proposed objective instead of a linear combination of reward and cost metrics. First, in a real-world system we need an explicit control over the parameter α that is not instance-dependent to understand and defend the trade-off between the various objectives. From the perspective of a product manager, the *indifference* among the arms within $(1 - \alpha)$ of the arm with the highest mean reward, is a transparent and interpretable compromise to lower costs. Second, for the intelligent SMS routing application discussed earlier, different sets of aggregators operate in different regions. Thus, separate λ values would need to be configured for each region, making the process cumbersome.

Further, the setting considered in this paper is mathematically not equivalent to taking a parametrized linear combination of reward and cost metrics. In particular, for any specified subsidy factor α, the value λ required in the linear objective function, for i_* to be the optimal arm would depend on the cost and reward distributions of the arms. Therefore, using a single value of λ and relying on standard MAB algorithms would not lead to the desired outcome for our problem.

1.2 Related Work

Our problem is closely related to the MAB with multiple objectives line of work, which has attracted considerable attention in recent times. The existing literature on multi-objective MAB can be broadly classified into three different categories.

Bandits with Knapsacks (BwK). Bandits with knapsacks (BwK), introduced in the seminal work of Badanidiyuru et al. (2018) is a general framework that considers the standard MAB problem under the presence of additional budget/resource constraints. The BwK problem encapsulates a large number of constrained bandit problems that naturally arise in many application domains including dynamic pricing, auction bidding, routing and scheduling ((see Tran-Thanh et al. (2012); Agrawal and Devanur (2014); Immorlica et al. (2019)). In this formulation, the agent has access to a set of d finite resources and K arms, each associated with a reward distribution. Upon playing arm a at time t, the agent realizes a reward of r_t and incurs a penalty of $c_t^{(i)}$ for resource i, all drawn from a fixed, but unknown distribution corresponding to the arm. The objective of the agent is to maximize the cumulative reward before one of the resources is completely depleted. Although appealing in many applications, BwK formulation requires *hard* constraint on resources (cost in our setting) and hence, cannot be easily generalized to our problem. In particular, in the cost subsidized MAB problem, the equivalent budget limits depend on the problem instance and therefore cannot be determined a priori.

Pareto Optimality and Composite Objective. The second formulation is focused on identifying Pareto optimal alternatives and uniformly choosing among these options (see Drugan and Nowe (2013); Yahyaa et al. (2014); Paria et al. (2018); Yahyaa and Manderick (2015)). These approaches do not apply to our problem, since some of the Pareto alternatives could have extreme values for one of the metrics, for example having very low cost and low quality or extremely high cost and quality, making them undesirable for the applications discussed earlier. Closely related to this line of work is the set of works that focus on a composite objective by appropriately weighting the different metrics (see Paria et al. (2018); Yahyaa and Manderick (2015)). Such formulations also do not immediately apply for our problem, since in the SMS and ad applications discussed earlier, it is not acceptable to drop the quality beyond the allowed level irrespective of the cost savings we could obtain. Furthermore, in the SMS application, the trade-offs between quality and costs could vary from region to region, making it hard to identify a good set of weights for

the composite objective (see Section 1.1).

Conservative Bandits and Bandits with Safety Constraints. Two other lines of work that are recently receiving increased attention, particularly from practitioners are *bandits with safety constraints* (*e.g.,* Daulton et al. (2019); Amani et al. (2020); Galichet et al. (2013)) and *conservative bandits* (*e.g.,* Wu et al. (2016); Kazerouni et al. (2017)). In both these formulation, the algorithm chooses one of the arms and receives a reward and a cost associated with it. The goal of the algorithms is to maximize the total reward obtained while ensuring that either the chosen arm is within a pre-specified threshold (when costs of arms are unknown a priori) or reward of the arm is at least a specified fraction of a known benchmark arm. Neither of these models exactly capture the requirements of our applications: a) we do not have a hard constraint on the acceptable cost of a pulled arm. In particular, choosing low quality aggregators to avoid high costs (even for a few rounds) could be disastrous since it leads to bad user experience on the platform and eventual churn, and b) the equivalent benchmark arm in our case i.e. the arm with the highest mean reward is not known a priori.

Best Arm Identification. Apart from the closely related works mentioned above, our problem of identifying the *cheapest* arm whose expected reward is within an acceptable margin from the highest reward can be formulated as a stylized version of the *best-arm identification problem* (Katz-Samuels and Scott (2019); Jamieson and Nowak (2014); Chen et al. (2014); Cao et al. (2015); Chen et al. (2016)). However, in many settings and particularly applications discussed earlier, the agent's objective is optimizing cumulative reward and not just identifying the *best arm*.

1.3 Our Contributions

Novel Problem Formulation. In this work, we propose a stylized model, *MAB with a cost subsidy* and introduce new performance metrics that uniquely capture the salient features of many real world online learning problems involving multiple objectives. For this problem, we first show that naive generalization of popular algorithms like Upper Confidence Bound (UCB) and Thompson Sampling (TS) could lead to poor performance on the metrics. In particular, we show that the naive generalization of TS for this problem would lead to a linear cost regret for some problem instances.

Lower Bound. We establish a fundamental limit on the performance of *any online algorithm* for our problem. More specifically, we show that any online learning algorithm will incur a regret of $\Omega(K^{1/3}T^{2/3})$ on either the cost or the quality metric (refer to (1)), further establishing the *hardness of our problem* relative to the standard MAB problem, for which it is possible to design algorithms that achieve worst case regret bound of $\tilde{O}(\sqrt{KT})$. We introduce a novel reduction technique to derive the above lower bound, which is of independent interest.

Cost Subsidized Explore-Then-Commit. We present a simple algorithm, based on the *explore-then-commit* (ETC) principle and show that it achieves near-optimal performance guarantees. In particular, we establish that our algorithm achieves a worst-case bound of $O(K^{1/3}T^{2/3}\sqrt{\log T})$ for both cost and quality regret. A key challenge in generalizing the ETC algorithm for this problem arises from having to balance between two asymmetric objectives. We also discuss generalizations of the algorithm for settings where cost of the arms is not known a priori. Furthermore, we consider a special scenario of bounded costs, where naive generalizations of TS and UCB work reasonably well and establish worst case regret bounds.

Numerical Simulation. Lastly, we perform extensive simulations to understand various regimes of the problem parameters and compare different algorithms. More specifically, we consider scenarios where naive generalizations of UCB and TS, which have been adapted in real life implementations (see Daulton et al. (2019)) perform well and settings where they perform poorly, which should be of interest to practitioners.

1.4 Outline

The rest of this paper is structured as follows. In Section 2, we show that the naive generalization to TS or UCB algorithms perform poorly and in Section 3, we establish lower bounds on performance of any algorithm for MAB with cost subsidy problem. In Section 4, we present a variation of the ETC algorithm, and show that it achieves a near-optimal regret bound of $\tilde{O}(K^{1/3}T^{2/3})$ for both the metrics. In section 5, we show that with additional assumptions it is possible to show improved performance bounds for naive generalization of existing algorithms. Finally, in section 6 we perform numerical simulations to explore various regimes of the instance-space.

2 Performance of Existing MAB Algorithms

In this section, we consider a natural extension of two popular MAB algorithms, TS and UCB for our problem and show that such adaptations perform poorly. This highlights the challenges involved in developing good algorithms for the MAB problem with cost sub-

sidy. In particular, we establish theoretically that for some problem instances the TS variant incurs a linear cost regret and observe similar performance for the UCB variant empirically. Our key focus on TS in this section is primarily motivated by the superior performance that have been observed over a stream of recent papers in the context of TS versus more traditional approaches such as UCB (see Scott (2010); Oliver and Li (2011); May et al. (2012); Agrawal et al. (2017)).

We present the details of TS and UCB adaptations in Algorithm 1, which we will refer to as Cost-Subsidized TS(CS-TS) and Cost-Subsidized UCB(CS-UCB) respectively. These extensions are inspired by Daulton et al. (2019), which demonstrates empirical efficacy on a related (but different) problem. Briefly, in the CS-TS(CS-UCB) variation, we follow the standard TS (UCB) algorithm and obtain a quality score which is a sample from the posterior distribution (upper confidence bound) for each arm. We then construct a feasible set of arms consisting of arms whose quality scores are greater than $1-\alpha$ fraction of the highest quality score. Finally, we pull the cheapest arm among the feasible set of arms.

Algorithm 1: Cost Subsidized TS and UCB Algorithms

Result: Arm I_t to be pulled in each round $t \in [T]$
Input : T, K, prior distribution for mean rewards of all arms $\{\nu_i\}_{i=1}^K$, reward likelihood function $\{L_i\}_{i=1}^K$
$T_i(1) = 0\ \forall i \in [K]$;
for $t \in [K]$ **do**
  $I_t = t$;
  Play arm I_t and observe reward r_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\}\ \forall i \in [K]$;
end
for $t \in [K+1, T]$ **do**
  for $i \in [K]$ **do**
    $\hat{\mu}_i(t) \leftarrow \left(\sum_{\tau=1}^{t-1} r_\tau \mathbf{I}\{I_\tau = i\}\right)/T_i(t)$;
    $\beta_i(t) \leftarrow \sqrt{(2\log T)/T_i(t)}$;
    UCB: $\mu_i^{score}(t) \leftarrow \min\{\hat{\mu}_i(t) + \beta_i(t), 1\}$;
    TS: Sample $\mu_i^{score}(t)$ from the posterior distribution of arm i,
    $\nu_i\big(\cdot\,|\{r_s\}_{s\in\{1,2,\cdots t-1\}\text{ s.t. } I_s=i}, L_i\big)$;
  end
  $m_t = \arg\max_i \mu_i^{score}(t)$;
  $Feas(t) = \{i : \mu_i^{score}(t) - (1-\alpha)\mu_{m_t}^{score} \geq 0\}$;
  $I_t = \arg\min_{i\in Feas(t)} c_i$;
  Play arm I_t and observe reward r_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\}\ \forall i \in [K]$;
end

We will now show that CS-TS with Gaussian priors and posteriors (i.e. Gaussian distribution with mean $\hat{\mu}_i(t)$ and variance $1/T_i(t)$) described in Algorithm 1 incurs a linear cost regret in the worst case. More precisely, we prove the following result.

Theorem 1. *For any given K, α, T there exists an instance ϕ of problem such that* $\mathsf{Quality\_Reg}_{\mathsf{CS-TS}} + \mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T, \alpha, \phi)$ *is* $\Omega(T)$.

Proof Sketch. The proof closely follows the lower bound argument in Agrawal and Goyal (2017b). We briefly describe the intuition behind the result. Consider a scenario where the highest reward arm is expensive arm while all other arms are cheap and have rewards marginally above the smallest tolerated reward. In the traditional MAB problem, the anti-concentration property of the Gaussian distribution (see Agrawal and Goyal (2017b)) ensures samples from *good arm* would be large enough with sufficient frequency, ensuring appropriate exploration and good performance. However, in our problem, the anti-concentration property would result in playing the expensive arm too often since the difference in the mean qualities is small, incurring a linear cost regret while achieving zero quality regret. A complete proof of the theorem is provided in Appendix C. □

The poor performance of the algorithm is not limited only to the above instance and usage of Gaussian prior. More generally, the CS-TS and CS-UCB algorithms seem to perform poorly whenever the mean reward of the optimal arm is very close to the smallest tolerated reward. We illustrate this through another empirical example. Consider the following instance with two arms each having Bernoulli rewards and $T = 10,000$. The costs of the two arms are $c_1 = 0$ and $c_2 = 1$. The expected qualities are $\mu_1 = 0.5(1-\alpha)+1/\sqrt{T}, \mu_2 = 0.5$ with $\alpha = 0.1$. The prior of the mean reward of both the arms is a Beta(1,1) distribution. Here, the quality regret will be zero irrespective of which arm is played. But both CS-TS and CS-UCB incur significant cost regret as shown in Figure 1. (In the figure, we also plot the performance of the key algorithm we propose in the paper (Algorithm 2) and note that it has much superior performance as compared to CS-TS and CS-UCB.)

3 Lower Bound

In this section, we establish that any policy must incur a regret of $\Omega(K^{1/3}T^{2/3})$ on at least one of the regret metrics. More precisely, we prove the following result.

Theorem 2. *For any given α, K, T and (possibly randomized) policy π, there exists an instance ϕ of problem* (1) *with $K+1$ arms such*

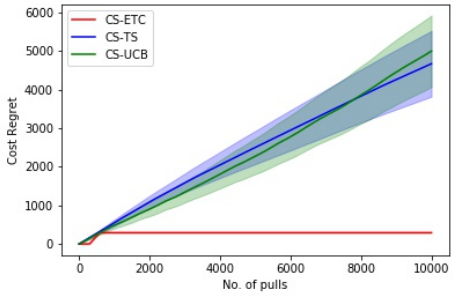

Figure 1: Cost regret of various algorithms for an instance where the mean reward of the optimal arm is very close to the smallest tolerated reward. CS-TS and CS-UCB incur significant regret. But CS-ETC attains low cost regret. The width of the error bands is two standard deviations based on 50 runs of the simulation.

that $\mathsf{Quality\_Reg}_\pi(T,\alpha,\phi) + \mathsf{Cost\_Reg}_\pi(T,\alpha,\phi)$ *is* $\Omega\left((1-\alpha)^2 K^{\frac{1}{3}} T^{\frac{2}{3}}\right)$ *when* $0 \le \alpha \le 1$ *and* $1 \le K \le T$.

3.1 Proof Overview

We consider the following families of instances to establish the lower bound. More specifically, we first prove the result for $\theta = 0$ and then establish a reduction for $\theta = \alpha$ to the special case of $\theta = 0$.

Definition 1 (Family of instances $\Phi_{\theta,p,\epsilon}$)**.** *Define a family of instances* $\Phi_{\theta,p,\epsilon}$ *consisting of instances* $\Phi^0_{\theta,p,\epsilon}, \Phi^1_{\theta,p,\epsilon}, \cdots \Phi^K_{\theta,p,\epsilon}$ *each with* $K+1$ *Bernoulli arms indexed by* $0, 1, \cdots, K$*. For the instance* $\Phi^0_{\theta,p,\epsilon}$*, the costs and mean reward of the* j*-th arm are*

$$c_j^{\Phi^0_{\theta,p,\epsilon}} = \begin{cases} 0 & j = 0 \\ 1 & j \neq 0 \end{cases}, \qquad \mu_j^{\Phi^0_{\theta,p,\epsilon}} = \begin{cases} p & j = 0 \\ \frac{p}{1-\theta} & j \neq 0 \end{cases}.$$

for $0 \le j \le K$*. For the instance* $\Phi^a_{\theta,p,\epsilon}$ *with* $1 \le a \le K$*, the costs and mean rewards of the* j*-th arm are*

$$c_j^{\Phi^a_{\theta,p,\epsilon}} = \begin{cases} 0 & j = 0 \\ 1 & j \neq 0 \end{cases}, \qquad \mu_j^{\Phi^a_{\theta,p,\epsilon}} = \begin{cases} p & j = 0 \\ \frac{p+\epsilon}{1-\theta} & j = a \\ \frac{p}{1-\theta} & otherwise \end{cases}.$$

for $0 \le j \le K$*, where* $0 \le \theta < 1, 0 < p \le 1/2, \epsilon > 0$ *and* $(p+\epsilon)/(1-\theta) < 1$.

Lemma 1. *For any given* p, K, T *and any (possibly randomized) policy* π*, there exists an instance* ϕ *(from the family* $\Phi_{0,p,\epsilon}$*) such that* $\mathsf{Quality\_Reg}_\pi(T,0,\phi) + \mathsf{Cost\_Reg}_\pi(T,0,\phi)$ *is* $\Omega\left(pK^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ *when* $0 < p \le 1/2$ *and* $1 \le K \le T$.

Lemma 1 establishes that when $\alpha = 0$, any policy must incur a regret of $\Omega(K^{1/3}T^{2/3})$ on an instance from the family $\Phi_{0,p,\epsilon}$. To prove Lemma 1, we argue that any online learning algorithm will not be able to differentiate the instance $\Phi^0_{0,p,\epsilon}$ from the instance $\Phi^a_{0,p,\epsilon}$ for $1 \le a \le K$ and therefore, must either incur a high cost regret if the algorithm does not select 0^{th} arm frequently or high quality regret if the algorithm selects 0^{th} arm frequently. More specifically, any online algorithm would require $O(1/\epsilon^2)$ samples or rounds to distinguish instance $\Phi^0_{0,p,\epsilon}$ from instance $\Phi^a_{0,p,\epsilon}$ for $1 \le a \le K$. Hence, any policy π can avoid high quality regret by *exploring sufficiently* for $O(1/\epsilon^2)$ rounds, incurring a cost regret of $O(1/\epsilon^2)$ or incur zero cost regret at the expense of $O(T\epsilon)$ regret on the reward metric. This suggests a trade-off between $1/\epsilon^2$ and $T\epsilon$, which are of the same magnitude at $\epsilon = T^{-1/3}$ resulting in the aforementioned lower bound. The complete proof generalizes techniques from the standard MAB lower bound proof and is provided in Appendix B. □

Now, we generalize the above result for $\alpha = 0$ to any α for $0 \le \alpha \le 1$. The main idea in our reduction is to show that if there exists an algorithm π_α for $\alpha > 0$ such that $\mathsf{Quality\_Reg}_\pi(T,\alpha,\phi) + \mathsf{Cost\_Reg}_\pi(T,\alpha,\phi)$ is $o(K^{1/3}T^{2/3})$ on every instance in the family $\Phi_{\alpha,p,\epsilon}$, then we can use π_α as a sub-routine to construct an algorithm π for problem (1) such that $\mathsf{Quality\_Reg}_\pi(T,0,\phi) + \mathsf{Cost\_Reg}_\pi(T,0,\phi)$ is $o(K^{1/3}T^{2/3})$ on every instance in $\Phi_{0,p,\epsilon}$, thus contradicting the lower bound of Lemma 1. This will prove Theorem 2 by contradiction. In order to construct the aforementioned sub-routine, we leverage techniques from *Bernoulli factory* (Keane and O'Brien (1994); Huber (2013)) to generate a sample from a Bernoulli random variable with parameter $\mu/(1-\alpha)$ using samples from a Bernoulli random variable with parameter μ, for any $0 < \mu < 1 - \alpha < 1$. We provide the exact sub-routine and complete proof in Appendix B.

4 Explore-then-commit based algorithm

We propose an explore-then-commit algorithm, named Cost-Subsidized Explore-Then-Commit (CS-ETC), to have better worst case performance guarantees as compared to the extensions of the TS and UCB algorithms. As the name suggests, first this algorithm plays each arm for a specified number of rounds. After sufficient exploration, the algorithm continues in a UCB-like fashion. In every round, based on the upper and lower confidence bounds on the reward of each arm, a feasible set of arms is constructed as an estimate of all arms having mean reward greater than the smallest tolerated reward. The lowest cost arm in this feasible set is then pulled. This is detailed in Algorithm 2. The key question that arises in this algorithm is how many exploration rounds are needed before exploitation can begin. We establish that $O\left((T/K)^{2/3}\right)$ rounds are sufficient for exploration in the following result (proof in

Appendix C).

Algorithm 2: Cost-Subsidized Explore-Then-Commit

Result: Arm I_t to be pulled in each round $t \in [T]$
Input : K, T, no. of exploration pulls per arm τ
$T_i(1) = 0 \ \forall i \in [K]$
Pure exploration phase:
for $t \in [1, K\tau]$ **do**
  $I_t = t \mod K$;
  Pull arm I_t to obtain reward r_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\} \ \forall i \in [K]$;
end
UCB phase:
for $t \in [K\tau + 1, T]$ **do**
  $\hat{\mu}_i(t) \leftarrow \left(\sum_{\tau=1}^{t-1} r_\tau \mathbf{I}\{I_\tau = i\}\right) / T_i(t) \ \forall i \in [K]$;
  $\beta_i(t) \leftarrow \sqrt{(2 \log T) / T_i(t)} \ \forall i \in [K]$;
  $\mu_i^{\mathsf{UCB}}(t) \leftarrow \min\{\hat{\mu}_i(t) + \beta_i(t), 1\} \ \forall i \in [K]$;
  $\mu_i^{\mathsf{LCB}}(t) \leftarrow \max\{\hat{\mu}_i(t) - \beta_i(t), 0\} \ \forall i \in [K]$;
  $m_t = \arg\max_i \mu_i^{\mathsf{LCB}}(t)$;
  $Feas(t) = \{i : \mu_i^{\mathsf{UCB}}(t) \geq (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t)\}$;
  $I_t = \arg\min_{i \in Feas(t)} c_i$;
  Pull arm I_t to obtain reward r_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\} \ \forall i \in [K]$;
end

Theorem 3. *For an instance ϕ with K arms, when the number of exploration pulls of each arm $\tau = (T/K)^{2/3}$, then the sum of cost and quality regret incurred by CS-ETC(Algorithm 2) on any instance ϕ i.e.* $\mathsf{Quality\_Reg}_{CS-ETC}(T, \alpha, \phi) + \mathsf{Cost\_Reg}_{CS-ETC}(T, \alpha, \phi)$ *is* $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

The key reason that sufficient exploration is needed for our problem is that there can be arms with mean rewards very close to each other but significantly different costs. If cost regret were not of concern, then playing either arm would have led to satisfactory performance by giving low quality regret. But the need for performing well on both cost and quality regrets necessitates differentiating between the two arms and finding the one with the cheapest cost among the arms with mean reward above the smallest tolerated reward.

The regret guarantee mainly stems from the exploration phase of the algorithm. In fact, an algorithm which estimates the optimal arm only once after the exploration phase and pulls that arm for the remaining time will have the same regret upper bound as CS-ETC. But we empirically observed that the non-asymptotic performance of this algorithm is worse as compared to Algorithm 2.

5 Performance With Constraints on Costs and Rewards

In this section, we present some extensions of the previous results.

5.1 Consistent Cost and Quality

The lower bound result in Theorem 2 is motivated by an extreme instance where arms with very similar mean rewards have very different costs. This raises the following question - can better performing algorithms be obtained if the rewards and costs are *consistent* with each other? We show that this is indeed the case. Motivated by the instance which led to the worst case performance, we consider a constraint which gives an upper bound on the difference in costs of every pair of arms by a multiple of the difference in the qualities of these arms. Under this constraint, CS-UCB has good performance as per the following result with the proof in Appendix C.

Theorem 4. *If for an instance ϕ with K arms, $|c_i - c_j| \leq \delta|\mu_i - \mu_j| \ \forall i, j \in [K]$ and any (possibly unknown) $\delta > 0$, then* $\mathsf{Quality\_Reg}_{CS-UCB}(T, \alpha, \phi) + \mathsf{Cost\_Reg}_{CS-UCB}(T, \alpha, \phi)$ *is* $O((1+\delta)\sqrt{KT \log T})$.

Note that, in general, δ can be unknown. Hence, even with the above assumption on consistency of cost and quality, a priori any algorithm cannot get a bound on the quality difference between arms, only by virtue of knowing their costs.

5.2 Unknown Costs

In some applications, it is possible that the costs of the arms are also unknown and random. Hence, in addition to the mean reward, the mean costs also need to be estimated. Without loss of generality, we assume that the distribution of the random cost of each arm has support [0,1]. Not knowing the cost of the arm, does not fundamentally change the regret minimization problem we have discussed in the above sections. Clearly, the lower bound result is still valid. Algorithm 2 can be generalized to the unknown costs setting with a minor modification in the UCB phase of the algorithm. The modified UCB phase is described in Algorithm 4 in Appendix D. In this algorithm, we maintain confidence bounds on the costs of each arm. Instead of picking the arm with the lowest cost among all feasible arms, the algorithm now picks the arm with the lowest lower confidence bound on cost. Theorem 3 holds for this modified algorithm also.

Similarly, when costs and quality are consistent as described in Section 5.1, the CS-UCB algorithm can be modified to pick the arm with the lowest lower confi-

Parameter	Value
Mean reward of arm 1 (μ_1)	0.5
Mean reward of arm 2 (μ_2)	0.3-0.6
Cost of arm 1 (c_1)	1
Cost of arm 2 (c_2)	0
Subsidy factor (α)	0.1
Time horizon (T)	5000

Table 1: Parameter values

dence bound on cost and Theorem 4 holds.

6 Numerical Experiments

In the previous sections, we have shown theoretical results on the worst case performance of different algorithms for (1). Now, we illustrate the empirical performance of these algorithms. We shed light on which algorithm performs better in what regime of parameter values. The key quantity which differentiates the performance of different algorithms is how close the mean rewards of different arms are to each other. We consider a setting with two Bernoulli arms and vary the mean reward of one arm (the cheaper arm) while keeping the other quantities (reward distribution of the other arm and costs of both arms) fixed. The values of these parameters are described in Table 1. The reward in each round follows a Bernoulli distribution whereas the cost is a known fixed value. The cost and quality regret at time T of the different algorithms are plotted in Figure 2.

We observe that the performance of the CS-TS and CS-UCB are close to each other for the entire range of mean reward values. To compare the performance of these algorithms with CS-ETC, we focus on how close the mean reward of the lower mean reward arm is to the smallest tolerated reward. When $\mu_2 \leq 0.5$ ($\mu_2 > 0.5$), the lowest tolerated reward is 0.45 ($0.9\mu_2$). In terms of quality regret, when μ_2 is much smaller than 0.45, CS-TS and CS-UCB perform much better than CS-ETC. This is because the number of exploration rounds in the CS-ETC algorithm is fixed (independent of the difference in mean rewards of the two arms) leading to higher quality regret when μ_2 is much smaller than 0.45. On the other hand, because of the large difference in μ_2 and 0.45, CS-TS and CS-UCB algorithms are easily able to find the optimal arm and incur low quality regret. The cost regret of all algorithms is 0 because the optimal arm is the expensive arm.

When μ_2 is close to (and less than) 0.45, CS-TS and CS-UCB incur much higher cost regret as compared to CS-ETC. This is in line with the intuition established

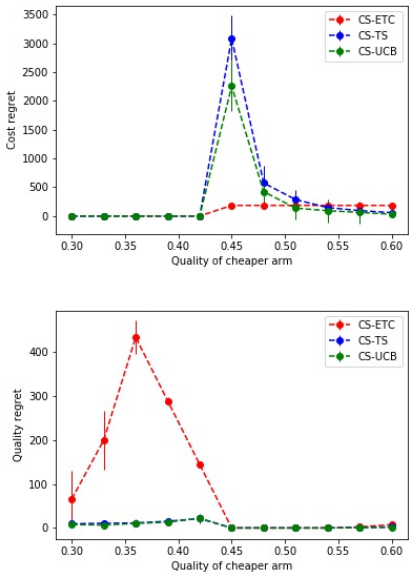

Figure 2: Performance of algorithms with varying mean reward of the cheaper arm. The length of the error bars correspond to two standard deviations in regret obtained by running the experiment 50 times.

in Section 2. Here, CS-TS and CS-UCB are unable to effectively conclude that the second (cheaper) arm is optimal. Thus, they end up pulling the first (expensive) arm many times leading to high cost regret. On the other hand, CS-ETC, after the exploration rounds is able to correctly identify the second arm as the optimal arm.

Thus, we recommend using the CS-TS/CS-UCB algorithm when the mean rewards of arms are well differentiated and CS-ETC when the mean rewards are close to one another (as is often the case in the SMS application). This is in line with the notion that algorithms which perform well in the worst case might not have good performance for an average case.

6.1 Conclusion and Future Work

In this paper, we have proposed a new variant of the MAB problem which factors costs associated with playing an arm and introduces new metrics that uniquely capture the features of multiple real world applications. We argue about the *hardness* of this problem by establishing fundamental limits on performance of any online algorithm and also demonstrating that traditional MAB algorithms perform poorly from a both theoretical and empirical standpoint. We present a simple near-optimal algorithm and through numerical simulations, we prescribe ideal algorithmic choice for different problem regimes.

An important question that naturally arises from this work is developing an algorithm for the adversarial variant of the MAB with cost subsidy problem. In particular, it is not immediately clear if EXP3 (Auer et al. (2002b)) family of algorithms, that are popular for non-stochastic MAB problem can be generalized to setting where the reward distribution is not stationary.

Acknowledgements

We would like to thank Nicolas Stier-Moses for introducing us to the SMS routing application and helping us in formulating it as an MAB problem. We would also like to thank three anonymous reviewers for their detailed feedback which helped in sharpening our arguments.

References

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653.

S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5), 2017a.

S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017b.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. A near-optimal exploration-exploitation approach for assortment selection. *Proceedings of the 2016 ACM Conference on Economics and Computation (EC)*, pages 599–600, 2016.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pages 76–78, 2017.

S. Amani, M. Alizadeh, and C. Thrampoulidis. Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3562–3566. IEEE, 2020.

Amazon. Amazon auto-targeting, 2019. URL `https://tinyurl.com/yx9lyfwq`.

A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *J. ACM*, 65(3):13:1–13:55, Mar. 2018. ISSN 0004-5411. doi: 10.1145/3164539. URL `http://doi.acm.org/10.1145/3164539`.

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

M. Canlas, K. P. Cruz, M. K. Dimarucut, P. Uyengco, G. Tangonan, M. L. Guico, N. Libatique, and C. Pineda. A quantitative analysis of the quality of service of short message service in the philippines. In *2010 IEEE International Conference on Communication Systems*, pages 710–714. IEEE, 2010.

W. Cao, J. Li, Y. Tao, and Z. Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 1036–1044, 2015.

N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2014.

L. Chen, A. Gupta, and J. Li. Pure exploration of multi-armed bandit under matroid constraints. In *Conference on Learning Theory*, pages 647–669, 2016.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

S. Daulton, S. Singh, V. Avadhanula, D. Dimmery, and E. Bakshy. Thompson sampling for contextual bandit problems with auxiliary safety constraints. *arXiv preprint arXiv:1911.00638*, 2019.

M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013.

Facebook. Facebook targeting expansion, 2016. URL `https://tinyurl.com/y3ss2j8g`.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

Google. Google auto-targeting, 2014. URL `https://tinyurl.com/y3c4bdaj`.

M. Huber. Nearly optimal bernoulli factories for linear functions. *arXiv preprint arXiv:1308.1562*, 2013.

N. Immorlica, K. A. Sankararaman, R. Schapire, and A. Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219, 2019.

K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

J. Katz-Samuels and C. Scott. Top feasible arm identification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1593–1601, 2019.

A. Kazerouni, M. Ghavamzadeh, Y. A. Yadkori, and B. Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3910–3919, 2017.

M. Keane and G. L. O'Brien. A bernoulli factory. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 4(2):213–219, 1994.

R. Koningstein. Suggesting and/or providing targeting information for advertisements, July 6 2006. US Patent App. 11/026,508.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934, 2015.

MarketWatch. Marketwatch a2p report, 2020. URL `https://rb.gy/0w96oi`.

B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, (13): 2069–2106, 2012.

X. Meng, P. Zerfos, V. Samanta, S. H. Wong, and S. Lu. Analysis of the reliability of a nationwide short message service. In *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pages 1811–1819. IEEE, 2007.

C. Oliver and L. Li. An empirical evaluation of thompson sampling. *In Advances in Neural Information Processing Systems (NIPS)*, 24:2249?2257, 2011.

O. Osunade and S. O. Nurudeen. Route optimization for delivery of short message service in telecommunication networks.

Ovum. Sustaining a2p sms growth while securing mobile network, 2017. URL `https://rb.gy/qqonzd`.

B. Paria, K. Kandasamy, and B. Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. *arXiv preprint arXiv:1805.12168*, 2018.

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

A. Sankararaman, A. Ganesh, and S. Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

S. L. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237. URL `http://dx.doi.org/10.1561/2200000068`.

A. Slivkins and J. W. Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *ACM SIGecom Exchanges*, 12(2):4–23, 2014.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

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.

Twilio and Uber. Uber built a great ridesharing experience with sms & voice, 2020. URL `https://customers.twilio.com/208/uber/`.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262, 2016.

S. Yahyaa and B. Manderick. Thompson sampling for multi-objective multi-armed bandits problem. In *Proceedings*, page 47. Presses universitaires de Louvain, 2015.

S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Annealing-pareto multi-objective multi-armed bandit algorithm. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–8. IEEE, 2014.

P. Zerfos, X. Meng, S. H. Wong, V. Samanta, and S. Lu. A study of the short message service of a nationwide cellular network. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 263–268, 2006.

Multi-armed Bandits with Cost Subsidy: Supplementary Material

Outline The supplementary material of the paper is organized as follows.

- Appendix A contains technical lemmas used in subsequent proofs.
- Appendix B contains a proof of the lower bound.
- Appendix C contains proofs related to the performance of various algorithms presented in the paper.
- Appendix D gives a detailed description of the CS-ETCalgorithm when the costs of the arms are unknown and random.

A Technical Lemmas

Lemma 2 (Taylor's Series Approximation)**.** *For* $x > 0,\ \ln(1+x) \geq x - \frac{x^2}{1-x^2}$.

Proof. For $x > 0$,

$$\begin{aligned}
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
&\geq x - \frac{x^2}{2} - \frac{x^4}{4} - \cdots && \text{(because } x > 0\text{)} \\
&\geq x - x^2 - x^4 \\
&= x - x^2(1 + x^2 + x^4 + \cdots) \\
&= x - \frac{x^2}{1-x^2}.
\end{aligned}$$

□

Lemma 3 (Taylor's Series Approximation)**.** *For* $x > 0,\ \ln(1-x) \geq -x - \frac{x^2}{1-x}$.

Proof. For $x > 0$,

$$\begin{aligned}
\ln(1-x) &= -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
&\geq -x - x^2 - x^3 - x^4 - \cdots && \text{(because } x > 0\text{)} \\
&= -x - x^2(1 + x + x^2 + \cdots) \\
&= -x - \frac{x^2}{1-x}.
\end{aligned}$$

□

Lemma 4 (Pinsker's inequality)**.** *Let* $Ber(x)$ *denote a Bernoulli distribution with mean* x *where* $0 \leq x \leq 1$*. Then,* $KL(Ber(p); Ber(p+\epsilon)) \leq \frac{4\epsilon^2}{p}$ *where* $0 < p \leq \frac{1}{2}$*,* $0 < \epsilon \leq \frac{p}{2}$ *and* $p + \epsilon < 1$ *and the KL divergence between two Bernoulli distributions with mean* x *and* y *is given as* $KL(Ber(x); Ber(y)) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$.

Proof. $KL(Ber(p); Ber(p+\epsilon)) = p \ln \frac{p}{p+\epsilon} + (1-p) \ln \frac{1-p}{1-p-\epsilon}$

$$\begin{aligned}
KL(Ber(p); Ber(p+\epsilon)) &= p \ln \frac{p}{p+\epsilon} + (1-p) \ln \frac{1-p}{1-p-\epsilon} \\
&= -p \ln \left(1 + \frac{\epsilon}{p}\right) - (1-p) \ln \left(1 - \frac{\epsilon}{1-p}\right) \\
&\leq -p \left(\frac{\epsilon}{p} - \frac{\frac{\epsilon^2}{p^2}}{1 - \frac{\epsilon^2}{p^2}}\right) - (1-p) \left(-\frac{\epsilon}{1-p} - \frac{\frac{\epsilon^2}{(1-p)^2}}{1 - \frac{\epsilon}{1-p}}\right) \\
&\qquad \text{(Using Lemmas 2 and 3.)}
\end{aligned}$$

Thus,

$$\begin{aligned}
KL(Ber(p); Ber(p+\epsilon)) &\leq -\epsilon + \frac{\epsilon^2}{p\left(1 - \frac{\epsilon^2}{p^2}\right)} + \epsilon + \frac{\epsilon^2}{(1-p)\left(1 - \frac{\epsilon}{1-p}\right)} \\
&\leq \frac{\epsilon^2}{p\left(1 - \frac{1}{4}\right)} + \frac{\epsilon^2}{(1-p)\left(1 - \frac{1}{2}\right)} \qquad \text{(because } \frac{\epsilon}{1-p} \leq \frac{\epsilon}{p} \leq \frac{1}{2}\text{)} \\
&= \frac{4\epsilon^2}{3p} + \frac{2\epsilon^2}{1-p} \\
&\leq \frac{2\epsilon^2}{p} + \frac{2\epsilon^2}{1-p} \\
&\leq \frac{2\epsilon^2}{p} + \frac{2\epsilon^2}{p} \qquad \text{(because } p \leq \frac{1}{2}\text{)} \\
&= \frac{4\epsilon^2}{p}.
\end{aligned}$$

□

B Proof of Lower Bound

Proof of Lemma 1. In the family of instances $\Phi_{\theta,p,\epsilon}$, the costs of the arms are same across instances. Arm 0 is the cheapest arm in all the instances. With this, we define a modified notion of quality regret which penalizes the regret only when this cheap arm is pulled as

$$\mathsf{Mod\_Quality\_Reg}_{\pi}(T, \alpha, \boldsymbol{\mu}, \boldsymbol{c}) = \sum_{t=1}^{T} \max\{\mu_{m*} - \mu_{\pi_t}, 0\} \mathbf{I}(c_{i_t} = 0). \tag{2}$$

An equivalent notation for denoting the modified regret of policy π on an instance I of the problem is $\mathsf{Mod\_Quality\_Reg}_{\pi}(T, \alpha, I)$. This modified quality regret is at most equal to the quality regret. For proving the lemma, we will show a stronger result that there exists an instance $\phi_{0,p,\epsilon}$ such that $\mathsf{Mod\_Quality\_Reg}(T, 0, \phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}(T, 0, \phi_{0,p,\epsilon})$ is $\Omega\left(pK^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ which will imply the required result.

Let us first consider any deterministic policy (or algorithm) π. For a deterministic algorithm, the number of times an arm is pulled is a function of the observed rewards. Let the number of times arm j is played be denoted by N_j and let the total number of times any arm with cost 1 i.e. an expensive arm is played be $N_{exp} = 1 - N_0$. For any a such that $1 \leq a \leq K$, we can use the proof of Lemma A.1 in Auer et al. (2002b), with function $f(\mathbf{r}) = N_{exp}$ to get

$$\mathbb{E}^a[N_{exp}] \leq \mathbb{E}^0[N_{exp}] + 0.5T\sqrt{2\mathbb{E}^0[N_a] KL(Ber(p); Ber(p+\epsilon))}$$

where $\mathbb{E}^j$ is the expectation operator with respect to the probability distribution defined by the random rewards in instance $\Phi^j_{0,p,\epsilon}$. Thus, using Lemma 4, we get,

$$\mathbb{E}^a[N_{exp}] \leq \mathbb{E}^0[N_{exp}] + 0.5T\sqrt{\mathbb{E}^0[N_a] 8\epsilon^2/p}. \tag{3}$$

Now, let us look at the regret of the algorithm for each instance in the family $\Phi_{0,p,\epsilon}$. We have

1. $\mathsf{Cost\_Reg}_\pi(T,\alpha,\Phi^0_{0,p,\epsilon}) = \mathbb{E}^0[N_{exp}],\ \mathsf{Mod\_Quality\_Reg}_\pi(T,\alpha,\Phi^0_{0,p,\epsilon}) = 0$
2. $\mathsf{Cost\_Reg}_\pi(T,\alpha,\Phi^a_{0,p,\epsilon}) = 0,\ \mathsf{Mod\_Quality\_Reg}_\pi(T,\alpha,\Phi^a_{0,p,\epsilon}) = \epsilon\,(T - \mathbb{E}^a[N_{exp}])$.

Now, define randomized instance $\phi_{0,p,\epsilon}$ as the instance obtained by randomly choosing from the family of instances $\Phi_{0,p,\epsilon}$ such that $\phi_{0,p,\epsilon} = \Phi^0_{0,p,\epsilon}$ with probability $1/2$ and $\phi_{0,p,\epsilon} = \Phi^a_{0,p,\epsilon}$ with probability $1/2K$ for $1 \le a \le K$. The expected regret of this randomized instance is

$$
\begin{aligned}
&\mathbb{E}\left[\mathsf{Mod\_Quality\_Reg}_\pi(T,0,\phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,0,\phi_{0,p,\epsilon})\right] \\
&= \frac{1}{2}\left(\mathsf{Mod\_Quality\_Reg}_\pi(T,\alpha,\Phi^0_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,\alpha,\Phi^0_{0,p,\epsilon})\right) + \\
&\quad \frac{1}{2K}\sum_{a=1}^{K}\left(\mathsf{Mod\_Quality\_Reg}_\pi(T,\alpha,\Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,\alpha,\Phi^a_{0,p,\epsilon})\right) \\
&= \frac{1}{2}\mathbb{E}^0[N_{exp}] + \frac{1}{2K}\sum_{a=1}^{K}\epsilon(T - \mathbb{E}^a[N_{exp}]) \\
&\ge \frac{1}{2}\mathbb{E}^0[N_{exp}] + \frac{1}{2K}\sum_{a=1}^{K}\epsilon\left(T - \mathbb{E}^0\left[N_{exp}\right] - \frac{1}{2}T\sqrt{\mathbb{E}^0[N_a]\frac{8\epsilon^2}{p}}\right) \qquad \text{(using (3))} \\
&= \frac{1}{2}\left[\epsilon T + (1-\epsilon)\sum_{a=1}^{K}\mathbb{E}^0[N_a] - \frac{T\epsilon}{2K}\sum_{a=1}^{K}\sqrt{\frac{8\epsilon^2}{p}\mathbb{E}^0[N_a]}\right] \\
&= \frac{1}{2}\sum_{a=1}^{K}\left[\frac{\epsilon T}{K} + (1-\epsilon)(\mathbb{E}^0[N_a])^2 - T\mathbb{E}^0[N_a]\epsilon^2\frac{\sqrt{2}}{K\sqrt{p}}\right] \\
&= \frac{1}{2}\sum_{a=1}^{K}\left[\left(\sqrt{1-\epsilon}\mathbb{E}^0[N_a] - \frac{\epsilon^2 T}{2K}\sqrt{\frac{2}{p(1-\epsilon)}}\right)^2 + \frac{\epsilon T}{K} - \frac{\epsilon^4 T^2}{2pK^2(1-\epsilon)}\right] \\
&\ge \frac{1}{2}\sum_{a=1}^{K}\frac{\epsilon T}{K} - \frac{\epsilon^4 T^2}{2pK^2(1-\epsilon)} \\
&= \frac{\epsilon T}{2} - \frac{\epsilon^4 T^2}{4pK(1-\epsilon)}
\end{aligned}
$$

Taking $\epsilon = \frac{p}{2}(\frac{K}{T})^{\frac{1}{3}}$, we get $\mathbb{E}\left[\mathsf{Mod\_Quality\_Reg}_\pi(T,0,\phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,0,\phi_{0,p,\epsilon})\right]$ is $\Omega(pK^{1/3}T^{2/3})$ when $K \le T$.

Using Yao's principle, for any randomized algorithm π, there exists an instance $\Phi^j_{0,p,\epsilon}$ with $0 \le j \le K$ such that $\mathsf{Mod\_Quality\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon})$ is $\Omega(pK^{1/3}T^{2/3})$. Also, since $\mathsf{Mod\_Quality\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon}) \le \mathsf{Quality\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon})$, we have $\mathsf{Quality\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T,0,\Phi^j_{0,p,\epsilon})$ is $\Omega(pK^{1/3}T^{2/3})$. □

Proof of Theorem 2. **Notation**: For any instance ϕ, we define the arms m^ϕ_* and i^ϕ_* as $m^\phi_* = \arg\max_i \mu^i_\phi$ and $i^\phi_* = \arg\min_i c^i_\phi$ s.t. $q_{i_\phi} \ge (1-\theta)q_{m^*_\phi}$. When the instance is clear, we will use the simplified notation i_* and m_* instead of i^ϕ_* and m^ϕ_*.

Proof Sketch: Lemma 1 establishes that when $\alpha = 0$, for any given policy, there exists an instance on which the sum of quality and cost regret are $\Omega(K^{1/3}T^{2/3})$. Now, we generalize the above result for $\alpha = 0$ to any α for $0 \le \alpha \le 1$. The main idea in our reduction is to show that if there exists an algorithm π_α for $\alpha > 0$ that achieves $o(K^{1/3}T^{2/3})$ regret on every instance in the family $\Phi_{\alpha,p,\epsilon}$, then we can use π_α as a subroutine

to construct an algorithm π_0 for problem (1) that achieves $o(K^{1/3}T^{2/3})$ regret on every instance in the family $\Phi_{0,p,\epsilon}$, thus contradicting the lower bound of Lemma 1. This will prove the theorem by contradiction. In order to construct the aforementioned sub-routine, we leverage techniques from *Bernoulli factory* to generate a sample from a Bernoulli random variable with parameter $\mu/(1-\alpha)$ using samples from a Bernoulli random variable with parameter μ, for any $\mu, \alpha < 1$.

Aside on Bernoulli Factory: The key tool we use in constructing the algorithm π_0 from π_α is *Bernoulli factory* for the linear function. The Bernoulli factory for a specified scaling factor $C > 1$ i.e. *BernoulliFactory*(C) uses a sequence of independent and identically distributed samples from $Ber(r)$ and returns a sample from $Ber(Cr)$.The key aspect of a Bernoulli factory is the number of samples needed from $Ber(r)$ to generate a sample from $Ber(Cr)$. We use the Bernoulli factory described in Huber (2013) which has a guarantee on the expected number of samples τ from $Ber(r)$ needed to generate a sample from $Ber(Cr)$. In particular, for a specified $\delta > 0$,

$$\sup_{r \in [0, \frac{1-\delta}{C}]} E[\tau] \leq \frac{9.5C}{\delta}. \tag{4}$$

Detailed proof: For some value of p, ϵ (to be specified later in the proof) such that $0 \leq p < 1$ and $0 \leq \epsilon \leq p/2$, consider the family of instances $\Phi_{\alpha,p,\epsilon}$ and $\Phi_{0,p,\epsilon}$. Let π_α be any algorithm for the family $\Phi_{\alpha,p,\epsilon}$. Using π_α, we construct an algorithm π_0 for the family $\Phi_{0,p,\epsilon}$. This algorithm is described in Algorithm 3. We will use $I_l^\alpha = \pi_\alpha([(I_1^\alpha, r_1), (I_2^\alpha, r_2), \cdots (I_{l-1}^\alpha, r_{l-1})])$ to denote the arm pulled by algorithm π_α at time l after having observed rewards r_l $\forall 1 \leq i < l$ through arm pulls I_l $\forall 1 \leq i < l$. The function *BernoulliFactory*(C) returns two values - a random sample from the distribution $Ber(Cr)$ and the number of samples of $Ber(r)$ needed to generate this random sample.

Algorithm 3: Derived Algorithm π_0

Result: Arm I_t^0 to be pulled in each round t, total number of arm pulls T
input : Algorithm π_α, L - Number of arm pulls for algorithm π_α
$l = 1, t = 1$;
for $l \in [L]$ **do**
  $I_l^\alpha = \pi_\alpha([(I_1^\alpha, r_1), (I_2^\alpha, r_2), \cdots (I_{l-1}^\alpha, r_{l-1})])$;
  if $I_l^\alpha = 0$ **then**
    Pull arm 0 to obtain outcome r_l ;
    $I_t^0 = I_l^\alpha = 0$;
    $U_l = \{t\}$;
  else
    Call $r_l, n =$ *BernoulliFactory*$(\frac{1}{1-\alpha})$ on samples generated from repeated pulls of the arm I_l^α ;
    $U_l = \{t, t+1 \cdots t+n-1\}$;
    $I_t^0 = I_{t+1}^0 = \cdots I_{t+n-1}^0 = I_l^\alpha$;
  end
  $S_l = |U_l|$;
  $l = l + 1$;
  $t = t + S_l$;
end
$T = t$

Now, let us analyze the expected modified regret incurred by algorithm π_0 on an instance $\Phi_{0,p,\epsilon}^a$ for any $0 \leq a \leq K$ where the expectation is with respect to the random variable T, total number of arm pulls.

Similarly, we analyze the cost regret incurred by algorithm π_0 on an instance $\Phi_{0,p,\epsilon}^a$ for any $0 \leq a \leq K$.

$$
\begin{aligned}
&\mathbb{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(T,0,\Phi^a_{0,p,\epsilon})\right] + \mathbb{E}\left[\mathsf{Cost\_Reg}_{\pi_0}(T,0,\Phi^a_{0,p,\epsilon})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_t^0=0\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{L}\sum_{t\in U_l}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_t^0=0\}\right] + \mathbb{E}\left[\sum_{l=1}^{L}\sum_{t\in U_l} c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{L} S_l\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha=0\}\right] + \mathbb{E}\left[\sum_{l=1}^{L} S_l\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \\
&= \sum_{l=1}^{L}\mathbb{E}\left[\mathbb{E}\left[S_l|I_l^\alpha\right]\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha=0\}\right] + \mathbb{E}\left[\sum_{l=1}^{L}\mathbb{E}[S_l|I_l^\alpha]\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \\
&\le \sum_{l=1}^{L}\mathbb{E}\left[\frac{9.5}{\delta(1-\alpha)}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha=0\}\right] + \mathbb{E}\left[\sum_{l=1}^{L}\frac{9.5}{\delta(1-\alpha)}\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \qquad \text{(Using (4))} \\
&= \frac{9.5}{\delta(1-\alpha)}\sum_{l=1}^{L}\mathbb{E}\left[\left((1-\alpha)\mu_{m^*}^{\Phi^a_{\alpha,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha=0\}\right] + \frac{9.5}{\delta(1-\alpha)}\sum_{l=1}^{L}\mathbb{E}\left[c_{i_*}^{\Phi^a_{\alpha,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{\alpha,p,\epsilon}}\right] \\
&\quad \text{(Because costs of arms are same in all instances, } i_*^{\Phi^a_{\alpha,p,\epsilon}} = i_*^{\Phi^a_{0,p,\epsilon}} = a \text{ and } \mu_{m^*}^{\Phi^a_{0,p,\epsilon}} = (1-\alpha)\mu_{m^*}^{\Phi^a_{\alpha,p,\epsilon}}\ ) \\
&= \frac{9.5}{\delta(1-\alpha)}\mathsf{Quality\_Reg}_{\pi_\alpha}(L,\alpha,\Phi^a_{\alpha,p,\epsilon}) + \frac{9.5}{\delta(1-\alpha)}\mathsf{Cost\_Reg}_{\pi_\alpha}(L,\alpha,\Phi^a_{\alpha,p,\epsilon}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathsf{Quality\_Reg}_{\pi_\alpha}(L,\alpha,\Phi^a_{\alpha,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_\alpha}(L,\alpha,\Phi^a_{\alpha,p,\epsilon}) \\
&\ge \frac{\delta(1-\alpha)}{9.5}\mathbb{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(T,0,\Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(T,0,\Phi^a_{0,p,\epsilon})\right] \qquad (5) \\
&\ge \frac{\delta(1-\alpha)}{9.5}\mathbb{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(L,0,\Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(L,0,\Phi^a_{0,p,\epsilon})\right] \qquad \text{(because } L\le T\text{)} \\
&\ge \frac{\delta(1-\alpha)}{9.5}\left(\mathsf{Mod\_Quality\_Reg}_{\pi_0}(L,0,\Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(L,0,\Phi^a_{0,p,\epsilon})\right) \qquad (6)
\end{aligned}
$$

Using Lemma 1 and choosing $p = \frac{1-\alpha}{3}, \delta = \frac{1}{2}, \epsilon = \frac{p}{2}(\frac{K}{T})^{1/3}$, we get for any randomized algorithm π_α, there exists instance $\Phi^b_{\alpha,p,\epsilon}$ (for some $0 \le b \le K$) such that $\mathsf{Quality\_Reg}_{\pi}(T,\alpha,\Phi^b_{\alpha,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi}(T,\alpha,\Phi^b_{\alpha,p,\epsilon})$ is $\Omega\left((1-\alpha)^2K^{1/3}T^{2/3}\right)$.

□

C Performance of Algorithms

We use the following fact in the proof of Theorem 1.

Fact 1. *(Abramowitz and Stegun, 1948) For a Normal random variable Z with mean m and variance σ^2, for any z,*

$$\Pr\left(|Z-m| > z\sigma\right) > \frac{1}{4\sqrt{\pi}}\exp(-\frac{7z^2}{2}).$$

Proof of Theorem 1. This proof is inspired by the lower bound proof in Agrawal and Goyal (2017b). For any given α, K and T, we construct an instance on which the CS-TS algorithm (Algorithm 1) gives linear regret in cost.

Consider an instance ϕ with K arms where the costs and mean reward of the j-th arm are

$$c_j = \begin{cases} 0 & j=0 \\ 1 & j\neq 0 \end{cases}, \qquad \mu_j = \begin{cases} (1-\alpha)q + \frac{d}{\sqrt{T}} & j=0 \\ q & j\neq 0 \end{cases}$$

where $q = \frac{d}{(1-\alpha)\sqrt{T}}$ for some $0 < d < \min\{\sqrt{T}/2, (1-\alpha)\sqrt{T}\}$. Moreover, the reward of each arm is deterministic though this fact is not known to the agent. As in the SMS application, we assume that the cost rewards of all arms are known a priori to the agent.

Let the prior distribution that the agent assumes over the mean reward of each arm be $\mathcal{N}(0, \sigma_0^2)$ for some prior variance σ_0^2 . Further, the agent assumes that the observed qualities to be normally distributed with noise variance σ_n^2. As such at the start of period t, the agent will consider a normal posterior distribution for each arm i with mean

$$\hat{\mu}_i(t) = \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2} + T_i(t)}\mu_i \tag{7}$$

and variance

$$\sigma_i(t)^2 = \left(\frac{1}{\sigma_0^2} + \frac{T_i(t)}{\sigma_n^2}\right)^{-1}. \tag{8}$$

As $d < q\alpha\sqrt{T}$, the highest quality across all arms is q. Thus, note that all arms are *feasible* in terms of quality i.e. have their quality within $(1-\alpha)$ factor of the best quality arm. Hence, quality regret $\mathsf{Quality\_Reg}_{\mathsf{CS-TS}}(t, \alpha, \phi) = 0\ \forall t > 0$ (for any algorithm) on this instance.

The first arm is the optimal arm (i_*). Thus, the cost regret equals the number of times any arm but the first arm is pulled. In particular, let

$$R_c(T) = \sum_{t=1}^{T} \max\{c_{I_t} - c_{i_*}, 0\} = \sum_{t=1}^{T} \mathbf{1}\{I_t \neq 1\},$$

so that $\mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T, \alpha, I) = \mathbb{E}\left[R_C(T)\right]$.

Define the event $A_{t-1} = \{\sum_{i\neq 1} T_i(t) \leq sT\sqrt{K}\}$ for a fixed constant $s > 0$. For any t, if the event A_{t-1} is not true, then $R_c(T) \geq R_c(t) \geq sT\sqrt{K}$. We can assume that $\Pr(A_{t-1}) \geq 0.5\ \forall t \leq T$. Otherwise

$$\begin{aligned}
\mathsf{Cost_Reg}_{\mathsf{CS-TS}}(T, \alpha, \phi) &= \mathbb{E}\left[R_C(T)\right] \\
&\geq 0.5E[R_C(T)|A_{t-1}^c] \\
&= \Omega(T\sqrt{K}).
\end{aligned}$$

Now, we will show that whenever A_{t-1} is true, probability of playing a sub-optimal arm is at least a constant. For this, we show that the probability that $\mu_1^{score}(t) \leq \mu_1$ and $\mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}$, for some $1 < i \leq K$ is lower bounded by a constant.

Now, given any history of arm pulls $\mathcal{F}_{t-1}$ before time t, $\mu_1^{score}(t)$ is a Gaussian random variable with mean $\hat{\mu}_1(t) = \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2}+T_i(t)}\mu_1$. By symmetry of Gaussian random variables, we have

$$\begin{aligned}
\Pr\left(\mu_1^{score}(t) \leq \mu_1 \middle| \mathcal{F}_{t-1}\right) &\geq \Pr\left(\mu_1^{score}(t) \leq \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2} + T_i(t)}\mu_1 \middle| \mathcal{F}_{t-1}\right) \\
&= \Pr\left(\mu_1^{score}(t) \leq \hat{\mu}_1(t) \middle| \mathcal{F}_{t-1}\right) \\
&= 0.5.
\end{aligned}$$

Based on (7) and (8), given any realization F_{t-1} of $\mathcal{F}_{t-1}$, $\mu_i^{score}(t)$ for $i \neq 1$ are independent Gaussian random variables with mean $\hat{\mu}_i(t)$ and variance $\sigma_i(t)^2$. Thus, we have

$$
\begin{aligned}
&\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{1}{1-\alpha}\left(q(1-\alpha) + \frac{d}{\sqrt{T}}\right) - \hat{\mu}_i(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{d}{(1-\alpha)\sqrt{T}} + \frac{1}{1 + T_i(t)\frac{\sigma_0^2}{\sigma_n^2}} q \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&\geq \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{d}{(1-\alpha)\sqrt{T}} + q \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{2d}{(1-\alpha)\sqrt{T}} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ (\mu_i^{score}(t) - \hat{\mu}_i(t))\frac{1}{\sigma_i(t)} \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ Z_i(t) \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right)
\end{aligned}
$$

where $Z_i(t)$ are independent standard normal variables for all i, t. Thus,

$$
\begin{aligned}
&\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= 1 - \Pr\left(\forall i \neq 1,\ Z_i(t) \leq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= 1 - \Pi_{i\neq 1}\left(1 - \Pr\left(\ Z_i(t) \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right)\right) \\
&\geq 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\frac{1}{\sigma_i(t)^2}\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right) \qquad \text{(Using Fact 1)} \\
&= 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\left(\frac{1}{\sigma_0^2} + \frac{T_i(t)}{\sigma_n^2}\right)\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right) \\
&\geq 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_n^2}\right)T_i(t)\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right),
\end{aligned}
$$

The last inequality follows from the fact that $T_i(t) \geq 1$.

Now, when the event A_{t-1} holds, we have $\sum_{i\neq 1} T_i(t) \leq sT\sqrt{K}$. Thus, the right hand side would be minimized when $T_i(t) = \frac{sT}{\sqrt{K}}$, $\forall i \neq 1$. Substituting this value of $T_i(t)$, the right hand side reduces to $g(K) = 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-14\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_n^2}\right)\frac{s\sqrt{K}d^2}{(1-\alpha)^2}\right)\right)$. Thus, $\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right) \geq g(K)$ whenever F_{t-1} is such that A_{t-1} holds.

Probability of playing any sub-optimal arm at time t is,

$$\begin{aligned}
\Pr\left(\exists i \neq 1,\ I_t = i\right) &\geq \Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right) \\
&= \mathbb{E}\left[\Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right)\middle|\mathcal{F}_{t-1}\right] \\
&\geq \mathbb{E}\left[\Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right)\middle|\mathcal{F}_{t-1}, A_{t-1}\right]\Pr(A_{t-1}) \\
&\geq \frac{1}{2}\cdot g(K)\cdot\frac{1}{2}
\end{aligned}$$

Thus, at every time instant t, the probability of playing a sub-optimal is lower bounded by $\frac{g(K)}{4}$. This implies that the cost regret $\mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T,\alpha,\phi) \geq 0.25Tg(K)$.

□

Proof of Theorem 3. This algorithm has two phases - pure exploration and UCB. In the first phase, the algorithm pulls each arm a specified number of times (τ). In the second phase, the algorithms maintains upper and lower confidence bounds on the mean reward of each arm. Then, it estimates a *feasible* set of arms and pulls the cheapest arm in this set.

We will define the *clean event* $\mathcal{E}$ in this proof as the event that for every time $t \in [T]$ and arm $i \in [K]$, the difference between the mean reward and the empirical mean reward does not exceed the size of the confidence interval ($\beta_i(t)$) i.e. $\mathcal{E} = \{|\hat{\mu}_i(t) - \mu_i| \leq \beta_i(t),\ \forall i \in [K],\ t \in [T]\}$.

Define $\hat{t} = K\tau + 1$ as the first round in the UCB phase of the algorithm. Further, define instantaneous cost and quality regret as the regret incurred in the t-th arm pull:

$$\begin{aligned}
\mathsf{Quality_Reg}_{\pi}^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) &= \mathbb{E}\left[\max\{(1-\alpha)\mu_{m_*} - \mu_{\pi_t}, 0\}\right], \\
\mathsf{Cost_Reg}_{\pi}^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) &= \mathbb{E}\left[\max\{c_{\pi_t} - c_{i_*}, 0\}\right],
\end{aligned} \tag{9}$$

where the expectation is over the randomness in the policy π.

Let us first assume that the clean event holds. As both the instantaneous regrets are upper bounded by 1, $\sum_{t=1}^{K\tau}\mathsf{Quality\_Reg}_{\pi}^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) \leq K\tau$ and $\sum_{t=1}^{K\tau}\mathsf{Cost\_Reg}_{\pi}^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) \leq K\tau$.

Now, let us look at the UCB phase of the algorithm. Here, $\forall\ \hat{t} \leq t \leq T$, we have

$$\mu_{i_*}^{\mathsf{UCB}}(t) \geq \mu_{i_*} \geq (1-\alpha)\mu_{m_*} \geq (1-\alpha)\mu_{m_t} \geq (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t).$$

Here, the first and fourth inequality are because of the clean event. The second and third inequality are from the definition of i_* and m_* respectively.

Thus from the inequality above, the optimal arm i_* is in the set $Feas(t), \forall \hat{t} \leq t \leq T$. This implies that the arm pulled in each time step in the UCB phase, is either the optimal arm or an arm cheaper than it. Thus, instantaneous cost regret is zero for all time steps in the UCB phase of the algorithm.

Now, let us look at the quality regret in the UCB phase i.e. for any $\hat{t} \leq t \leq T$. We have

$$\mu_{I_t} + 2\beta_{I_t}(t) \geq \mu_{I_t}^{\mathsf{UCB}}(t) \geq (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t) \geq (1-\alpha)\mu_{m_*}^{\mathsf{LCB}}(t) \geq (1-\alpha)\left(\mu_{m_*} - 2\beta_{m_*}(t)\right) \geq (1-\alpha)\mu_{m_*} - 2\beta_{m_*}(t)$$

The first and fourth inequality hold because the clean event holds. The second and third inequalities follow from the definition of I_t and m_t respectively. Thus,

$$\mathsf{Quality\_Reg}_{\pi}^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}|\mathcal{E}) = (1-\alpha)\mu_{m_*} - \mu_{I_t} \leq 2\left(\beta_{I_t}(t) + \beta_{m_*}(t)\right) \leq 2\left(\sqrt{\tfrac{2\log T}{\tau}} + \sqrt{\tfrac{2\log T}{\tau}}\right) = 4\sqrt{\tfrac{2\log T}{\tau}}.$$

The total regret incurred by the algorithm is the sum of the instantaneous regrets across all time steps in the exploration and the UCB phase. Thus,

$\mathsf{Quality\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq K\tau + 4(T - K\tau)\sqrt{\frac{2\log T}{\tau}} \leq K\tau + 4T\sqrt{\frac{2\log T}{\tau}}$ and $\mathsf{Cost\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq K\tau$. Substituting $\tau = (T/K)^{2/3}$, we conclude that both cost and quality regret are $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

Now, when the clean event does not hold, the cost and quality regret are at most T each. The probability that the clean event does not hold is at most $2/T^2$ (Lemma 1.6 in Slivkins (2019)). Thus, the expected cost and quality regret obtained by averaging over the clean event holding and not holding is $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

□

Proof of Theorem 4. As in the previous proof, we will define the *clean event* $\mathcal{E}$ as the event that for every time $t \in [T]$ and arm $i \in [K]$, the difference between the mean reward and the empirical mean reward does not exceed the size of the confidence interval $(\beta_i(t))$ i.e. $\mathcal{E} = \{|\hat{\mu}_i(t) - \mu_i| \leq \beta_i(t),\ \forall i \in [K],\ t \in [T]\}$. Also, define the quality and cost gap of each arm as $\Delta_{\mu,i} = \max\{(1-\alpha)\mu_{m^*} - \mu_i, 0\}$ and $\Delta_{c,i} = \max\{c_{i_*} - c_i, 0\}$.

When the clean event does not hold, both cost and quality regrets are upper bounded by T. Let us look at the case when the clean event holds and analyze the cost and quality regret.

Quality Regret: Let t_i be the last time t when $i \in Feas(t)$ i.e. $t_i = \max\{K,\ \max\{t : i \in Feas(t)\}\}$. Thus, $T_i(T) = T_i(t_i)$.

Consider any arm i which would incur a quality regret on being pulled i.e. arm i such that $\mu_i < (1-\alpha)\mu_{m_*}$. We have

$$\mu_i + 2\beta_i(t_i) \geq \mu_i^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_{t_i}}^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_*}^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_*}.$$

The first and fourth inequality hold because of the clean event. The third inequality is from the definition of m_{t_i}.

Thus, $(1-\alpha)\mu_{m_*} - \mu_i \leq 2\beta_i(t_i)$. Using the definition of $\beta_i(t_i)$, we get $T_i(T) = T_i(t_i) \leq \frac{8\log T}{\Delta_{\mu,i}^2}$.

Using Jensen's inequality,

$$\begin{aligned}
\left(\frac{\sum_{i=1}^K T_i(T)\Delta_{\mu,i}}{T}\right)^2 &\leq \frac{\sum_{i=1}^K T_i(T)\Delta_{\mu,i}^2}{T} \\
&= \frac{\sum_{i=1:\ \Delta_{\mu,i}>0}^K T_i(T)\Delta_{\mu,i}^2}{T} \\
&\leq \sum_{i=1:\Delta_{\mu,i}>0}^K \frac{8\log T}{\Delta_{\mu,i}^2}\frac{\Delta_{\mu,i}^2}{T} \\
&= \frac{8K\log T}{T}
\end{aligned}$$

Thus, $\mathsf{Quality\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq \sqrt{8KT\log T}$.

Cost Regret: Let i be an arm such that $c_i > c_{i_*}$. Let $\tilde{t}_i$ be the last time when arm i is pulled. Thus, $i_* \notin Feas(\tilde{t}_i)$. We have, $\mu_{i_*} \leq \mu_{i_*}^{\mathsf{UCB}}(\tilde{t}_i) < \mu_i^{\mathsf{UCB}}(\tilde{t}_i) \leq \mu_i(\tilde{t}_i) + 2\sqrt{(2\log T)/T_i(\tilde{t}_i)}$. Thus,

$$T_i(T) = T_i(\tilde{t}_i) < \frac{8\log T}{(\mu_{i_*} - \mu_i)^2} \leq \frac{8\delta^2\log T}{(c_{i_*} - c_i)^2} = \frac{8\delta^2\log T}{\Delta_{c,i}^2}.$$

Using Jensen's inequality as for the case of quality regret, we get, $\mathsf{Cost\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq \sqrt{8\delta^2 KT\log T}$.

Note that the probability of the clean event is at least $1 - 2/T^2$ (Lemma 1.6 in Slivkins (2019)). Thus, the sum of the expected cost and quality regret by averaging over the clean event holding and not holding is $O((1+\delta)\sqrt{KT\log T})$.

□

D Algorithm with Unknown and Random Costs

Algorithm 4: CS-ETC with unknown costs

Result: Arm I_t to be pulled in each round $t \in [T]$
input : K, T, Number of exploration pulls τ
$T_i(1) = 0 \ \forall i \in [K]$
Pure exploration phase:
for $t \in [1, Kf(K,T)]$ **do**
  $I_t = t \mod K$;
  Pull arm I_t to obtain reward r_t and cost χ_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\} \ \forall i \in [K]$;
end
UCB phase:
for $t \in [Kf(K,T)+1, T]$ **do**
  $\hat{\mu}_i(t) \leftarrow \frac{\sum_{\tau=1}^{t-1} r_\tau \mathbf{I}\{I_\tau = i\}}{T_i(t)} \ \forall i \in [K]$;
  $\hat{c}_i(t) \leftarrow \frac{\sum_{\tau=1}^{t-1} \chi_\tau \mathbf{I}\{I_\tau = i\}}{T_i(t)} \ \forall i \in [K]$;
  $\beta_i(t) \leftarrow \sqrt{\frac{2 \log T}{T_i(t)}} \ \forall i \in [K]$;
  $\mu_i^{\mathsf{UCB}}(t) \leftarrow \min\{\hat{\mu}_i(t) + \beta_i(t), 1\} \ \forall i \in [K]$;
  $\mu_i^{\mathsf{LCB}}(t) \leftarrow \max\{\hat{\mu}_i(t) - \beta_i(t), 0\} \ \forall i \in [K]$;
  $c_i^{\mathsf{LCB}}(t) \leftarrow \max\{\hat{c}_i(t) - \beta_i(t), 0\} \ \forall i \in [K]$;
  $m_t = \arg\max_i \mu_i^{\mathsf{LCB}}(t)$;
  $Feas(t) = \{i : \mu_i^{\mathsf{UCB}}(t) > (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t)\}$;
  $I_t = \arg\min_{i \in Feas(t)} c_i^{\mathsf{LCB}}$;
  Pull arm I_t to obtain reward r_t and cost χ_t;
  $T_i(t+1) = T_i(t) + \mathbf{1}\{I_t = i\} \ \forall i \in [K]$;
end