
Cost-Ordered Feasibility for Multi-Armed Bandits with Cost Subsidy

Ishank Juneja Carlee Joe-Wong Osman Yağın
 Department of Electrical and Computer Engineering
 Carnegie Mellon University
 Pittsburgh, PA 15213
 {ijuneja,cjoewong,oyagan}@andrew.cmu.edu

Abstract

The classic multi-armed bandit (MAB) problem tackles the challenge of accruing maximum reward while making decisions under uncertainty. However, in applications, often the goal is to minimize cost subject to a constraint on the minimum permissible reward, an objective captured by multi-armed bandits with cost-subsidy (MAB-CS). Of interest to this paper is the setting where the quality (reward) constraint is specified relative to the unknown best reward and the cost of each arm is known. We characterize the expected sub-optimal samples required by any policy by proving instance-dependent lower bounds that offer new insight into the problem and are a strict generalization of prior bounds. Then, we propose an algorithm called Cost-Ordered Feasibility (COF) that leverages our insight and intelligently combine samples from all arms to gauge the feasibility of a cheap arm. Thereafter, we analyze COF to establish instance-dependent upper bounds on its expected cumulative cost and quality regret, i.e., relative to the cheapest feasible arm. Finally, we empirically validate the merits of COF, comparing it to baselines from the literature through extensive simulation experiments on the MovieLens and Goodreads datasets as well as representative synthetic instances. Not only does our paper develop qualitatively better theoretical regret upper bounds, but COF also convincingly demonstrates improved empirical performance.

1 Introduction

Multi-armed bandits (MABs) [20] is a framework for online sequential decision making that models decisions as arms that can be pulled. Pulling or sampling each arm leads to the observation of a scalar reward, the generative process for which is apriori unknown. The stationary stochastic bandits [4] framework in addition imposes the structure that the observed rewards be random variables that are independent and identically distributed. Classically in MABs, the goal is to maximize the cumulative reward across all the sampling decisions made over the problem horizon T . However, in practice, the costs associated with every sampling decision must be considered as part of the decision making. In particular, the goal in cost sensitive applications, like machine learning inference, is to minimize the cumulative cost subject to a constraint on the observed reward or quality level. For this reason, model routers are needed to map queries to the most appropriate LLM from a large selection of open source, proprietary, and task-specialized models [21]. The cost of these LLMs may vary over orders of magnitude, therefore routing must balance between expected quality and cost [16, 28, 31, 32].

Cost-subsidy framework: The quality-constrained cost-minimization setting is captured by the recently introduced multi-armed bandits with cost-subsidy (MAB-CS) framework [24, 17]. MAB-CS captures the quality constraint through threshold μ_{CS} . For a K -armed bandit instance with arms $\mathcal{A} = \{a_1, a_2, \dots, a_K\}$, the expected reward from sampling arm k is denoted μ_k . The constraint on the expected reward gives rise to a set of feasible arms $\mathcal{S} = \{a_k \in \mathcal{A} \mid \mu_k \geq \mu_{CS}\}$. In our work

μ_{CS} is specified in relation to the unknown best reward as $\mu_{CS} = (1 - \alpha)\mu^*$, where $\alpha \in (0, 1)$ is the *subsidy factor*, and μ^* is the best reward defined as $\max_{a_k \in \mathcal{A}} \mu_k$. As is convention in bandits, we use i^* to denote the arm bearing μ^* . If the cost associated with sampling arm a_k is denoted c_k , then in MAB-CS, the optimal arm a^* is defined to be the cheapest arm from the set of feasible arms. Mathematically $a^* = \arg \min_{a_i \in \mathcal{S}} c_i$ and reward borne by a^* is denoted μ_{a^*} . Application to the problem of selecting communication modalities in advertising [15] demonstrates the utility of MAB-CS. Available modalities like text message, postal mail, and doorstep solicitation shall each have an unknown expected quality (i.e., effectiveness or response rate) and a known deployment cost [23]. The business may want to program their response rate to be a fraction of the best unknown rate across modalities. This relative rate then serves as the quality constraint subject to which cost must be minimized. The MAB-CS framework is immensely useful for applications where no preset threshold has to be satisfied but the run-time performance of the system cannot be permitted to fall too far below the best available alternatives, no matter their cost. In the cost-subsidy framework decisions are made to minimize the metrics of cumulative cost and quality regret which are formally defined in Section 2. The former captures the excess cost paid over c_{a^*} and the latter captures violations of the quality constraint. The core challenge in jointly minimizing both regret metrics is exploring for cheap feasible arms while simultaneously estimating the feasibility threshold $(1 - \alpha)\mu^*$.

Gaps and challenges in prior work: Prior work by Sinha et al. [24] identified *worst-case* lower bounds on cost and quality regret for the MAB-CS problem. While these lower bounds formally establish that cost and quality regret minimization in the framework is a strictly more challenging problem than regret minimization in conventional MABs, they do not provide insight into the design of practically useful algorithms. Other research has introduced lower bounds on cost and quality regret that inform algorithm design by revealing the dependence between the problem instance and the minimum samples required under expectation [17]. However, their singular treatment of the best reward arm i^* leaves open room for improved general analysis. Moreover, Juneja et al. [17] introduced regret-minimization policies for settings in which μ_{CS} is either a known threshold, the subsidized reward of a specified reference arm, or the subsidized best reward. While their algorithms for the first two settings were order-wise optimal, their approach for the subsidized best reward setting studied in our work was not. A primary obstacle in regret minimization for this setting is determining if a candidate arm has reward that is feasible relative to $(1 - \alpha)\mu^*$. Prior work [17] addressed this challenge through a two stage approach where first i^* is identified through a best arm identification (BAI) scheme. Then, the feasibility of candidate arms is determined through pairwise comparisons. In this paper to estimate the $(1 - \alpha)\mu^*$ criterion we develop a method that progresses in a manner inspired by BAI without converging onto a single arm, avoiding the accrual of unnecessary samples.

Our contributions and their organization is as follows. In Section 2, we formalize the MAB-CS framework and provide instance-dependent lower bounds on the expected number of samples of all sub-optimal arms. We include an *entirely novel result* that jointly lower bounds the expected samples from a collection of arms. The joint bound is a structure that emerges naturally from the fact that an arm can be certified as infeasible once its reward is worse than the subsidized reward of *any* arbitrary arm, not just i^* . In Section 3 we present the *Cost-Ordered Feasibility (COF) algorithm*, which aggregates the signal from the comparisons between a candidate feasible arm and all other arms. Section 3 also includes an *upper bound on the expected cumulative cost and quality regrets* from COF. We validate the superiority of our approach on both regret metrics through *extensive simulation experiments on recommendation systems datasets* as well as on hand designed problem instances in Section 4. In Section 5, we contextualize MAB-CS, and our contributions in particular, into the broader multi-objective MAB literature. Finally in Section 6, we summarize our contributions and chart the course for future research.

2 Theoretical Framework and Lower Bounds

In this section we first formalize cost and quality regret and formulate them in terms of the expected samples of sub-optimal arms. Then, we present theoretical lower bounds on the expected number of sub-optimal arm samples incurred by any policy.

Cost and Quality Regret: Under the stationary stochastic bandit setting [4], at time t we observe scalar reward $r_t \geq 0$ by sampling arm k_t . Further $\mathbb{E}[r_t] = \mu_{k_t}$, and r_t is distributed per the stationary distribution ν_{k_t} . In the MAB framework a policy π maps observation histories to a decision. For a K -armed bandit instance with arms $\mathcal{A}$, the policy represents the mapping $\pi : \mathcal{H}_t \rightarrow \mathcal{A}$, where

$\mathcal{H}_t := \{k_1, r_1, \dots, k_{t-1}, r_{t-1}\}$ denotes the complete observation history. The goal in MAB-CS is to design π that converges onto sampling a^* in a manner that balances between the sampling of feasible arms and accrual of least cost. These objectives are captured respectively by cost regret and quality regret. They are defined in Equation 1 for a problem instance ν .

$$\text{Cost_Reg}(T, \nu, \pi) = \sum_{t=1}^T \mathbb{E}_\pi[(c_{k_t} - c_{a^*})^+], \quad \text{Quality_Reg}(T, \nu, \pi) = \sum_{t=1}^T \mathbb{E}_\pi[(\mu_{\text{CS}} - \mu_{k_t})^+]. \quad (1)$$

Here, $x^+ := \max\{x, 0\}$ is the zero-clipping operator, k_t is a random variable denoting the policy decisions, and $\mathbb{E}_\pi$ denotes expectation over bandit arm choices. The zero-clipping ensures that the incremental regret from any policy decision is non-negative. Notably, only the sampling of the optimal arm a^* has both zero incremental cost and quality regret. In practice, the zero-clipping of incremental quality regret $(\mu_{\text{CS}} - \mu_{k_t})$ penalizes only violations of the feasibility constraint while ignoring reward above μ_{CS} . Since the goal is to converge onto sampling optimal arm a^* , cost regret with zero-clipping only penalizes samples of arms more expensive than a^* while overlooking the cost accrual from cheaper arms. We denote the incremental cost and quality gaps from sampling arm a_i by $\Delta_{C,i} := c_i - c_{a^*}$ and $\Delta_{Q,i} := \mu_{\text{CS}} - \mu_i$ respectively. Using a standard regret decomposition result, in Equation 2 we rewrite regret in terms of the expected number of samples of sub-optimal arms.

$$\text{Cost_Reg}(T, \nu, \pi) = \sum_{i=1}^K \Delta_{C,i}^+ \mathbb{E}[n_i(T)], \quad \text{Quality_Reg}(T, \nu, \pi) = \sum_{i=1}^K \Delta_{Q,i}^+ \mathbb{E}[n_i(T)]. \quad (2)$$

The reformulation in Equation 2 permits the regret analysis of any policy π by simply bounding the expected number of samples of sub-optimal arms. Moreover, lower bounds on expected samples imply lower bounds on regret under the decomposition. For the remainder of the paper we assume without loss of generality that bandit arms are indexed in non-decreasing order of cost. Further we characterize the sub-optimal arms as cheap and expensive per Definition 2.1

Definition 2.1 (Cheap and expensive arms). We define *cheap arms* $\mathcal{A}^-$ to be arms with cost lower than c_{a^*} : $\mathcal{A}^- := \{a_k \in \mathcal{A} \mid k < a^*\}$. And *expensive arms* to have cost higher than c_{a^*} : $\mathcal{A}^+ := \{a_k \in \mathcal{A} \mid k > a^*\}$.

Lower Bounds for MAB-CS: This section discusses problem instance-dependent lower bounds on the expected number of samples of sub-optimal arms for the class of consistent policies. The restriction imposed by consistency is available in Appendix B and amounts to a technical condition to exclude policies that perform unevenly well on certain instances while incurring high regret on others. In addition to informing us about the least samples accrued by a policy for an arm, these lower bounds serve to elucidate the challenges that need to be overcome by any policy that works to minimize regret. The complete proofs of our lower bounds are available in Appendix C and the current section provides proof sketches.

Results from information theory relate the expected number of samples drawn by a policy for two bandit instances with overlapping arms $\mathcal{A}$, but different arm reward distributions [13]. The origin of sample lower bounds then lies in the construction of such an instance pair that differ minimally in their reward distributions but have distinct optimal arms. In the remainder of this section, we first discuss lower bounds for cheap arms $a_k \in \mathcal{A}^-$ and expensive arms $a_k \in \mathcal{A}^+$. Finally a joint lower bound on the samples of arms whose expected rewards satisfy a certain constraint is presented². A proof sketch for Theorem 2.3 builds intuition ahead of the more involved proof sketch for Theorem 2.5.

Theorem 2.2 (Lower Bound for Cheap Arms). *Expected number of samples for cheap arm a_k satisfy,*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[n_k(T)]}{\log T} \geq \frac{2}{\Delta_{Q,k}^2}, \quad \forall a_k \in \mathcal{A}^-.$$

Theorem 2.3 (Lower Bound for Expensive Arms). *Expected number of samples for expensive arm a_k satisfy,*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[n_k(T)]}{\log T} \geq \frac{2(1-\alpha)^2}{(\mu_{a^*} - (1-\alpha)\mu_k)^2}, \quad \forall a_k \in \mathcal{A}^+.$$

¹The individual lower bounds for cheap arms, Theorem 2.2, and expensive arms, Theorem 2.3, have appeared in prior work [17] and are re-stated for completeness of presentation.

²While the results of Section 2 are for arm reward distributions that are Gaussian with unit variance, the generalization to other distribution families such as Bernoulli is straightforward [20].

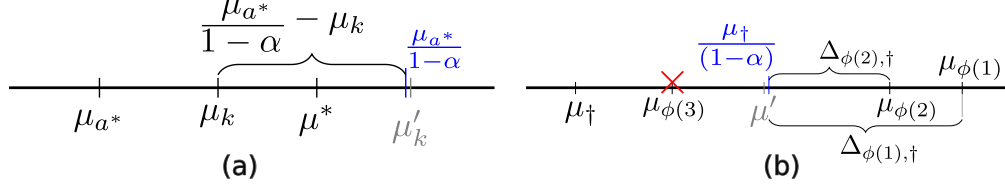


Figure 1: (a) Arm a^* must be deemed feasible by each expensive arm $a_k \in \mathcal{A}^+$ introducing dependence on $\mu_{a^*}/(1 - \alpha) - \mu_k$. (b) Each arm in $\mathcal{A}^\dagger$ is diminished to make $\mu_\dagger$ feasible. $\phi(i)$ denotes the index of the i^{th} highest reward arm; $a_{\phi(1)}$ and $a_{\phi(2)}$ are in $\mathcal{A}^\dagger$ while $a_{\phi(3)}$ is not.

Proof sketch for Theorem 2.3: The proof for Theorem 2.3 can be illustrated using Fig. 1(a). We construct a perturbed bandit instance ν' such that the expected reward μ'_k of the expensive arm $a_k \in \mathcal{A}^+$ under consideration is enhanced to be more than $\mu_{a^*}/(1 - \alpha)$, while all others are held constant. The result is that a^* is no longer optimal nor feasible in instance ν' giving rise to Theorem 2.3. Next, an analogous sketch reveals a joint lower bound on the expected samples of certain high reward arms that can **eliminate** infeasible arms through pairwise comparisons.

Definition 2.4 (Best reward cheap arm $a_\dagger$). $\mu_\dagger := \max_{a_i \in \mathcal{A}^-} \mu_i$ is the highest reward among cheap arms and is borne by arm $a_\dagger$.

Theorem 2.5 (Joint Lower Bound). *Expected number of samples of arms in $\mathcal{A}^\dagger$ satisfy,*

$$\liminf_{T \rightarrow \infty} \frac{\sum_{a_i \in \mathcal{A}^\dagger} \Delta_{i, \dagger}^2 \mathbb{E}[n_i(T)]}{\log T} \geq 2(1 - \alpha)^2.$$

Where $\Delta_{k, \dagger} := (1 - \alpha)\mu_k - \mu_\dagger$ represent the gap between the subsidized reward of an arbitrary arm a_k and $\mu_\dagger$. $\mathcal{A}^\dagger := \{a_k \in \mathcal{A} \mid \Delta_{k, \dagger} > 0\}$ represents arms with reward sufficient to eliminate $a_\dagger$.

Proof sketch for Theorem 2.5: Intuitively, $\mu_\dagger$ represents the reward that shall be most challenging to deem infeasible. The set $\mathcal{A}^\dagger$ represents arms that are capable of eliminating $a_\dagger$ by virtue of their rewards being higher than $\mu_\dagger/(1 - \alpha)$, i.e. the inflated $\mu_\dagger$. More concretely, in Fig. 1b, we construct perturbed instance ν' by reducing the reward of all the arms lying to the right of inflated $\mu_\dagger$ to below $\mu_\dagger/(1 - \alpha)$. Since arm $a_\dagger \in \mathcal{A}^-$ is cheaper than a^* , the role of optimal arm is usurped by $a_\dagger$ in ν' . Perturbing ν in this manner originates Theorem 2.5 which is a strict generalization of a bound from prior work for the case when $\mathcal{A}^\dagger$ only contains i^* [17].

Takeaways from lower bounds: Collectively the lower bounds inform us of the differences in expected rewards that must be resolved in MAB-CS and motivate the design of our algorithmic approach. Theorem 2.2 informs us that any consistent policy must check the feasibility of each cheap arm separately against μ_{CS} . Since the position of a^* in the cost-ordered line-up is not known apriori, this structure motivates an episodic and cost-ordered check on the feasibility of arms. Theorem 2.3 exposes that a minimal requirement on the samples of an expensive arm $a_k \in \mathcal{A}^+$ will come from a comparison between $(1 - \alpha)\mu_k$ and the expected reward of the optimal arm μ_{a^*} . Finally, Theorem 2.5 reveals that the disqualification of an arm is solved optimally by combining together the samples of all arms that are capable of disqualifying it.

3 Cost-Ordered Feasibility

The pseudocode for our Cost Ordered Feasibility (COF) algorithm is presented in Algorithm 1. Our algorithm accepts as input a K -armed MAB instance with collection of arms $\mathcal{A}$. COF also requires as input the problem horizon T , and an error tolerance factor δ . In practice, we tune δ using horizon T to secure the regret guarantee of Theorem 3.2. COF evaluates the feasibility of candidate arm a_ℓ by pitting it against a set of gating arms $\mathcal{G}_\ell$. We introduce candidate arms to COF in the sequence in which they are indexed, i.e. in increasing (non-decreasing) order of their costs. Starting from the cheapest arm a_1 , COF is designed to deem each arm as either feasible or infeasible with sufficient confidence. If determined feasible, we solely sample that arm thereafter. However, if infeasible we move on to the next cheapest arm.

Algorithm 1 Cost Ordered Feasibility (COF)

Inputs: K armed bandit instance $\mathcal{A} = \{a_1, \dots, a_K\}$ indexed in non-decreasing order of their costs $c_1 \leq \dots \leq c_K$; Horizon T ; Subsidy factor α ; Error tolerance δ .

Initialize: For each $a_k \in \mathcal{A}$ set $n_k = 0$, $\hat{\mu}_k = 0$, $\text{UCB}_k = 0$, $\text{LCB}_k = 0$; Candidate arm $a_\ell = a_1$.

- 1: **for** $a_k \in \mathcal{A}$: Sample a_k ; update n_k , $\hat{\mu}_k$, UCB_k , LCB_k
- 2: **while** $\sum_{a_k \in \mathcal{A}} n_k < T$ **do**
- 3: $\mathcal{G}_\ell \leftarrow \{a_i \in \{a_{\ell+1}, \dots, a_K\} \mid (1-\alpha)\text{UCB}_i \geq \text{LCB}_\ell\}$
- 4: **if** $\mathcal{G}_\ell = \emptyset$ **then**
- 5: Deem a_ℓ as optimal; sample a_ℓ thereafter
- 6: **else**
- 7: **for each** $a_k \in \mathcal{A}$ **do**
- 8: $\epsilon_{k,\ell} \leftarrow \epsilon(n_k, \hat{\mu}_k, \text{UCB}_\ell, \alpha)$ $\triangleright$ Computed as per Expression 40
- 9: **if** $\prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell} \leq \delta$ **then** $\triangleright$ Combining (aggregating) samples
- 10: Deem a_ℓ infeasible. Next evaluate $a_{\ell+1}$
- 11: **else**
- 12: **if** $n_\ell < \max_{a_i \in \mathcal{G}_\ell} n_i$ **then**
- 13: Sample a_ℓ ; update n_ℓ , $\hat{\mu}_\ell$, UCB_ℓ , LCB_ℓ $\triangleright$ Exclusive sampling
- 14: **else**
- 15: Sample a_ℓ ; update n_ℓ , $\hat{\mu}_\ell$, UCB_ℓ , LCB_ℓ $\triangleright$ BAI-filter
- 15: **and** sample $a_i \in \mathcal{G}_\ell \mid \text{UCB}_i > \max_{a_j \in \mathcal{G}_\ell} \text{LCB}_j$; update n_i , $\hat{\mu}_i$, UCB_i , $\text{LCB}_i \forall$ sampled a_i

Confidence bound scheme: Whenever an arm is sampled by COF, its sample count is incremented, and its empirical mean, UCB (Upper Confidence Bound), and LCB (Lower Confidence Bound) are updated to reflect the observed reward. For any arm a_k , $\hat{\mu}_k$ is the running sample mean of observed rewards, and the radius for the confidence bounds $\beta_k(\delta)$ is determined by the error tolerance level δ and is equal to $\sqrt{\log(1/\delta)/2n_k}$. UCB_k and LCB_k are then given by $\hat{\mu}_k + \beta_k(\delta)$ and $\hat{\mu}_k - \beta_k(\delta)$. Intuitively, $(\text{LCB}_k, \text{UCB}_k)$ represents a high probability interval for the true expected mean μ_k of arm a_k . In COF we design tests that check both the feasibility and infeasibility of candidate arm a_ℓ relative to the quality constraint $(1-\alpha)\mu^*$. The design of both COF's feasibility and infeasibility criteria use the insight that for two arbitrary arms if $\text{LCB}_k > \text{UCB}_m$, then $\mu_k > \mu_m$ with high confidence.

Determining feasibility: a_ℓ is determined to be **feasible** only once we find that its expected reward exceeds the $(1-\alpha)$ subsidized reward of every arm more expensive than itself. This condition is checked while resetting gating arms $\mathcal{G}_\ell$ (Line 3). The set $\mathcal{G}_\ell$ represents arms whose subsidized rewards have not yet been shown to be worse than μ_ℓ with sufficient confidence. We deem arm a_ℓ to be feasible once we find $\mathcal{G}_\ell$ to be empty (Line 4). Since cheaper candidates are evaluated prior to more expensive ones, the first candidate deemed feasible is treated as optimal and is sampled for the remaining time-slots (Line 5). Arms cheaper than a_ℓ , that have already been deemed infeasible in prior episodes, are excluded from $\mathcal{G}_\ell$ since their low reward makes them non-informative about μ^* .

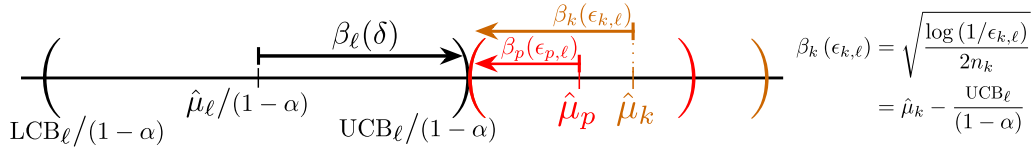


Figure 2: **Left:** The empirical mean of neither arm a_p nor arm a_k is sufficiently larger than $\text{UCB}_\ell/(1-\alpha)$ to eliminate a_ℓ . Since $\hat{\mu} - \text{UCB}_\ell/(1-\alpha) < \beta_\ell(\delta)$ for both a_p, a_k . However if $\epsilon_{p,\ell}, \epsilon_{k,\ell}$ aggregated are less than δ , then together they can eliminate a_ℓ . **Right:** $\epsilon_{k,\ell}$ is derived using a back calculation.

Determining infeasibility: Previous approaches have worked to first perform a best arm identification (BAI) round to lock in on i^* , and then proceed to make pairwise comparisons by introducing arms in cost-order [17]. We forgo a BAI round which makes determining the infeasibility criteria more statistically challenging. Declaring infeasibility requires that candidate a_ℓ lose to the subsidized

reward of any *single* arm. Moreover, as we understand from the discussion in Section 2, the infeasibility comparison is best made not by comparing candidate a_ℓ to any single arm, but rather by pooling together the confidence in the infeasibility of a_ℓ based on its comparisons with all arms. Corresponding to the comparison between a_ℓ and the subsidized reward of arm a_k we compute an upper bound on the probability that $\mu_\ell > (1 - \alpha)\mu_k$ and denote it $\epsilon_{k,\ell}$ (Line 8). Once the product of $\epsilon_{k,\ell}$ across all comparisons drops below tolerance δ , we deem a_ℓ to be infeasible (Lines 9-10). We reference this as the **combining samples** feature of COF. To compute $\epsilon_{k,\ell}$, we perform a back calculation, Fig. 2, and obtain the formula stated in Expression 40 of Appendix D.2.

By line 10 of COF (Algorithm 1), if a_ℓ cannot be conclusively deemed feasible or infeasible we continue to sample arm a_ℓ and all arms in $\mathcal{G}_\ell$ in lines 12-15. The sampling scheme of COF has a filter on $\mathcal{G}_\ell$ (line 15) built in. The filter on sampling arms in $\mathcal{G}_\ell$ pauses the sampling of arms that are with sufficient confidence not the best reward arm. This is done since the point of uniformly sampling all arms in $\mathcal{G}_\ell$ is to obtain an estimate of the feasibility threshold $\mu_{CS} = (1 - \alpha)\mu^*$. However, if we find that the UCB of a gating arm's reward has fallen below the LCB of another gating arm, we can say with confidence that the arm with low UCB is not the best reward arm. We call this feature the **BAI-filter** and it leads to bandit arms having unequal samples.

An arm whose sampling was paused by the BAI-filter may subsequently be evaluated for feasibility as a candidate feasible arm. If COF finds that the samples of candidate a_ℓ are trailing those of the most sampled arm in $\mathcal{G}_\ell$, COF will sample a_ℓ exclusively until a_ℓ catches up. We call this feature **exclusive sampling** (Line 13) and its presence reduces the overall number of sub-optimal samples required by COF. The principle of sample matching for faster resolution between rewards has roots in probability theory and often appears in the MAB context when arms with wider confidence intervals are preferentially sampled [12]. The *combining samples* and *exclusive sampling* featured by COF are deliberate design choices and their benefits on empirical performance are studied in Appendix A.

Expected regret: Under normative progression, COF deems cheap arms $a_k \in \mathcal{A}^-$ infeasible during episodes leading up to episode a^* . Then, it deems a^* feasible and samples it for the remaining time. Due to the formulation of cost and quality regret in terms of the expected number of samples in Equation 2 we can bound COF's regret by upper bounding the expected number of suboptimal samples i.e. expected samples of cheap and expensive arms. We find that for cheap arm a_ℓ the sample upper bound can be traced to episode ℓ when its feasibility is evaluated, and for an expensive arm a_k , the bound emanates from a_k 's sampling during episodes $\dagger$ and a^* .

Definition 3.1 (The quantities $\gamma_k^\dagger$ and $\gamma_k^{a^*}$). For any expensive arm $a_k \in \mathcal{A}^+$ $\gamma_k^\dagger, \gamma_k^{a^*}$ are defined as,

$$\gamma_k^\dagger := \min \left\{ \frac{(3\sqrt{A} + 1)^2 \log T}{\sum_{i=1}^A \Delta_{\phi(i),\dagger}^2}, \frac{16 \log T}{\Delta_k^2} \right\}, \quad \gamma_k^{a^*} := \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_k)^2}. \quad (3)$$

Where $\phi(k)$ represents the index of the k^{th} highest reward arm, $\Delta_{\phi(k),\dagger} = (1 - \alpha)\mu_{\phi(k)} - \mu_\dagger$ is the gap between the subsidized reward of $a_{\phi(k)}$ and the reward of $a_\dagger$, and $A \leq |\mathcal{A}^\dagger|$ are the number of top-reward arms from $\mathcal{A}^\dagger$ that participate in deeming $a_\dagger$ infeasible during episode $\dagger$.

$\gamma_k^\dagger$ represents the bounding requirement from samples of a_k during episode $\dagger$ and $\gamma_k^{a^*}$ represents the corresponding quantity during episode a^* , both under normative progression of COF. Inside the min operation for $\gamma_k^\dagger$ the first term is the number of times unfiltered arms in $\mathcal{G}_\ell$ are sampled before $a_\dagger$ is deemed infeasible, and the second term is the number of samples beyond which sampling of a_k is paused by the BAI-filter. The min reflects that the overall samples of arm a_k during episode $\dagger$ will come from the earlier of these two terminations. $\gamma_k^{a^*}$ represents the bound on samples of a_k from episode a^* . The BAI-filter does not play a role here since every arm in $\mathcal{G}_\ell$ must be eliminated by a_ℓ for a^* to be deemed feasible. We use $\gamma_k^\dagger, \gamma_k^{a^*}$ to state the cost and quality regret bounds for COF.

Theorem 3.2. For a bandit instance ν with arms $\mathcal{A}$ the expected cumulative cost and quality regret over horizon T for COF are upper bounded as,

$$\begin{aligned} \mathbb{E}[\text{Cost_Reg}(T, \nu)] &\leq \sum_{a_k \in \mathcal{A}^+} \max\{\gamma_k^\dagger, \gamma_k^{a^*}\} \Delta_{C,k}^+ + K \sum_{a_k \in \mathcal{A}^+} \Delta_{C,k}^+, \\ \mathbb{E}[\text{Quality_Reg}(T, \nu)] &\leq \sum_{a_k \in \mathcal{A}^-} \frac{16 \log T}{\Delta_{Q,k}^+} + \sum_{a_k \in \mathcal{A}^+} \max\{\gamma_k^\dagger, \gamma_k^{a^*}\} \Delta_{Q,k}^+ + K \sum_{a_k \in \mathcal{A}} \Delta_{Q,k}^+. \end{aligned}$$

Order wise perspective on regret: On studying Theorem 3.2 we find that while expensive arms $\mathcal{A}^+$ contribute to both cost and quality regret, cheap arms $\mathcal{A}^-$ only contribute to quality regret since $\Delta_{C,k}^+ = 0$ for them. The leading $O(\log T)$ term for cost regret stems from episodes $\dagger$ and a^* and the overall contribution from expensive arm is the larger of the two. Both cheap and expensive arms may contribute to quality regret, accordingly two $O(\log T)$ terms appear in its upper bound. The trailing $O(1)$ terms in both cost and quality regret stem from the unlikely outcome that COF does not follow its normative trajectory.

Comparison with lower bounds: On dividing the leading $O(\log T)$ term for the cheap arms by the incremental quality regret $\Delta_{Q,k}$ it is apparent that the dependence of the expected samples on the reward gaps is $1/\Delta_{Q,k}^2$. This is precisely the dependence in the lower bound of Theorem 2.2 making COF *order-optimal for cheap arms*. The story on the optimality for expensive arms is more complex. The sample bound for expensive arms varies loosely as the max among $1/(\mu_{a^*} - (1 - \alpha)\mu_k)^2$ and $1/\sum_{i=1}^A \Delta_{\phi(i),\dagger}^2$. The former term has a tight match with the individual lower bound for expensive arms stated in Theorem 2.3, however the story with the latter term is not as simple. The joint lower bound of Theorem 2.5 by itself allows for flexibility over the distribution of samples among arms in $\mathcal{A}^\dagger$. Imposing uniform sampling as a constraint in Theorem 2.5, we see the dependence $1/\sum_{i=1}^{|\mathcal{A}^\dagger|} \Delta_{\phi(i),\dagger}^2$ emerge. COF does not in general leverage all the arms in $\mathcal{A}^\dagger$ to eliminate $a_\dagger$. Moreover, for expensive arms the reciprocal sum of gap squared dependence is contingent on the arm being sampled to the end of episode $a_\dagger$. This need not happen as is revealed by the min in $\gamma_k^\dagger$ with the BAI-filter sample bound. Which is why for expensive arms there is a *mismatch* between the lower bound and COF’s regret upper bound. This discrepancy arises in part from there being room for improvement in the sample lower bounds, particularly from incorporating the reward gaps $\Delta_k = \mu^* - \mu_k$. We construct an example to demonstrate this in Appendix C.

Bounds compared to prior work: When it comes to regret upper bounds, COF offers an improvement over the bounds proved for prior MAB-CS algorithms. ETC-CS is an algorithm proposed by Sinha et al. [24] that operates with a fixed exploration budget. Although order-optimal on worst-case regret, due to its exploration being non-adaptive, ETC-CS ends up with $O(T^{2/3})$ instance dependent regret; order-wise worse than COF. PE-CS does have $O(\log T)$ regret, and like COF, the algorithm is order-optimal for cheap arms [17]. Due to different constant factors, a precise ordering between the COF and PE-CS bounds cannot be established. Therefore, we compare their bounds based on their dependence on reward gaps from the bandit instance. For expensive arms, PE-CS always accrues samples that vary as $1/\Delta_k^2$. Moreover, for arm i^* PE-CS has a regret upper-bound dependence on the smallest gap: $\max_{a_k \in \mathcal{A}} 1/\Delta_k^2$. We improve on PE-CS in cases when COF can deem $a_\dagger$ infeasible sooner than the BAI-filter gets to a_k . When $\gamma_k^{a^*}$ dominates, we can see that the gap $\mu_{a^*}/(1 - \alpha) - \mu_k$ is a strictly larger gap than $\Delta_k = \mu^* - \mu_k$ making the instance-dependence of COF superior.

4 Experiments

To validate COF empirically, we conduct experiments on the Goodreads [29] and MovieLens [14] datasets from the recommendation systems literature. Goodreads contains crowd-sourced book ratings over 2.3 million unique books from 870,000 users. The books are organized into eight genres and each genre has between 36,514 and 335,449 books annotated with it. The number of aggregated reviews per genre are between 150,000 and 3.5 million. For MovieLens we use the 25M variant which consists of 25 million movie ratings over 62,000 movies rated by 162,000 users. Each movie is tagged with one or more genres with a total of 20 unique genres.

We simulate a scenario where a streaming service needs to make a decision over the genre of content served to their customer while minimizing the costs incurred from serving feasible quality content, a documented real-world objective [1, 25, 36]. Each genre is modeled as a bandit arm with a Bernoulli reward distribution and known cost. The reward represents the users quality of experience and the cost represents royalties paid out by the platform. The expected reward of an arm is computed as the average of ratings given to content annotated with a certain genre. The costs for our experiments are not native to the original datasets and are sampled uniformly at random between 0 and 1. Following this method we obtain Goodreads and MovieLens bandit instances detailed in Appendix A. We simulate COF and four algorithms from the literature on these instances. These are PE-CS [17], UCB-

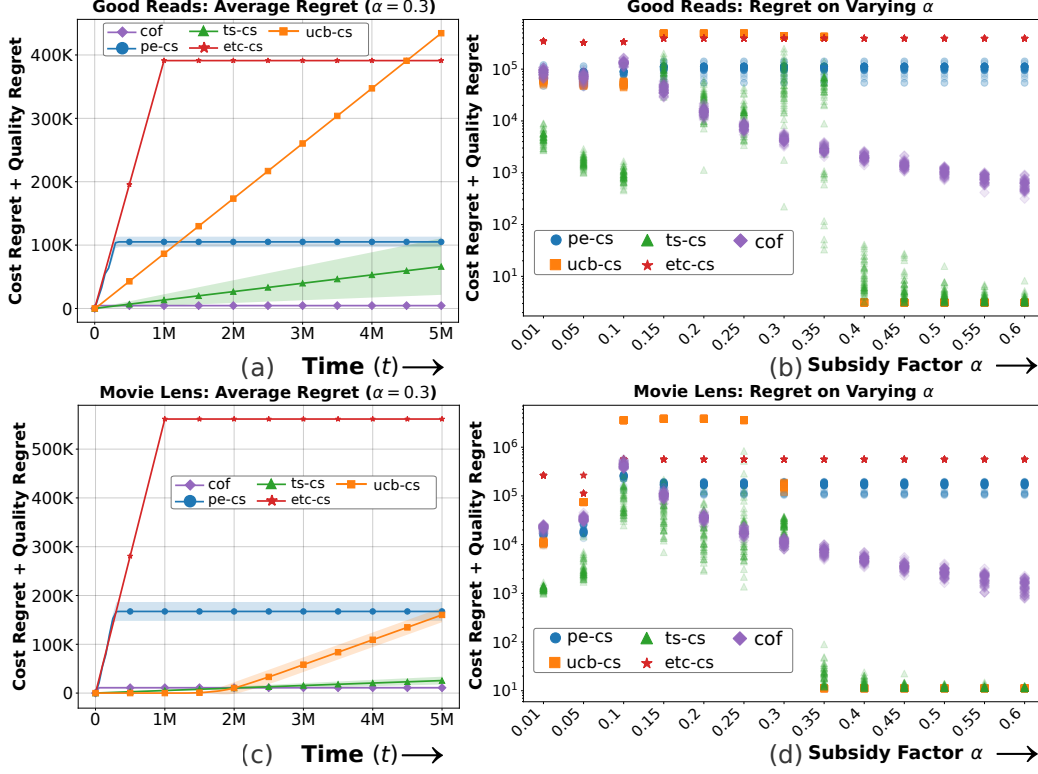


Figure 3: **Panels (a) and (c)** show the evolution of summed cost and quality regret over a horizon of 5 million samples for COF and other algorithms. Data is averaged over 50 runs and error bars represent the 20-80 percentile band. Left panels are for subsidy factor $\alpha = 0.3$. **Panels (b), (d)** contain more comprehensive results for α varying between 0.01 and 0.60. Each α column has 50 markers per algorithm corresponding to 50 independent runs **Top: Goodreads** and **Bottom: MovieLens**.

CS, TS-CS, and ETC-CS [24]. The results from the simulations on the Goodreads and MovieLens bandit instances are available in Fig. 3.

Among the algorithms competing with COF, only **ETC-CS** and **PE-CS** have theoretical regret guarantees. Sinha et al. [24] do not offer precise criteria for selecting the exploration budget for ETC-CS so we pick it to be 20% of the available horizon. Overall, from Fig. 3 we see that COF is better than both ETC-CS and PE-CS, consistent with our expectations from Section 3. ETC-CS does not adapt to a problem instance and needs a conservatively chosen exploration budget. This leads to consistently high regret on both Goodreads and MovieLens. PE-CS explores in an instance-adaptive manner, however it runs a best arm identification (BAI) routine to get the feasibility threshold μ_{CS} . Even as the feasibility criteria becomes looser with increasing subsidy factor α , the terminal regret of PE-CS remains trapped on a high plateau. COF on the other hand does not suffer from such plateauing. For the smallest values of α we find that PE-CS outperforms COF by a small margin. Small alpha values in effect make accruing a large number of samples of the best reward arm i^* necessary, so when it comes to this **small-alpha regime** PE-CS has a slight advantage over COF by virtue of tighter confidence intervals [3].

UCB-CS and **TS-CS** work by constructing a set of empirically feasible arms and then choosing to sample the cheapest arm from this set. These approaches lack guarantees on the correctness of the empirically feasible set and can in general include cheap arms in the set. While for some α values UCB-CS and TS-CS are successful at outperforming COF, their lack of reliability is evident in their outright failure on other scenarios. UCB-CS exhibits **linear regret** for several cases in both the Goodreads and MovieLens simulations. TS-CS fares better overall, however it too lacks reliability and has several outliers that drive a linear regret trend. For larger α , 0.40 and beyond, UCB-CS and TS-CS fair well despite their unprincipled nature due to all arms being feasible. This means that their empirically feasible sets cannot accidentally include a cheap arm.

The regret vs. time trends of Fig. 3(a) and (c) reveal how regret accrues over time. Optimal arm a^* is the only arm that incurs zero incremental cost and quality regret. Regardless of the experiment or the specific independent run, a principled algorithm should start with an exploration phase where it accrues regret. Then, once the arm a^* has been identified with sufficient confidence, the regret should flatten out. We see this for COF, PE-CS, and ETC-CS. However for TS-CS and UCB-CS we either find there to be a linear trend driven by all independent runs, UCB-CS on Goodreads (for $0.15 \leq \alpha \leq 0.35$), or by a subset of runs driving a less steep but still linear regret trend, TS-CS on Goodreads ($\alpha = 0.3, 0.35$) and MovieLens ($\alpha = 0.25$). On display is the fragility of UCB-CS. Once a cheap arm incorrectly makes into the empirically feasible set of UCB-CS, the algorithm will commence sampling it exclusively and not be able to recover from its mistake; Fig. 3(c) starting about $t = 2$ million. Additional experiments demonstrating the **impact of ablating** the *exclusive-sampling* and *combining samples* features of COF are documented in Section A.

5 Related Work

Multi-armed bandits have a rich literature. In the following we first discuss cost-sensitive and multi-objective bandit frameworks. Then we highlight other problems where, similar to MAB-CS, the criteria for exploration are defined relative to unknown expected rewards in the bandit instance.

Cost sensitive MABs: Broadly, costs manifest as costs from *sampling* an arm [5, 27] or as costs caused by *observing* reward [11]. The MAB-CS framework only incurs sampling costs that are deterministic and arm dependent. Most prior work from the sampling costs category [5, 7, 8, 26, 27, 34, 35] deals with a fixed budget. However, budget constrained MABs are inappropriate for real-world decision-making settings where accrued cost simply has to be minimized while honoring a constraint on reward.

Multi-objective problems: Prior work by Kanarios et al. [19] minimizes accumulated cost regardless of budget, and works with the objective of identifying the best reward arm using the least cost. However, Kanarios et al. [19] have no consideration for feasible reward or regret. Drugan and Nowe [9] introduced the multi-objective framework of MO-MAB that can incorporate the dual objectives of cost and quality regret. However, their framework works to exploit any pareto-optimal bandit arm and does not provision a feasibility criteria like MAB-CS.

Broader exploration objectives: Examples of MAB frameworks where the usual goal of exploring to identify high reward arms is fused with another constraint include not dropping below baseline rewards [10, 33], identifying all ϵ -optimal arms [2, 22], and identifying top m reward arms [6, 12, 18]. The statistical principles leveraged by these methods overlap with our work, especially probability aggregation [18]. However these methods are fundamentally different to MAB-CS since they lack the notion of cost. In our work we want to explore just enough to determine the cheapest feasible arm while being agnostic to the complete set of feasible arms. Our method prioritizes cheaper arms and only evaluates an arm’s feasibility if cheaper arms have been deemed infeasible with high confidence.

6 Conclusions and Future Work

In this paper we worked on multi-armed bandits with cost subsidy (MAB-CS) with the minimum tolerated quality set to be a fraction of the unknown best reward. First, we developed lower bounds on the expected number of samples needed from sub-optimal arms. Then, we developed the COF algorithm to solve MAB-CS, analyzed its expected cost and quality regret, and validated COF through experiments. We found a discrepancy between the instance dependence of our joint sample lower bound, and COF’s corresponding upper bound, which we shall address in future research.

COF leverages a set of gating arms to determine reward feasibility and is constrained to uniformly sample the gating arms retained by the BAI-filter. Depending on the application it may be unwise to fixate on minimizing the number of sub-optimal samples and instead cost-biased sampling may be more suitable. Our joint lower bound result reveals flexibility on gating sample origins. If the highest reward arms are too expensive, COF should lean on more samples from less expensive arms of sufficient quality. A query contextual variant of COF with cost-biased sampling would be the appropriate scheme for LLM routers due to models having an uneven cost distribution with a large spread. We motivate the need for developing a cost-biased variants of COF using a synthetic example in Appendix A and leave formal analysis and its contextual generalization to future work.

Acknowledgments and Disclosure of Funding

References

- [1] Luis Aguiar, Joel Waldfogel, and Axel Zeijen. Platform power struggle: Spotify and the major record labels. Technical report, National Bureau of Economic Research, 2024.
- [2] Aymen Al Marjani, Tomas Kocak, and Aurélien Garivier. On the complexity of all ε -best arms identification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 317–332. Springer, 2022.
- [3] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- [4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
- [5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), mar 2018. ISSN 0004-5411. doi: 10.1145/3164539. URL <https://doi.org/10.1145/3164539>.
- [6] Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR, 2013.
- [7] Semih Cayci, Atilla Eryilmaz, and R. Srikant. Budget-constrained bandits over general cost and reward distributions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4388–4398. PMLR, 2020. URL <https://proceedings.mlr.press/v108/cayci20a.html>.
- [8] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 232–238, 2013. doi: 10.1609/aaai.v27i1.8637.
- [9] Madalina M Drugan and Ann 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.
- [10] Yihan Du, Siwei Wang, and Longbo Huang. A one-size-fits-all solution to conservative bandit problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7254–7261, 2021.
- [11] Eray Can Elumar, Cem Tekin, and Osman Yağan. Multi-armed bandits with costly probes. *IEEE Transactions on Information Theory*, 71(1):618–643, 2025. doi: 10.1109/TIT.2024.3506866.
- [12] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in neural information processing systems*, 25, 2012.
- [13] Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–408, 2019.
- [14] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL <https://doi.org/10.1145/2827872>.
- [15] Hooshang Tadrissi Javan, Amir Khanlari, Omid Motamedi, and Hadi Mokhtari. A hybrid advertising media selection model using ahp and fuzzy-based ga decision making. *Neural Computing and Applications*, 29(4):1153–1167, 2018.
- [16] Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, et al. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*, 2025.

- [17] Ishank Juneja, Carlee Joe-Wong, and Osman Yağan. Pairwise elimination with instance-dependent guarantees for bandits with cost subsidy. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/forum?id=eB7T1bqthA>.
- [18] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: theory and practice. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 511–518, 2010.
- [19] Kellen Kanarios, Qining Zhang, and Lei Ying. Cost aware best arm identification. *arXiv preprint arXiv:2402.16710*, 2024.
- [20] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- [21] Benjamin Laufer, Hamidah Oderinwale, and Jon Kleinberg. Anatomy of a machine learning ecosystem: 2 million models on hugging face, 2025. URL <https://arxiv.org/abs/2508.06811>.
- [22] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ -good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33:20707–20718, 2020.
- [23] Jordan W Moffett, Judith Anne Garretson Folse, and Robert W Palmatier. A theory of multi-format communication: mechanisms, dynamics, and strategies. *Journal of the Academy of Marketing Science*, 49(3):441–461, 2021.
- [24] Abhishek Sinha, Aditya Gopalan, and Shie Mannor. Multi-armed bandits with cost subsidy. In *Proceedings of the 38th International Conference on Machine Learning*, volume 130 of *Proceedings of Machine Learning Research*, pages 9734–9744. PMLR, 2021.
- [25] Spotify for Artists. How royalties work on spotify, 2026. URL <https://artists.spotify.com/royalties-guide>. Accessed: 2026-04-27.
- [26] Long Tran-Thanh, Archie Chapman, Enrique Munoz De Cote, Alex Rogers, and Nicholas R Jennings. Epsilon-first policies for budget-limited multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1211–1216, 2010.
- [27] Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1134–1140, 2012.
- [28] Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Causal llm routing: End-to-end regret minimization from observational data. *arXiv preprint arXiv:2505.16037*, 2025.
- [29] Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94, 2018.
- [30] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1248. URL <https://aclanthology.org/P19-1248/>.
- [31] Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A Rossi, and Hoda Eldardiry. Learning to route llms from bandit feedback: One policy, many trade-offs. *arXiv preprint arXiv:2510.07429*, 2025.
- [32] Fangzhou Wu and Sandeep Silwal. Efficient training-free online routing for high-volume multi-llm serving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [33] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari. Conservative bandits. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <https://proceedings.mlr.press/v48/wu16.html>.

- [34] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 3960–3966. AAAI Press, 2015.
- [35] Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In Geoffrey Holmes and Tie-Yan Liu, editors, *Asian Conference on Machine Learning*, volume 45 of *Proceedings of Machine Learning Research*, pages 317–332. PMLR, 2016. URL <https://proceedings.mlr.press/v45/Xia15.html>.
- [36] Kevin Zielnicki, Guy Aridor, Aurelien Bibaut, Allen Tran, Winston Chou, and Nathan Kallus. The value of personalized recommendations: Evidence from netflix. Technical Report 12257, CESifo Working Paper Series, 2025.

A Additional Experiments and Details

In this section we include all the empirical results and documentation supporting the experiments for reproducibility. All variants of COF were run by setting error tolerance parameter $\delta = \frac{K^2}{T^2}$ where K is the number of bandit arms. Although this setting of δ differs from the $\delta = \frac{1}{T^2}$, the theoretical bounds for our choice of δ are identical to those for $\delta = \frac{1}{T^2}$ and only differ on constant factors.

A.1 Details on Main Paper Experiments

Statement on the use of datasets in this paper

We use the Goodreads Book Graph dataset released by Wan and McAuley and the MovieLens 25M dataset released by GroupLens. We cite the requested dataset publications for Goodreads [29, 30] and MovieLens [14]. Both datasets are used only for non-commercial academic research. We do not redistribute the raw datasets, and we comply with the stated usage terms: the Goodreads dataset is made available for academic use only and may not be redistributed or used commercially; the MovieLens 25M dataset may be used for research purposes with acknowledgment, without redistribution, without implying endorsement by the University of Minnesota or GroupLens, and without commercial or revenue-bearing use unless permission is obtained.

Instructions to recreate experiments

We have shared a folder of supplemental materials with the structure:

```
|_ data
|_ scripts
|_ source
```

- data contains bandit instance data in plain text format
- scripts contain bash files needed to generate all the log files and the plots
- source contains the python source code implementing the algorithm and plotting logic.

Scripts contains the following bash scripts.

```
|_ compress_comb_samp.sh
|_ compress_exclusive_samp.sh
|_ plot_comb_samp_bespoke.sh
|_ plot_comb_samp.sh
|_ plot_exclusive_samp_bespoke.sh
|_ plot_exclusive_samp.sh
|_ plot_only.sh
|_ plot_regret_vs_time.sh
|_ readme.md
|_ run_comb_samp.sh
|_ run_exclusive_samp.sh
|_ run_good_reads_ablation.sh
|_ run_good_reads.sh
|_ run_movie_lens_ablation.sh
|_ run_movie_lens.sh
```

- To generate Fig. 3(a) perform the following sequence of steps:
 1. Run `./scripts/run_good_reads.sh` from the project parent directory. All resource path based commands from here on out will be relative to the parent directory. This will save the log file into `results/run_logs`
 2. Run `./scripts/plot_regret_vs_time.sh` after making sure the correct destination of log files is hard coded into the bash file
- To generate Fig. 3(b) perform the following sequence of steps:
 1. Run `./scripts/run_good_reads.sh` from the project parent directory. All resource path based commands from here on out will be relative to the parent directory. This will save the log file into `results/run_logs`
 2. Run `./scripts/plot_only.sh` after making sure the correct destination of log files is hard coded into the bash file

- To generate Fig. 3(c) perform the following sequence of steps:
 1. Run `./scripts/run_movie_lens.sh` from the project parent directory. All resource path based commands from here on out will be relative to the parent directory. This will save the log file into `results/run_logs`
 2. Run `./scripts/plot_regret_vs_time.sh` after making sure the correct destination of log files is hard coded into the bash file
- To generate Fig. 3(d) perform the following sequence of steps:
 1. Run `./scripts/run_movie_lens.sh` from the project parent directory. All resource path based commands from here on out will be relative to the parent directory. This will save the log file into `results/run_logs`
 2. Run `./scripts/plot_only.sh` after making sure the correct destination of log files is hard coded into the bash file

The other bash scripts in `./scripts` are used to generate the plots available here in this supplemental experiments section.

Computational resources and wall-clock execution time

To run our bandit experiments we used a machine with the below configuration: Using this configuration, 50 independent runs

OS	Access Methods	GPU	Cores	Memory (GB)	Scratch Storage (GB)
RHEL8	ssh	NVidia Tesla T4	16	256	512

of the tested MAB-CS algorithms for a single value of α takes between 5-10 minutes of wall clock time to run depending on the choice of algorithm.

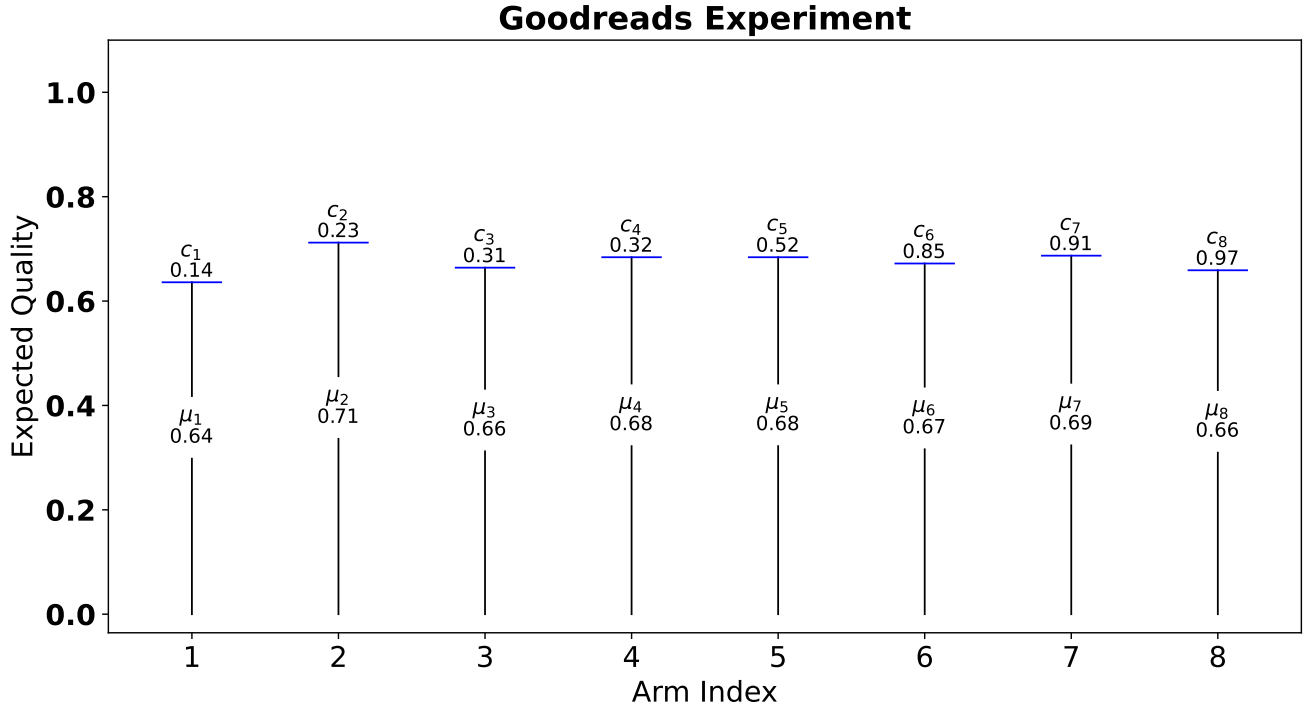


Figure 4: Good reads bandit instance

A.2 Ablation Experiments

Ablation instance ν_1

ν_1 with arm reward array $\mu = [0.15, 0.24, 0.96, 0.95, 0.99, 0.98, 0.97]$ and costs $1, 2, \dots, 7$. We use $\alpha = 0.8$ making $\mu_{CS} = 0.198$ and $a^* = 2$.

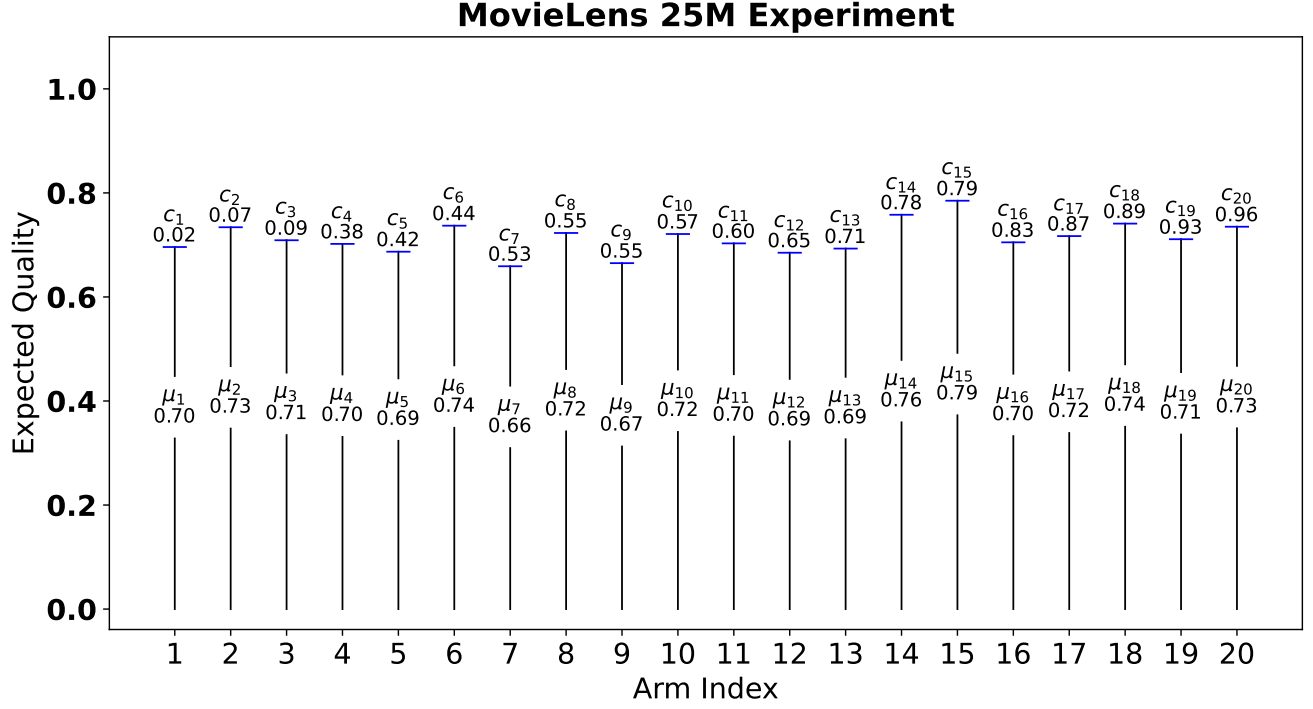


Figure 5: MovieLens bandit instance

Ablation instance ν_2

ν_2 with $\mu = [0.44, 0.46, 0.48, 0.7, 0.71, 0.704, 0.714, 0.702, 0.716, 0.708, 0.712, 0.706]$ and costs $1, 2, \dots, 12$. We use $\alpha = 0.3$ making $\mu_{CS} = 0.501$ and $a^* = 4$

Next we highlight the role of the *exclusive sampling* and *combining samples* features of COF by ablating them from COF and comparing the performance of the COF to its ablated variants on cost and quality regret. Towards this end we define two variants of COF. First COF (no exclusive) always samples both the arm a_ℓ currently under evaluation for feasibility and arms in $\mathcal{G}_\ell$ (those not excluded by the BAI-filter as discussed in Section 3). Second COF (no combine) does not aggregate the error probabilities (Line 9 of Algorithm 1) and instead checks if any gating arm individually is capable of eliminating a_ℓ . To highlight the role of exclusive sampling we design bandit instance where the second cheapest arm is optimal however is low reward leading to its sampling being paused early when it was a gating arm by the BAI-filter during the first episode that deems the cheapest arm infeasible. However a lot more of the samples of this second cheapest arm are needed to evaluate its own feasibility. We create 7-armed bandit instance ν_1 . Unablated COF has much better regret and the comparison of regret between COF and COF (no exclusive) is available in Fig. 6 panels (a) and (b) which represent two zoomed in time windows.

To illustrate the efficiency coming from combining samples (error probability aggregation) we choose a 12 armed bandit instance ν_2 . As seen from the results in Fig. 6 panels (c)-(e) unablated COF is able to deem arms a_1, a_2, a_3 with fewer samples and achieve substantially lower regret than its ablated variant that does not combine samples of expensive arms. We highlight here that while *combining samples* will always lower both the cost and quality regret of COF, this is not true for *exclusive sampling*. One example of problem where fewer exclusive sampling leads to worse regret is when fewer total samples from a pair of arms, with one of the arms being the optimal a^* comes at the cost of more samples from the arm that is sub-optimal among the two. However, we shall continue to have fewer total sub-optimal samples even in this example since without the catch-up caused by exclusive sampling, more samples of the sub-optimal arm will be needed to deem it infeasible.

A.3 Non-uniform sampling with COF

Experiment on cost-biased sampling. Bandit instance used was:

$$\mu = [0.38, 0.18, 0.50, 0.23, 0.44, 0.33, 0.28, 0.74, 0.80, 0.90, 0.96, 0.95, 0.48]$$

$$c = [1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, 10.0, 40.0, 80.0, 800.0, 850.0]$$

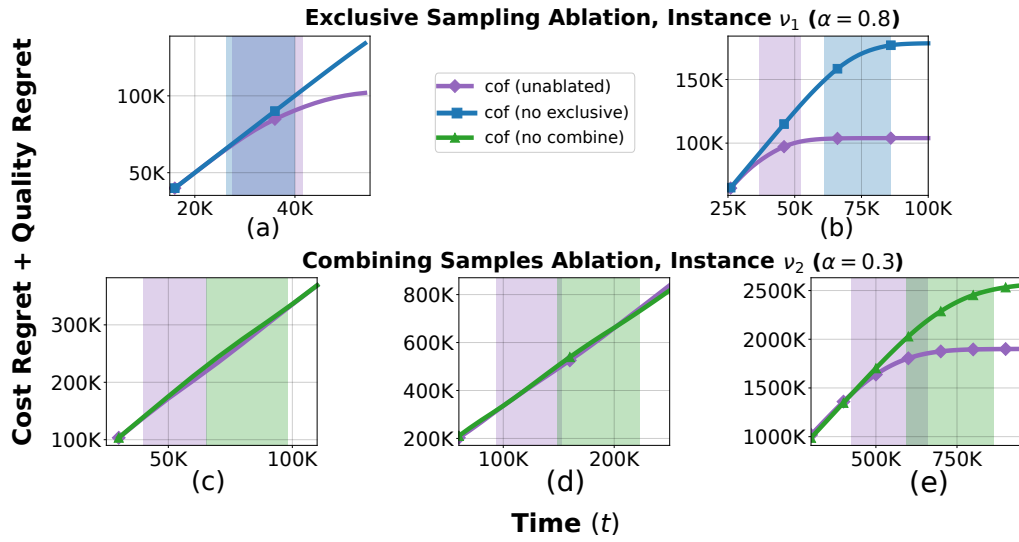


Figure 6: **Top:** Panels (a) and (b) depict windows from ablation on exclusive sampling experiment with instance ν_1 . Subsidy factor $\alpha = 0.8$, horizon $T = 200,000$ samples. The event rectangles in panels (a) and (b) represent the observed times at which a_1 was eliminated and a_2 was deemed optimal respectively. **Bottom:** Panels (c), (d), and (e) depict windows from ablation on combining samples experiment with instance ν_2 . Subsidy factor $\alpha = 0.3$, horizon $T = 1,000,000$. The event rectangles in panels (a), (b), and (c) represent the observed times at which arms a_1 , a_2 , and a_3 respectively were deemed infeasible. Data in figure is based on 1000 independent runs. All event rectangles capture one standard deviation of times at which the event occurred.

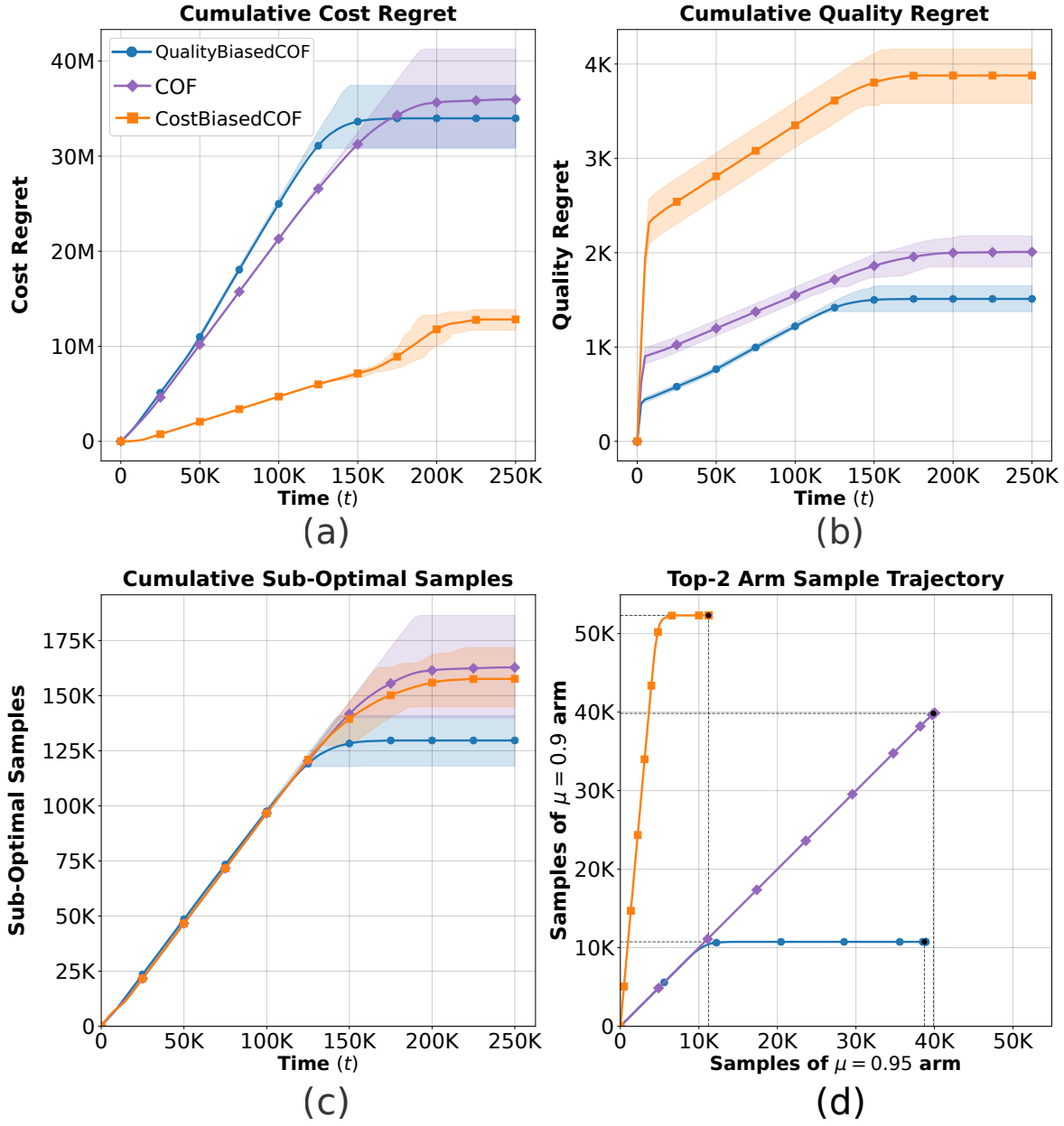


Figure 7: Variants of COF with uniform and non-uniform sampling

B Mathematical Preliminaries

We present here some mathematical preliminaries that appear repeatedly in the proofs of Theorem's stated in the main paper. These results act as building blocks for the regret lower and upper bound analyses. Instead of proving results that are standard in the MAB literature, we provide a reference to a source that prove them. First, in Section B.1 we state the tools leveraged in proving the lower bounds in Section C. Then, in Section B.2 we provide the results utilized in Section D to analyze the upper bounds on the expected number of sub-optimal arms by COF (Algorithm 1) and consequently the cost and quality regret incurred by COF. Throughout this Section, and for the analysis generally, we use the same notation as the main paper and introduce and reference new notation whenever necessary.

B.1 Preliminaries for Lower Bound Analysis

In this section we consider an MAB-CS instance with arms represented $\mathcal{A}$, and known costs of sampling each arm $a_i \in \mathcal{A}$. Just like in the main paper, arms are indexed in the non-decreasing order of their costs, and number of arms $|\mathcal{A}| = K$. An instance ν is completely specified as a collection of reward distributions $\{\nu_i\}_{a_i \in \mathcal{A}}$, and each reward distribution ν_i is assumed to be drawn from some common distribution family $\mathcal{M}$, and $\mathbb{E}[\nu_i]$ is denoted μ_i . The feasible arms $\mathcal{S} \subseteq \mathcal{A}$ are those whose expected reward $\mu_i \geq \mu_{CS}$. Since $\mu_{CS} = (1 - \alpha)\mu^*$, $\alpha \in [0, 1)$, means $\mathcal{S} := \{a_i \in \mathcal{A} \mid \mu_i \geq (1 - \alpha)\mu^*\}$. The optimal arm indexed a^* is then the least cost feasible arm $a_{a^*} = \arg \min_{a_i \in \mathcal{S}} c_i$. The MAB-CS instance dependent lower bound on expected samples of sub-optimal arms is for the class of *consistent policies*.

Definition B.1 (Consistent Policy π). A policy π is consistent if for all bandit instances ν and for all arms $i \neq a^*$, $\mathbb{E}[n_i(T)] = o(T^\gamma)$ for all $0 < \gamma \leq 1$.

Next we state the fundamental inequality based on an information theory result used frequently in the MAB literature for characterizing problem instance dependent lower bounds on the expected number of samples of sub-optimal arms.

Lemma B.2 (Fundamental inequality adapted from [13]). *Consider two MAB-CS instances ν and ν' over the same set of bandit arms $\mathcal{A}$. The reward distribution associated with sampling arm $a_i \in \mathcal{A}$ is ν_i in case of the first instance and ν'_i in case of the second. The expected number of samples $\mathbb{E}_\nu[n_i(T)]$ of an arm a_i under any policy interacting with instance ν over T sampling rounds must satisfy,*

$$\sum_{a_i \in \mathcal{A}} \mathbb{E}_\nu[n_i(T)] \text{KL}(\nu_i, \nu'_i) \geq \text{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]). \quad (4)$$

Where kl denotes the Kullback–Leibler (KL) divergence between two Bernoulli distributions, i.e.,

$$\forall p, q \in [0, 1]^2, \quad \text{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q},$$

and where Z is any random variable with support lying in $[0, 1]$. Z must be measurable with respect to the probability space that all random variables are defined on.

Lemma B.3 (Lower bound on $\text{kl}(p, q)$ from [13]). *The KL divergence between two Bernoulli distributions with parameters $p, q \in [0, 1]^2$ is lower bounded as,*

$$\text{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \leq (1 - p) \log \frac{1}{1 - q} - \log 2. \quad (5)$$

Next we define some key quantities needed for the lower bound analysis,

Definition B.4 (Key quantities $D_{\inf}, \tilde{D}_{\inf}$ from [17]). Let KL denote the Kullback–Leibler divergence between two probability distributions. Given a distribution $\nu_i \in \mathcal{M}$ and real number x , we define

$$D_{\inf}(\nu_i, x) = \inf\{\text{KL}(\nu_i, \nu'_i) \mid \nu'_i \in \mathcal{M}, \text{ and } \mathbb{E}[\nu'_i] > x\}, \quad (6)$$

$$\tilde{D}_{\inf}(\nu_i, x) = \inf\{\text{KL}(\nu_i, \nu'_i) \mid \nu'_i \in \mathcal{M}, \text{ and } \mathbb{E}[\nu'_i] \leq x\}. \quad (7)$$

Where $\mathbb{E}[\nu'_i]$ denotes the mean of distribution ν'_i . In the lower bound proofs of Section C, $D_{\inf}, \tilde{D}_{\inf}$ serve the role of quantifying the least perturbation to distributions $\{\nu_i\}_{a_i \in \mathcal{A}}$ needed to induce a change in optimal arm a^* . We use $D_{\inf}(\nu_i, x)$ when $\mu_i < x$, and $\tilde{D}_{\inf}(\nu_i, x)$ when $\mu_i > x$.

Lemma B.5 (Explicit formulas for $D_{\inf}, \tilde{D}_{\inf}$ from [20]). *When the family of reward distributions $\mathcal{M}$ is all Gaussian distributions with a common variance σ^2 , the key terms $D_{\inf}$ and $\tilde{D}_{\inf}$ take the form,*

$$D_{\inf}(\nu_i, x) = \tilde{D}_{\inf}(\nu_i, x) = \frac{(\mu_i - x)^2}{2\sigma^2}.$$

B.2 Preliminaries for Upper Bound Analysis

Definition B.6 (Subgaussian random variable). We say that X is σ -subgaussian if for any $\epsilon \geq 0$,

$$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Lemma B.7 (Bounded random variables are Subgaussian, example 5.6(c) in [20]). *If Random Variable $X \in [a, b]$ almost surely, then X is $\frac{b-a}{2}$ subgaussian.*

Lemma B.8 (Hoeffding Bound, Section 5.4 in [20]). *Let $X_1, X_2, \dots, X_n$ be n independent random variables, each bounded within the interval $[a, b] : a \leq X_i \leq b$. The empirical mean of these variables is given by,*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then Hoeffding's inequality states,

$$\begin{aligned} \mathbb{P}(\bar{X} - \mathbb{E}[X] \geq t) &\leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right), \\ \mathbb{P}(\bar{X} - \mathbb{E}[X] \leq -t) &\leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right). \end{aligned}$$

Corollary B.9 (Hoeffding Bound for Bernoulli random variables). *In the Hoeffding bound of Lemma B.8, when $\{X_i\}_{i=1}^n$ are independent and identically distributed Bernoulli random variables with $\mathbb{E}[X_i] = \mu$, then,*

$$\begin{aligned} \mathbb{P}(\hat{\mu} - \mu \geq t) &\leq \exp(-2nt^2), \\ \mathbb{P}(\hat{\mu} - \mu \leq -t) &\leq \exp(-2nt^2). \end{aligned}$$

Where $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ denotes the sample mean over n samples.

Lemma B.10 (Probability of inaccurate UCB or LCB for Bernoulli rewards). *For a bandit arm a_i with reward distributed $\text{Bern}(\mu_i)$, when the upper and lower confidence bound of the arm are defined with confidence radius $\beta_i(\delta) = \sqrt{\frac{\log(1/\delta)}{2n_i}}$ as $\text{UCB}_i = \hat{\mu}_i + \beta_i(\delta)$, and $\text{LCB}_i = \hat{\mu}_i - \beta_i(\delta)$, then we have,*

$$\begin{aligned} \mathbb{P}(\text{UCB}_i < \mu_i) &\leq \delta, \\ \mathbb{P}(\text{LCB}_i > \mu_i) &\leq \delta. \end{aligned}$$

Proof. Rearranging terms we can write,

$$\begin{aligned} \mathbb{P}(\text{UCB}_i < \mu_i) &= \mathbb{P}(\hat{\mu}_i + \beta_i(\delta) < \mu_i) \\ &= \mathbb{P}(\hat{\mu}_i - \mu_i < -\beta_i(\delta)) \\ &\leq \exp(-2n_i\beta_i^2(\delta)) = \delta. \end{aligned}$$

Where we have used the probability bound from Corollary B.9. The second part of the Lemma statement can be proved analogously. $\square$

Lemma B.11 (Iterated expectation lemma). *Let X be any integrable random variable over probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\{E_i\}_{i=1}^n$ be a collection of mutually exclusive and exhaustive measurable events. That is $\bigcup_{i=1}^n E_i = \Omega$ and $E_i \cap E_j = \emptyset, \forall i, j \in [n], i \neq j$. Then the following identity holds,*

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X | E_i] \mathbb{P}(E_i).$$

As a special case if the events are just some E and its complement E^c , then,

$$\mathbb{E}[X] = \mathbb{E}[X | E] \mathbb{P}(E) + \mathbb{E}[X | E^c] \mathbb{P}(E^c).$$

Proof. Define a sub σ -algebra of $\mathcal{F}$, $\mathcal{G} = \{\phi, E_1, E_2, \dots, E_n, \Omega\}$. Then,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] \text{ (Because } \mathcal{G} \subset \mathcal{F} \text{)} \quad (8)$$

$$= \sum_{i=1}^n \mathbb{E}[X \mid E_i] \mathbb{P}(E_i). \quad (9)$$

□

Lemma B.12 (Expectation is at most equal to larger of the conditioned expectations). *Let X be any integrable random variable over probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\{E_i\}_{i=1}^n$ be a collection of mutually exclusive and exhaustive measurable events. That is $\bigcup_{i=1}^n E_i = \Omega$ and $E_i \cap E_j = \phi$, $\forall i, j \in [n], i \neq j$. Then,*

$$\mathbb{E}[X] \leq \max_{i \in [n]} \{\mathbb{E}[X \mid E_i]\}. \quad (10)$$

Proof. Lemma B.12 can be considered a Corollary to Lemma B.11 as is illustrated by the following proof,

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X \mid E_i] \mathbb{P}(E_i) \text{ (From the proof of Lemma B.11).} \quad (11)$$

$$\leq \left(\sum_{i=1}^n \mathbb{P}(E_i) \right) \cdot \left(\max_{i \in [n]} \{\mathbb{E}[X \mid E_i]\} \right) \quad (12)$$

$$= \max_{i \in [n]} \{\mathbb{E}[X \mid E_i]\}. \quad (13)$$

□

In the analysis of the sample complexity of a sub-optimal arm $a_i \in \mathcal{A}$, $i \neq a^*$ incurred by COF (Algorithm 1) we upper bound its expected samples $\mathbb{E}[n_i(T)]$ over horizon T using Lemma B.11 and conditioning on an event that ensures the normative progression of the algorithm. We formalize this templated approach through Lemma B.13

Lemma B.13 (Bounding expected samples through conditional expectation). *For any event G that is measurable with respect to the probability space that all reward random variables are defined on,*

$$\mathbb{E}[n_i(T)] \leq \mathbb{E}[n_i(T) \mid G] + T \mathbb{P}(G^c).$$

Proof. The proof follows trivially from the iterated expectation Lemma B.11, and the fact that the total number of samples of any arm can be at most T .

$$\begin{aligned} \mathbb{E}[n_i(T)] &= \mathbb{E}[n_i(T) \mid G] \mathbb{P}(G) + \mathbb{E}[n_i(T) \mid G^c] \mathbb{P}(G^c) \quad (\text{Lemma B.11}) \\ &\leq \mathbb{E}[n_i(T) \mid G] + T \mathbb{P}(G^c) \quad (\mathbb{P}(G) \leq 1 \text{ and } n_i(T) \leq T). \end{aligned}$$

□

Lemma B.14. *For any collection of N non-negative numbers $\{x_i\}_{i=1}^N$, $x_i \geq 0$, their root mean square is larger than their arithmetic mean,*

$$\sqrt{\frac{\sum_{i=1}^N x_i^2}{N}} \geq \frac{\sum_{i=1}^N x_i}{N} \quad (14)$$

Proof. The proof follows from applying the Cauchy–Schwarz inequality to the N length vectors $\mathbf{x} = [x_1, x_2, \dots, x_N]$ and $\mathbf{1} = [1, 1, \dots, 1]$

$$\left(\sum_{i=1}^N x_i^2 \right) \cdot \left(\sum_{i=1}^N 1^2 \right) \geq \left(\sum_{i=1}^N x_i \cdot 1 \right)^2 \quad (\text{Cauchy-Schwarz inequality}) \quad (15)$$

$$\implies N \sum_{i=1}^N x_i^2 \geq \left(\sum_{i=1}^N x_i \right)^2 \quad (16)$$

$$\implies \frac{\sum_{i=1}^N x_i^2}{N} \geq \frac{1}{N^2} \left(\sum_{i=1}^N x_i \right)^2. \quad (17)$$

Since each $x_i \geq 0$, taking principal square roots on both sides establishes the stated result.

□

Lemma B.15 (Max-min Lemma). *For any function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,*

$$\max_x \min_y f(x, y) \leq \min_y \max_x f(x, y).$$

Provided the maximum and minimum are achievable.

Proof. For all $x, y \in \mathbb{R}$ by the definition of min, max,

$$\begin{aligned} \min_u f(x, u) &\leq f(x, y) \leq \max_t f(t, y) \\ \implies \min_u f(x, u) &\leq \min_s \max_t f(t, s) \\ \implies \max_v \min_u f(v, u) &\leq \min_s \max_t f(t, s). \end{aligned}$$

Where $u, v, s, t \in \mathbb{R}$ are arbitrary variables. □

Lemma B.16 (Regret Decomposition Lemma, Lemma 4.5 in [20]). *For any policy π and stochastic bandit environment ν with K arms, for horizon T , the Expected Cumulative Regret $\text{Reg}_\pi(T, \nu)$ of policy π in ν satisfies,*

$$\mathbb{E}[\text{Reg}_\pi(T, \nu)] = \sum_{i \in [K]} \Delta_i \mathbb{E}[n_i(T)].$$

This result may be trivially generalized to other notions of regret where the gap determining the incremental regret due to arm i is some arbitrary $\Delta_{X,i}$. In this case, the regret decomposition shall be,

$$\mathbb{E}[\text{Reg}_\pi^X(T, \nu)] = \sum_{i \in [K]} \Delta_{X,i} \mathbb{E}[n_i(T)].$$

In particular in our problem we have Cost and Quality regret which are,

$$\begin{aligned} \mathbb{E}[\text{Cost_Reg}(T, \nu)] &= \sum_{i \in [K]} \Delta_{C,i}^+ \mathbb{E}[n_i(T)] \\ \mathbb{E}[\text{Quality_Reg}(T, \nu)] &= \sum_{i \in [K]} \Delta_{Q,i}^+ \mathbb{E}[n_i(T)]. \end{aligned}$$

C Lower Bound Proofs

This section proves the MAB-CS lower bounds stated in Section 2 for a consistent policy (Definition B.1) on an MAB-CS instance with arms $a_i \in \mathcal{A}$ and reward distributions $\nu = \{\nu'_i\}_{a_i \in \mathcal{A}}$. For each proof we introduce a perturbed instance $\nu' = \{\nu'_i\}_{a_i \in \mathcal{A}}$.

The optimal arm is defined as usual $a_{a^*} = \arg \min_{a_i \in \mathcal{S}} c_i$ where $\mathcal{S} = \{a_i \in \mathcal{A} \mid \mu_i \geq (1 - \alpha)\mu^*\}$ is the set of feasible arms. Instance ν' shall be constructed in a manner that its optimal arm differs from the optimal arm of instance ν . In particular, in each proof we will use the fundamental inequality stated in Lemma B.2 with $Z = \frac{\mathbb{E}[n_k(T)]}{T}$ representing the sample fraction of a certain arm $a_k \in \mathcal{A}$ that is sub-optimal for ν but optimal for ν' .

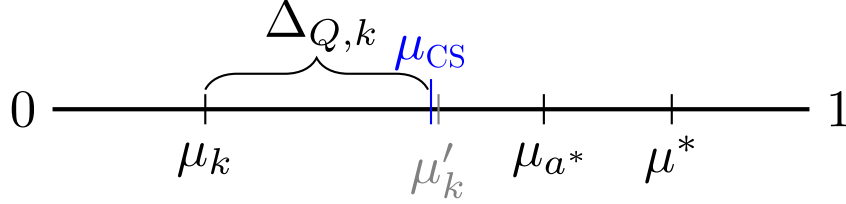


Figure 8: Cheap arms $a_k \in \mathcal{A}^-$ are infeasible as indicated by the gap $\Delta_{Q,k}$.

We outline the proof of Theorem 2.2 using the reward line illustration of Fig. 8. From the original bandit instance ν with optimal arm a^* that contains $a_k \in \mathcal{A}^-$, we construct a perturbed bandit instance ν' in which the expected reward μ'_k of a_k is enhanced to be more than the feasibility threshold μ_{CS} , and all other expected rewards are held constant. Since a_k is cheaper than a^* , the optimal arm in instance ν' shall be a_k . The lower bound of Theorem 2.2 stems from a policy solving both instances ν and ν' . No amount of reward enhancement can make an expensive arm optimal so their sample lower bound does not follow Theorem 2.2.

Theorem C.1 (Lower bound for samples of cheap arms). *For a bandit instance ν , over horizon T for all consistent policies the expected number of samples of a cheap arm are lower bounded as,*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[n_i(T)]}{\log T} \geq \frac{1}{D_{\inf}(\nu_i, (1 - \alpha)\mu^*)} \quad \forall a_i \in \mathcal{A}^-, \quad (18)$$

where $D_{\inf}$ is from Definition B.4, and $\mathcal{A}^-$ are the set of cheap arms with cost-ordered index less than a^* .

Proof. Given bandit instance ν , and cheap arm $a_i \in \mathcal{A}^-$ we construct instance ν' such that $\nu'_k = \nu_k \forall a_k \in \mathcal{A}$, $a_k \neq a_i$, and $\nu'_i \in \mathcal{M}$ is such that $\mathbb{E}[\nu'_i] = \mu_i \geq (1 - \alpha)\mu^*$. If a^* denotes the index of the optimal arm for ν , $c_i < c_{a^*}$, and arm a_i is optimal for ν' since expected reward of a_i is feasible and it is cheaper than a^* .

Theorem C.1 can be shown by applying Lemma B.2 to the instance pair ν, ν' with $[0, 1]$ bounded random variable $Z = \frac{[n_i(T)]}{T}$ denoting the sample fraction of arm a_i under any consistent policy.

$$\begin{aligned} \sum_{a_k \in \mathcal{A}} \mathbb{E}_\nu[n_k(T)] \text{KL}(\nu_k, \nu'_k) &\geq \text{kl}\left(\frac{\mathbb{E}_\nu[n_i(T)]}{T}, \frac{\mathbb{E}_{\nu'}[n_i(T)]}{T}\right) \\ \implies \mathbb{E}_\nu[n_i(T)] \text{KL}(\nu_i, \nu'_i) &\geq \left(1 - \frac{\mathbb{E}_\nu[n_i(T)]}{T}\right) \log\left(\frac{T}{T - \mathbb{E}_{\nu'}[n_i(T)]}\right) - \log 2. \end{aligned} \quad (19)$$

Where Equation 19 follows from instance ν and ν' being identical outside of arm a_i , and from the lower bound on the KL divergence between Bernoulli random variables (Lemma B.3).

For a consistent policy $\mathbb{E}_\nu[n_i(T)] = o(T^\gamma)$, and $\mathbb{E}_{\nu'}[n_i(T)] = T - o(T^\gamma)$, since total samples are T and a_i is optimal for ν' . So,

$$\liminf_{T \rightarrow \infty} \frac{1}{\log T} \left(1 - \frac{\mathbb{E}_\nu[n_i(T)]}{T}\right) \log\left(\frac{T}{T - \mathbb{E}_{\nu'}[n_i(T)]}\right) \geq 1 - \gamma, \quad \forall \gamma \in (0, 1].$$

Integrating bound in Equation 19 with the limiting bound on its right hand side,

$$\begin{aligned} \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[n_i(T)] \text{KL}(\nu_i, \nu'_i)}{\log T} &\geq 1 \\ \implies \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[n_i(T)]}{\log T} &\geq \frac{1}{\text{KL}(\nu_i, \nu'_i)} \quad \forall \nu'_i \in \mathcal{M} \text{ such that } \mathbb{E}[\nu'_i] = \mu_i \geq (1 - \alpha)\mu^*. \end{aligned}$$

Therefore the tightest limiting lower bound is the one stated in Theorem C.1. $\square$

Proof of Theorem 2.2. The proof of Theorem 2.2 follows in a straight forward manner from imposing the restriction that the distribution family $\mathcal{M}$ is all Gaussian distributions with unit variance on Theorem C.1. Under the stated restriction on $\mathcal{M}$, $D_{\text{inf}}(\nu_i, (1 - \alpha)\mu^*) = \frac{((1 - \alpha)\mu^* - \mu_i)^2}{2} = \Delta_{Q,i}^2/2 \quad \forall a_i \in \mathcal{A}^-$. Where the explicit formula for D_{inf} is from Lemma B.5. $\square$

Theorem C.2 (Lower bound for samples of expensive arms). *For a bandit instance ν , over horizon T for all consistent policies the expected number of samples of an expensive arm are lower bounded as,*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[n_i(T)]}{\log T} \geq \frac{1}{D_{\text{inf}}\left(\nu_i, \frac{\mu_{a^*}}{1 - \alpha}\right)} \quad \forall a_i \in \mathcal{A}^+, \quad (20)$$

where D_{inf} is from Definition B.4, and $\mathcal{A}^+$ are expensive arms with cost-ordered index more than a^* .

Proof. Along the lines of the proof of Theorem C.1, to prove a lower bound on the expected samples accrued of expensive arm $a_i \in \mathcal{A}^+$ we construct ν' such that $\nu'_k = \nu_k \quad \forall a_k \in \mathcal{A}$, $a_k \neq a_i$, and $\nu'_i \in \mathcal{M}$ is such that $\mathbb{E}[\nu'_i] = \mu_i > \frac{\mu_{a^*}}{1 - \alpha} \geq \mu^*$. Since arm a_i is more expensive than arm a_{a^*} , the only mechanism by which a perturbation to ν_i can upend the optimality of a_{a^*} is by raising the feasibility threshold μ_{CS} so that a_{a^*} is infeasible in instance ν' .

Let arm a_k denote the arm that is sub-optimal for ν but optimal for ν' and random variable $Z = \frac{n_k(T)}{T}$, $Z \in [0, 1]$ be the sample fraction of a_k by any policy. Using the fundamental inequality of Lemma B.2,

$$\mathbb{E}[n_i(T)] \text{KL}(\nu_i, \nu'_i) \geq \text{kl}\left(\frac{\mathbb{E}_\nu[n_k(T)]}{T}, \frac{\mathbb{E}_{\nu'}[n_k(T)]}{T}\right).$$

Similar to the proof of Theorem C.1, due to the consistency requirement, and the same policy operating on ν where a_k is sub-optimal, and on ν' where a_k is optimal, the following holds,

$$\liminf_{T \rightarrow \infty} \frac{\text{kl}\left(\frac{\mathbb{E}_\nu[n_k(T)]}{T}, \frac{\mathbb{E}_{\nu'}[n_k(T)]}{T}\right)}{\log T} \geq 1 - \gamma \quad \forall \gamma \in (0, 1].$$

Since the KL divergence $\text{KL}(\nu_i, \nu'_i)$ does not depend on T , the initial lower bound can be combined with the bound on $\text{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z])$ as,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[n_i(T)]}{\log T} \geq \frac{1}{\text{KL}(\nu_i, \nu'_i)} \quad \forall \nu'_i \in \mathcal{M} \text{ such that } \mathbb{E}[\nu'_i] = \mu'_i > \frac{\mu_{a^*}}{1 - \alpha}, \quad (21)$$

Theorem C.2 is then just the tightest lower bound on the left hand side quantity. $\square$

Proof of Theorem 2.3. Along the lines of the proof of Theorem 2.2, the proof of Theorem 2.3 follows from imposing the distribution family $\mathcal{M}$ being Gaussian with unit variance onto Theorem C.2. Under this restriction, $D_{\text{inf}}(\nu_i, \frac{\mu_{a^*}}{1 - \alpha}) = \frac{(\mu_{a^*} - (1 - \alpha)\mu_i)^2}{2(1 - \alpha)^2} \quad \forall a_i \in \mathcal{A}^+$. $\square$

Proof of Theorem 2.5. The joint lower bound is based on the insight that the optimality of a_{a^*} may be upended by a cheap arm $a_k \in \mathcal{A}^-$ not only by enhancing the reward of said cheap arm to $(1 - \alpha)\mu^*$, but also by diminishing the reward of any arm $a_i \in \mathcal{A}$ whose reward is more than $\frac{\mu_i}{1 - \alpha}$. Let $\mathcal{A}^k := \{a_i \in \mathcal{A} \mid (1 - \alpha)\mu_i > \mu_k\}$ denote the set of all arms standing in the way of cheap arm a_k being optimal. We construct ν' such that $\nu'_i = \nu_i \quad \forall a_i \notin \mathcal{A}^k$, and $\nu'_i \in \mathcal{M}$ is such that $\mathbb{E}[\nu'_i] = \mu'_i \leq \frac{\mu_k}{1 - \alpha} \quad \forall a_i \in \mathcal{A}^k$.

We start by applying the fundamental inequality of Lemma B.2 to the instance pair ν, ν' with random variable $Z = \frac{\mathbb{E}[n_k(T)]}{T}$. At the outset, we note that arm a_k is sub-optimal for instance ν , but optimal for ν' allowing us to leverage the consistency of

the policy in a manner identical to the proofs of Theorems C.1, and C.2.

$$\begin{aligned}
& \sum_{a_i \in \mathcal{A}^k} \mathbb{E}[n_k(T)] \text{KL}(\nu_i, \nu'_i) \geq \text{kl} \left(\frac{\mathbb{E}_{\nu}[n_k(T)]}{T}, \frac{\mathbb{E}_{\nu'}[n_k(T)]}{T} \right) \\
& \sum_{a_i \in \mathcal{A}^k} \mathbb{E}[n_k(T)] \text{KL}(\nu_i, \nu'_i) \\
\Rightarrow & \liminf_{T \rightarrow \infty} \frac{\sum_{a_i \in \mathcal{A}^k} \mathbb{E}[n_k(T)] \text{KL}(\nu_i, \nu'_i)}{\log T} \geq 1 \quad \forall \nu'_i \in \mathcal{M} \text{ such that } \mathbb{E}[\nu'_i] = \mu'_i \leq \frac{\mu_k}{1-\alpha}, \text{ for each } a_i \in \mathcal{A}^k \\
\Rightarrow & \liminf_{T \rightarrow \infty} \frac{\sum_{a_i \in \mathcal{A}^k} \mathbb{E}[n_k(T)] \tilde{D}_{\inf} \left(\nu_i, \frac{\mu_k}{1-\alpha} \right)}{\log T} \geq 1 \quad (\text{using Definition B.4 for } \tilde{D}_{\inf} \text{ to get the tightest lb}).
\end{aligned}$$

Once the restriction that the reward distribution family $\mathcal{M}$ is all Gaussian distributions with unit variance is imposed, we can leverage Lemma B.5 to replace $\tilde{D}_{\inf} \left(\nu_i, \frac{\mu_k}{1-\alpha} \right)$ with $\frac{((1-\alpha)\mu_i - \mu_k)^2}{2(1-\alpha)^2}$ to write,

$$\liminf_{T \rightarrow \infty} \frac{\sum_{a_i \in \mathcal{A}^k} ((1-\alpha)\mu_i - \mu_k)^2 \mathbb{E}[n_k(T)]}{\log T} \geq 2(1-\alpha)^2.$$

The result above represents a collection of $|\mathcal{A}^-| = a^* - 1$ joint lower bounds, one for each cheap arm $a_k \in \mathcal{A}^-$. In Section 2 we defined $a_{\dagger} = \arg \max_{a_i \in \mathcal{A}^-} \mu_i$ as the best reward cheap arm. This means that $\mathcal{A}^{\dagger} \subseteq \mathcal{A}^k \forall a_k \in \mathcal{A}^-$. Moreover, $((1-\alpha)\mu_i - \mu_{\dagger})^2 \leq ((1-\alpha)\mu_i - \mu_k)^2 \forall a_k \in \mathcal{A}^-$. These two facts put together imply that the joint lower bound stated in Theorem 2.5 is simply the tightest among these $|\mathcal{A}^-|$ overlapping lower bounds. $\square$

Remark C.3 (Joint lower bound is a generalization of [17]). As stated in Section 2 while we come to the same conclusion about the lower bound for cheap arms and the individual lower bound for expensive arms as [17], we highlight here how our Theorem 2.5 is a strict generalization of the lower bound on $\mathbb{E}[n_{i^*}(T)]$ shown in [17]. In [17] the lower bound based on reduction of reward for $\mathbb{E}[n_{i^*}(T)]$ is,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[n_{i^*}(T)]}{\log T} \geq \max_{i < a^*} \frac{2(1-\alpha)^2}{\Delta_{Q,i}^2} \text{ if } \mu_{\dagger} \geq (1-\alpha)\mu_{(2)}, \quad (22)$$

where $\mu_{(2)} := \max_{i \neq i^*} \mu_i$ is the second largest reward. The bound in Equation 22 is a special case of Theorem 2.5. When $\mu_{\dagger} \geq (1-\alpha)\mu_{(2)}$, only the best reward arm a_{i^*} stands in the way of cheap arm $a_{\dagger}$ being optimal and Theorem 2.5 reduces to,

$$\begin{aligned}
\liminf_{T \rightarrow \infty} \frac{((1-\alpha)\mu^* - \mu_k)^2 \mathbb{E}[n_k(T)]}{\log T} & \geq 2(1-\alpha)^2 \\
\Rightarrow \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[n_k(T)]}{\log T} & \geq \frac{2(1-\alpha)^2}{\Delta_{Q,\dagger}^2} \\
& = \max_{i < a^*} \frac{2(1-\alpha)^2}{\Delta_{Q,i}^2},
\end{aligned}$$

since $\mu_{\dagger} = \max_{a_i \in \mathcal{A}^-} \mu_i$.

Example to demonstrate room for improvement in lower bound

The following example illustrated how there is room for improvement in the lower bound on the expected number of samples of an expensive arm beyond the results proved in this paper. Using notation from the paper consider a four armed MAB-CS instance with arms $\mathcal{A} = \{a_{\dagger}, a^*, a_k, i^*\}$, and only best reward arm i^* being capable of deeming $a_{\dagger}$ infeasible. The lower bound on the expected samples required from arbitrary a_k can be tightened. Currently there is no dependence on the lower bound of $\mathbb{E}[n_k(T)]$ on the gap $\Delta_k = \mu^* - \mu_k$ even though in practice resolving Δ_k is necessary for directing sampling from a_k towards i^* to deem $a_{\dagger}$ infeasible.

D Analysis for COF

COF is carefully designed such that deviations from its intended execution have low probability. COF evaluates an arm a_ℓ as either feasible or infeasible by comparing it against all the arms more expensive than itself. First the feasibility criteria is checked while constructing the set of gating arms $\mathcal{G}_\ell$ (Line 3, Algorithm 1), then the infeasibility criteria is checked (Lines 7-10, Algorithm 1). If a_ℓ is found to be feasible then we sample it until the T slots run out. In contrast, once a_ℓ is deemed infeasible, we move on to evaluating the next cheapest $a_{\ell+1}$. If a decision about a_ℓ cannot yet be made with sufficient confidence, we further sample a_ℓ and arms in $\mathcal{G}_\ell$ (Lines 12-15, Algorithm 1).

Sampling of a_ℓ and arms $a_i \in \mathcal{G}_\ell$ is uniform barring two considerations. First, if the samples of a_ℓ lag behind those of arms in $\mathcal{G}_\ell$ then we exclusively sample a_ℓ until the disparity is amended. Second, if the UCB of an arm in $\mathcal{G}_\ell$ falls below the largest LCB among arms in $\mathcal{G}_\ell$ then its sampling is foregone in favor of arms that are better poised to gauge the feasibility of a_ℓ . We refer to the former feature as *exclusive sampling* and the latter as the *BAI-filter* on sampling. In this section we build up to the bounds on cost and quality regret for COF (Algorithm 1) that were stated in Theorem 3.2 by separately bounding the expected samples of cheap arms $\mathcal{A}^-$ and expensive arms $\mathcal{A}^+$.

COF operates in an episodic fashion where the ℓ^{th} episode is responsible for evaluating arm a_ℓ . The premise for analysis is that any arm a_i can only be sampled in episodes $1, 2, \dots, i$. In episodes $1, 2, \dots, i-1$ arm a_i may be sampled as part of $\mathcal{G}_\ell$ and in episode i as the candidate feasible arm. Analyzing $\mathbb{E}[n_i(T)]$ for a sub-optimal arm $i \neq a^*$ involves conditioning its number of samples on a carefully constructed event that secures normative sampling progression for the arm across episodes.

Definition D.1 (The event $E_{i,\ell}$). $E_{i,\ell}$ denotes the event that the terminal sample for arm i was sampled during episode i . To bound the samples of any sub-optimal arm $a_i \in \mathcal{A}$ we leverage the fact the collection of events $\{E_{i,\ell}\}$ are mutually exclusive and exhaustive.

D.1 Bound Samples of Cheap Arms

Lemma D.2 (Bound on number of samples of a cheap arm under COF). *When the error tolerance is $\delta = T^{-2}$, the expected number of samples of arm $a_i \in \mathcal{A}^-$ by COF over horizon T is upper bounded as,*

$$\mathbb{E}[n_i(T)] \leq \frac{16 \log T}{\Delta_{Q,i}^2} + 2.$$

Proof. We start by defining event G_i conditioning on which shall ensure our intended normative sampling for arm a_i .

$$G_i = \{|\hat{\mu}_i - \mu_i| < \beta_i(t, \delta) \ \forall t \in [T]\} \cap \{|\hat{\mu}_{i^*} - \mu^*| < \beta_{i^*}(t, \delta) \ \forall t \in [T]\}.$$

Or equivalently, in purely set-theoretic notation,

$$G_i = \left\{ \bigcap_{t=1}^T |\hat{\mu}_i - \mu_i| < \beta_i(t, \delta) \right\} \cap \left\{ \bigcap_{t=1}^T |\hat{\mu}_{i^*} - \mu^*| < \beta_{i^*}(t, \delta) \right\}. \quad (23)$$

Using Lemma B.13 with G_i we get,

$$\mathbb{E}[n_i(T)] \leq \mathbb{E}[n_i(T) \mid G_i] + T\mathbb{P}(G_i^c). \quad (24)$$

Henceforth we use notation $\mathbb{E}_G[n_i(T)] := \mathbb{E}[n_i(T) \mid G_i]$ for the conditional expectation. From Definition D.1 we know that the collection of events $\{E_{i,\ell}\}_{\ell=1}^i$ is mutually exclusive and exhaustive. Hence, we can apply Lemma B.12 to the conditional expectation $\mathbb{E}_G[n_i(T)]$ with the events $\{E_{i,\ell}\}_{\ell=1}^i$ to obtain,

$$\mathbb{E}_G[n_i(T)] \leq \max_{\ell=1, \dots, i} \mathbb{E}_G[n_i(T) \mid E_{i,\ell}]. \quad (25)$$

When G_i holds, due to its first clause related to the normativity of empirical mean μ_i we can upper bound UCB_i per,

$$\text{UCB}_i = \hat{\mu}_i + \beta_i(t, \delta) \leq \mu_i + 2\beta_i(t, \delta) \ \forall t \in [T]. \quad (26)$$

Similarly, due to the second clause related to the normativity of the empirical mean of the best reward arm $\hat{\mu}_{i^*}$ we can lower bound LCB_{i^*} per,

$$\text{LCB}_{i^*} = \hat{\mu}_{i^*} - \beta_{i^*}(t, \delta) \leq \mu^* - 2\beta_{i^*}(t, \delta) \ \forall t \in [T]. \quad (27)$$

Now we separately consider the consequences of Equations 26, 27 during an arbitrary episode $\ell < i$ and during episode i in bounding $\mathbb{E}_G[n_i(T) \mid E_{i,\ell}]$ and $\mathbb{E}_G[n_i(T) \mid E_{i,i}]$ respectively.

During Episode $\ell < i$

To begin with both $a_i, a_{i^*} \in \mathcal{G}_\ell$, and due to G_i the bounds on UCB_i and LCB_{i^*} stated in Equations 26 and 27 respectively hold. If the upper bound on UCB_i falls below the lower bound on LCB_{i^*} there will be no further samples of a_i due to the BAI-filter. This condition can be captured in terms of reward gap $\Delta_i = \mu^* - \mu_i$ based on the following simplification,

$$\begin{aligned} \mu_i + 2\beta_i(t, \delta) &< \mu^* - 2\beta_{i^*}(t, \delta) \\ \implies \beta_i(t, \delta) + \beta_{i^*}(t, \delta) &< \frac{\Delta_i}{2} \quad (\text{definition of gap } \Delta_i) \\ \implies \sqrt{\frac{1}{2n_i(t)} \log\left(\frac{1}{\delta}\right)} + \sqrt{\frac{1}{2n_{i^*}(t)} \log\left(\frac{1}{\delta}\right)} &< \frac{\Delta_i}{2} \quad (\text{definition of confidence radius } \beta(t, \delta)) \\ \sqrt{\frac{1}{2n(t)} \log\left(\frac{1}{\delta}\right)} &< \frac{\Delta_i}{4} \quad (\text{since } n_i(t) = n_{i^*}(t) = n(t) \text{ are matched until elimination}). \end{aligned}$$

Let $\tau_{i,\ell}$ represent both n_i, n_{i^*} beyond which there is a separation between the upper bound on UCB_i and the lower bound on LCB_{i^*} . Then,

$$\begin{aligned} \tau_{i,\ell} &= \min \left\{ n \mid \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} < \frac{\Delta_i}{4} \right\} \\ &= \frac{8 \log(1/\delta)}{\Delta_i^2} \quad (\text{rearranging terms and leveraging monotonicity in } n) \\ &= \frac{16 \log T}{\Delta_i^2} \quad (\text{since } \delta = T^{-2} \text{ in Lemma D.2}). \end{aligned}$$

We conclude that $\mathbb{E}_G[n_i(T) \mid E_{i,\ell}] \leq \tau_{i,\ell} = \frac{16 \log T}{\Delta_i^2}$.

During Episode i

During episode i , arm a_i is the candidate arm and a_{i^*} is in $\mathcal{G}_\ell$. Unlike in episodes $\ell < i$, sampling of a_i is not subject to the BAI-filter. To bound $\mathbb{E}_G[n_i(T) \mid E_{i,i}]$ we again leverage the bounds on UCB_i and LCB_{i^*} stated in Equations 26 and 27 respectively. However, instead of a lower bound LCB_{i^*} we use a lower bound on $(1 - \alpha)\text{LCB}_{i^*}$.

$$(1 - \alpha)\text{LCB}_{i^*} = (1 - \alpha)(\hat{\mu}_{i^*} - \beta_{i^*}(t, \delta)) \geq (1 - \alpha)(\mu^* - 2\beta_{i^*}(t, \delta)) \quad \forall t \in [T].$$

The key insight is that arm $a_i \in \mathcal{A}^-$ shall necessarily be deemed infeasible by COF by when $\epsilon_{i^*,i} < \delta$, making the overall product, $\prod_{a_k \in \mathcal{A}} \epsilon_{k,i} < \delta$ necessarily. Analogous to $\tau_{i,\ell}$, we find the smallest τ_i for which $\epsilon_{i^*,i} < \delta$ holds under G_i .

First we demonstrate through a rearrangement of definitions that $\text{UCB}_i < (1 - \alpha)\text{LCB}_{i^*} \implies \epsilon_{i^*,i} < \delta$.

$$\begin{aligned} \text{UCB}_i &< (1 - \alpha)\text{LCB}_{i^*} = (1 - \alpha)(\hat{\mu}_{i^*} - \beta_{i^*}(t, \delta)) \\ \implies \beta_{i^*}(t, \delta) &< \frac{\text{UCB}_i}{(1 - \alpha)} - \hat{\mu}_{i^*}. \end{aligned}$$

But we know that $\epsilon_{i^*,i}$ is such that,

$$\beta_{i^*}(t, \epsilon_{i^*,i}) = \frac{\text{UCB}_i}{(1 - \alpha)} - \hat{\mu}_{i^*}. \quad (28)$$

$$\implies \beta_{i^*}(t, \epsilon_{i^*,i}) > \beta_{i^*}(t, \delta). \quad (29)$$

$$\implies \epsilon_{i^*,i} < \delta. \quad (30)$$

Since $\epsilon_{k,i} \leq 1 \quad \forall a_k \in \mathcal{A}$, $\prod_{a_k \in \mathcal{A}} \epsilon_{k,i} < \delta$. Therefore, arm a_i is deemed infeasible by the time we accrue samples equal to those which can guarantee a separation between UCB_i and $(1 - \alpha)\text{LCB}_{i^*}$. From bounds in 26 and 28 this condition is,

$$\begin{aligned} \mu_i + 2\beta_i(t, \delta) &< (1 - \alpha)(\mu^* - 2\beta_{i^*}(t, \delta)) \\ \implies 2 + 2(1 - \alpha)\beta_i(t, \delta) &< \Delta_{Q,i}. \end{aligned}$$

The coalescing of the β terms follows from $n_i(t) \leq n_{i^*}(t) \implies \beta_i(t, \delta) \geq \beta_{i^*}(t, \delta)$ during episode a_i . Therefore, the desired bound τ_i on $\mathbb{E}_G[n_i(T) \mid E_{i,i}]$ is,

$$\begin{aligned}
\tau_i &= \min\{n_i(t) \mid 2 + 2(1 - \alpha)\beta_i(t, \delta) < \Delta_{Q,i}\} \\
&\leq \min\left\{n_i(t) \mid \beta_i(t, \delta) < \frac{\Delta_{Q,i}}{4}\right\} \quad (\text{since } \alpha \in [0, 1)) \\
&= \min\left\{n_i(t) \mid \sqrt{\frac{1}{2n_i(t)} \log\left(\frac{1}{\delta}\right)} < \frac{\Delta_{Q,i}}{4}\right\} \quad (\text{definition of } \beta_i(t, \delta)) \\
&= \frac{8 \log(1/\delta)}{\Delta_{Q,i}^2} \\
&= \frac{16 \log T}{\Delta_{Q,i}^2} \quad (\text{since } \delta = T^{-2} \text{ in Lemma D.2}).
\end{aligned}$$

For any cheap arm $a_i \in \mathcal{A}^-$, since $\mu_i < (1 - \alpha)\mu^*$, the quality gap $\Delta_{Q,i}^2 = (1 - \alpha)\mu^*$ is necessarily smaller than the reward gap $\Delta_i = \mu^* - \mu_i$. Therefore the bound in 25 can be resolved as,

$$\mathbb{E}[n_i(T) \mid G] \leq \max\{\tau_{i,\ell}, \tau_i\} = \frac{16 \log T}{\Delta_{Q,i}^2}.$$

To complete the proof of Lemma D.2 we must also bound the probability of G_i not holding $\mathbb{P}(G_i^c)$. Using De-morgan's law and the definition of G_i in Equation 23,

$$\begin{aligned}
\mathbb{P}(G_i^c) &= \mathbb{P}\left(\left\{\bigcup_{t=1}^T |\hat{\mu}_i - \mu_i| \geq \beta_i(t, \delta)\right\} \cup \left\{\bigcup_{t=1}^T |\hat{\mu}_{i^*} - \mu^*| \geq \beta_{i^*}(t, \delta)\right\}\right) \\
&\leq \sum_{t=1}^T \mathbb{P}(|\hat{\mu}_i - \mu_i| \geq \beta_i(t, \delta)) + \sum_{t=1}^T \mathbb{P}(|\hat{\mu}_{i^*} - \mu^*| \geq \beta_{i^*}(t, \delta)) \quad (\text{union bound}) \\
&\leq 2T\delta \quad (\text{using Lemma B.10}).
\end{aligned}$$

Plugging the bound on $\mathbb{P}(G_i^c)$ in Equation 24 and substituting in $\delta = T^{-2}$ we get the bound on $\mathbb{E}[n_i(T)]$ stated in Lemma D.2. $\square$

D.2 Bound Samples of Expensive Arms

The samples of an arbitrary expensive arm $a_i \in \mathcal{A}^+$ may only be accrued during episodes $\ell \in \{1, 2, \dots, a^*, \dots, i\}$. Under its normative progression, COF deems arm a^* feasible. Therefore the contribution to $n_i(T)$ from any episode beyond a^* is statistically insignificant. We establish this formally through Lemma D.3.

Lemma D.3 (Normative progression of COF ends in episode a^*). *Let $E_{i,\ell}$ be the event that the terminal sample of arm a_i is drawn during episode ℓ . Then the expected number of samples for an expensive arm $a_i \in \mathcal{A}^+$ under COF over horizon T can be upper bounded as,*

$$\mathbb{E}[n_i(T)] \leq \max \left\{ \max_{\ell < a^*} \{ \mathbb{E}[n_i(T) \mid E_{i,\ell}] \}, \mathbb{E}[n_i(T) \mid E_{i,a^*}, G_{a^*}] + 2T^2\delta \right\},$$

where G_{a^*} is the event,

$$G_{a^*} = \{ |\hat{\mu}_{a^*} - \mu_{a^*}| < \beta_{a^*}(t, \delta) \ \forall t \in [T] \} \cap \left\{ \prod_{a_k \in \mathcal{A}} \epsilon_{k,a^*}(t) > \delta \ \forall t \in [T] \right\}. \quad (31)$$

Proof. Let random variable Z denote the terminal episode in which arm a_i was sampled. Then Z can take values in $\{1, 2, \dots, a^*, \dots, i\}$ and the events $\{Z < a^*\}, \{Z \geq a^*\}$ are mutually exclusive and exhaustive. Applying Lemma B.12 to the expected samples $\mathbb{E}[n_i(T)]$ we have,

$$\mathbb{E}[n_i(T)] \leq \max \{ \mathbb{E}[n_i(T) \mid \{Z < a^*\}], \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}] \}.$$

Just like in the proof of Lemma D.2 let $E_{i,\ell}$ denote the event that the terminal sample of arm a_i was drawn during episode $\ell < a^*$. Then again using Lemma B.12,

$$\mathbb{E}[n_i(T) \mid \{Z < a^*\}] \leq \max_{\ell < a^*} \{ \mathbb{E}[n_i(T) \mid E_{i,\ell}] \}.$$

For the $Z \geq a^*$ case we introduce conditioning on the event G_{a^*} (Equation 31) that ensures the normative progression of COF during episode a^* .

$$G_{a^*} = \bigcap_{t=1}^T \{ |\hat{\mu}_{a^*} - \mu_{a^*}| < \beta_{a^*}(t, \delta) \} \cap \bigcap_{t=1}^T \left\{ \prod_{a_k \in \mathcal{A}} \epsilon_{k,a^*}(t) > \delta \right\},$$

where the latter clause is based on the infeasibility condition of COF. Now we can bound $\mathbb{E}[n_i(T) \mid \{Z \geq a^*\}]$ using Lemmas B.13 and the event G_{a^*} as,

$$\begin{aligned} \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}] &\leq \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}, G_{a^*}] + T\mathbb{P}(G_{a^*}^c \mid \{Z \geq a^*\}) \\ &= \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}, G_{a^*}] + T\mathbb{P}(G_{a^*}^c) \quad (G_{a^*} \text{ is not specified by episode}). \end{aligned} \quad (32)$$

The bound on $\mathbb{P}(G_{a^*}^c)$ follows as,

$$\begin{aligned} \mathbb{P}(G_{a^*}^c) &= \mathbb{P} \left(\bigcup_{t=1}^T \{ |\hat{\mu}_{a^*} - \mu_{a^*}| \geq \beta_{a^*}(t, \delta) \} \cup \bigcup_{t=1}^T \left\{ \prod_{a_k \in \mathcal{A}} \epsilon_{k,a^*}(t) \leq \delta \right\} \right) \quad (\text{De-morgan's rule on definition}) \\ &\leq \sum_{t=1}^T \mathbb{P}(|\hat{\mu}_{a^*} - \mu_{a^*}| \geq \beta_{a^*}(t, \delta)) + \sum_{t=1}^T \mathbb{P} \left(\prod_{a_k \in \mathcal{A}} \epsilon_{k,a^*}(t) \leq \delta \right) \quad (\text{union bound}) \\ &\leq 2T\delta \quad (\text{Lemma B.10 and by construction for } \epsilon_{k,a^*}). \end{aligned}$$

Plugging the bound on $\mathbb{P}(G_{a^*}^c)$ into 32,

$$\begin{aligned} \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}] &\leq \mathbb{E}[n_i(T) \mid \{Z \geq a^*\}, G_{a^*}] + 2T^2\delta \\ &= \mathbb{E}[n_i(T) \mid E_{i,a^*}, G_{a^*}] + 2T^2\delta. \end{aligned} \quad (33)$$

Where the final equivalence follows from the insight that the joint event $\{Z \geq a^*\}, G_{a^*}$ is equivalent to E_{i,a^*}, G_{a^*} since G_{a^*} precludes progression to any episode $Z > a^*$. Plugging in the bound 33 into 32 and combining the result with the simplification for episodes $\ell < a^*$ we get the bound stated in Lemma D.3. $\square$

In addition to Lemma D.3, we need some additional insights and definitions to analyze samples n_i during episodes $\ell < a^*$. The BAI-filter of COF (Algorithm 1, Line 15) ensures that any arm $a_i \in \mathcal{G}_\ell$ is not further sampled once we determine that the arm is not the highest reward arm i^* with sufficient confidence. As discussed in Section 3, the *BAI-filter* helps reduce regret because in the absence of the knowledge of cost gaps (costs are known but cost gaps $\Delta_{C,i}$ are unknown), the most regret efficient way of determining the infeasibility or feasibility of an arm are samples of arm i^* . Therefore once an arm is known to not be i^* , our limited sample budget is better off being allocated to other arms in $\mathcal{G}_\ell$.

In any particular episode $\ell < a^*$ the normative outcome is for a_ℓ to be deemed infeasible. Some arms $a_i, i > \ell$ will have their sampling curtailed by the BAI-filter, and others will continue to be sampled till a_ℓ is deemed infeasible. Prior to bounding the samples of an expensive arm during episode ℓ , we introduce some definitions $(\Delta_{k,\ell}, \mathcal{A}^\ell)$ and an intermediate result (Lemma D.7) to distinguish between the two possible fates of an expensive arms samples.

Definition D.4 (The gap $\Delta_{i,\ell}$ for MAB-CS). For an MAB-CS instance ν with bandit arms $a_i \in \mathcal{A}$ and subsidy factor α . If $\mathbb{E}[\nu_i] = \mu_i$, then we define $\Delta_{i,\ell} := (1 - \alpha)\mu_i - \mu_\ell$. Intuitively the $\Delta_{i,\ell}$ captures the ease with which the feasibility of arm a_ℓ relative to arm a_i may be resolved.

Definition D.5 (The set $\mathcal{A}^\ell$). The set $\mathcal{A}^\ell$ is the collection of all arms that have sufficient reward for deeming cheap arm $a_\ell \in \mathcal{A}^-$ infeasible. Mathematically,

$$\mathcal{A}^\ell := \{a_i \in \mathcal{A} \mid \Delta_{i,\ell} > 0\}, \quad |\mathcal{A}^\ell| := |\mathcal{A}^\ell|.$$

Where $\Delta_{i,\ell}$ is from Definition D.4.

Remark D.6 ($\mathcal{A}^\dagger \subseteq \mathcal{A}^\ell \forall \ell < a^*$). Let $a_\dagger := \arg \max_{i \mid a_i \in \mathcal{A}^-} \mu_i$ denote the highest reward cheap arm. All arms in $\mathcal{A}^\dagger$ will also have sufficient reward to deem a_ℓ with $\mu_\ell \leq \mu_\dagger$ infeasible and therefore will also be in $\mathcal{A}^\ell \forall \ell < a^*$.

Next we develop an intermediate result relating the sum of squared gaps to the number of sampling rounds. We shall later leverage Lemma D.7 to bound the sampling rounds needed to deem a cheap arm $a_\ell \in \mathcal{A}^-$ infeasible.

Lemma D.7 (Intermediate result relating samples and sum-squared gaps). Let δ be the error tolerance, and let $\beta := \sqrt{\frac{1}{2n} \log(\frac{1}{\delta})}$, $n \in \mathbb{N}$, $\delta \in \mathbb{R}^+$ denote the confidence radius. Then for any A -sized collection $\{\tilde{\Delta}_k\}_{k=1}^A$, $\tilde{\Delta}_k \in \mathbb{R}^+$,

$$\min \left\{ n \mid \sum_{k=1}^A (\tilde{\Delta}_k - 3\beta)^2 > \beta^2 \right\} \leq \frac{(3\sqrt{A} + 1)^2 \log(1/\delta)}{2 \tilde{\Delta}^2}.$$

Where $\tilde{\Delta}$ is the root of sum of squares of $\{\tilde{\Delta}_k\}_{k=1}^A$, $\tilde{\Delta} := \sqrt{\sum_{k=1}^A \tilde{\Delta}_k^2}$.

Proof. We shall prove the result by upper bounding the minimum number of samples under the constraint specified in Lemma D.7 by the minimum number of samples that satisfy a strictly stronger constraint.

$$\begin{aligned} \min \left\{ n \mid \sum_{k=1}^A (\tilde{\Delta}_k - 3\beta)^2 > \beta^2 \right\} &= \min \left\{ n \mid \sum_{k=1}^A \tilde{\Delta}_k^2 + 9A\beta^2 - 6\beta \sum_{k=1}^A \tilde{\Delta}_k > \beta^2 \right\} \\ &\leq \min \left\{ n \mid \tilde{\Delta}^2 + 9A\beta^2 - 6\sqrt{A}\tilde{\Delta}\beta > \beta^2 \right\} \quad (\tilde{\Delta} \text{ definition, and Lemma B.14}) \\ &= \min \left\{ n \mid (9A - 1)\beta^2 - 6\sqrt{A}\tilde{\Delta}\beta + \tilde{\Delta}^2 > 0 \right\}. \end{aligned}$$

Consider the quadratic expression $g(\beta) = (9A - 1)\beta^2 - 6\sqrt{A}\tilde{\Delta}\beta + \tilde{\Delta}^2$. For some $c_1, c_2 \in \mathbb{R}^+$ we create another related quadratic $h(\beta)$ as the factorization,

$$h(\beta) = \left((9A - 1)\beta - \frac{\tilde{\Delta}}{c_1} \right) \cdot \left(\beta - \frac{\tilde{\Delta}}{c_2} \right) = (9A - 1)\beta^2 - \tilde{\Delta} \left(\frac{9A - 1}{c_2} + \frac{1}{c_1} \right) \beta + \frac{\tilde{\Delta}^2}{c_1 c_2}.$$

If we can find c_1, c_2 such that $h(\beta) \leq g(\beta) \forall \beta$, then $h(\beta) > 0$ will imply $g(\beta) > 0$ and $\min\{n \mid g(\beta) > 0\} \leq \min\{n \mid h(\beta) > 0\}$. The solution region on β for the constraint $h(\beta) > 0$ is $\beta < \frac{\tilde{\Delta}}{(9A-1)c_1}$ or $\beta > \frac{\tilde{\Delta}}{c_2}$. The latter solution here provides a lower limit on β and for this reason is not meaningful for sample bounds in a bandit algorithm.

For $h(\beta) \leq g(\beta)$ to hold we need to pick $c_1, c_2 > 0$ such that,

$$\begin{aligned} \frac{1}{(9A-1)c_1} &\leq \frac{1}{c_2} \\ \frac{9A-1}{c_2} + \frac{1}{c_1} &\geq 6\sqrt{A}, \quad A \in \mathbb{N} \\ \frac{1}{c_1 c_2} &\leq 1. \end{aligned}$$

We can achieve this by plugging in the constraints on c_1, c_2 into a symbolic solver to achieve a valid factorization that satisfies all the constraints needed for $h(\beta) \leq g(\beta)$. Using a symbolic solver, one such factorization is,

$$\begin{aligned} c_1 &= \frac{3\sqrt{A}+1}{9A-1} \\ c_2 &= \frac{9A-1}{3\sqrt{A}+1}. \end{aligned}$$

Continuing to upper bound $\min\{n \mid g(\beta) > 0\}$ by plugging in the expression for c_1 ,

$$\begin{aligned} \min \left\{ n \mid \sum_{k=1}^A (\tilde{\Delta}_k - 3\beta)^2 > \beta^2 \right\} &\leq \min \left\{ n \mid \beta = \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} < \frac{\tilde{\Delta}}{3\sqrt{A}+1} \right\} \\ &= \frac{(3\sqrt{A}+1)^2}{2} \frac{\log(1/\delta)}{\tilde{\Delta}^2} \quad (\text{after rearranging constraint to be in } n). \end{aligned}$$

□

We now have the requisite tools needed to bound the samples of an expensive arm that end in an episode $\ell < a^*$.

Lemma D.8 (Bound on number of samples of an expensive arm in episode $\ell < a^*$). *The expected number of samples of expensive arm $a_i \in \mathcal{A}^+$, conditioned on its final sample being drawn in episode ℓ , over horizon T is upper bounded as,*

$$\mathbb{E}[n_i(T) \mid E_{i,\ell}] \leq \min \left\{ \frac{8 \log(1/\delta)}{\Delta_i^2}, \tau_\ell(\delta) \right\} + KT^2\delta.$$

Inside of the min, the first operand represents the case of the sampling of a_i being curtailed by the BAI-filter and the second case represents arm a_i being sampled to the end of episode ℓ . $\tau_\ell(\delta) := \frac{(3\sqrt{A}+1)^2}{2} \frac{\log(1/\delta)}{\sum_{i=1}^A \Delta_{\phi(i),\ell}^2}$, where $\phi(i)$ denotes the index of the i^{th} highest reward arm in $\mathcal{A}^\ell$, and A represents the number of highest-reward arms $\{a_{\phi(1)}, a_{\phi(2)}, \dots, a_{\phi(A)}\} \subseteq \mathcal{A}^\ell$ that collectively deem a_ℓ infeasible.

Proof. We define an event $G_{i,\ell}$ conditioning on which secures normative progression for sampling of arm a_i in episode ℓ .

$$G_{i,\ell} = \underbrace{\bigcap_{t=1}^T \{|\hat{\mu}_i - \mu_i| < \beta_i(t, \delta)\}}_{\text{Clause A}} \cap \underbrace{\bigcap_{t=1}^T \{|\hat{\mu}_\ell - \mu_\ell| < \beta_\ell(t, \delta)\}}_{\text{Clause B}} \cap \underbrace{\bigcap_{t=1}^T \left\{ \bigcap_{a_k \in \mathcal{A}^\ell} \{|\hat{\mu}_k - \mu_k| < \beta_k(t, \delta)\} \right\}}_{\text{Clause C}}.$$

$G_{i,\ell}$ consists of three clauses. Clause A secures the normative progression of the sampling of arm i with respect to the BAI-filter. Clauses B and C in conjunction ensure the normative outcome of arm a_ℓ being deemed infeasible. To be precise, Clause A implies,

$$\text{UCB}_i = \hat{\mu}_i + \beta_i(t, \delta) \leq \mu_i + 2\beta_i(t, \delta) \quad \forall t \in [T]. \quad (34)$$

Clause B implies,

$$\text{UCB}_\ell = \hat{\mu}_\ell + \beta_\ell(t, \delta) \leq \mu_\ell + 2\beta_\ell(t, \delta) \quad \forall t \in [T]. \quad (35)$$

And clause C implies,

$$\text{LCB}_k = \hat{\mu}_k - \beta_k(t, \delta) \geq \mu_k - 2\beta_k(t, \delta) \quad \forall a_k \in \mathcal{A}^\ell, \forall t \in [T]. \quad (36)$$

Since $i^* \in \mathcal{A}^\ell \forall \ell < a^*$, clause C includes,

$$\text{LCB}_{i^*} = \hat{\mu}_{i^*} - \beta_{i^*}(t, \delta) \geq \mu^* - 2\beta_{i^*}(t, \delta) \quad \forall t \in [T]. \quad (37)$$

Define the shorthand $\mathbb{E}_\ell[n_i(T)] := \mathbb{E}[n_i(T) \mid E_{i,\ell}]$. Then we show that conditioned on $G_{i,\ell}$, the sampling of arm a_i during episode ℓ will be curtailed either by the BAI-filter or by the arm a_ℓ being deemed infeasible. An overall upper bound on $\mathbb{E}_\ell[n_i(T) \mid G_{i,\ell}]$ thus will be from the earlier of the two possible fates. Applying Lemma B.13 to $\mathbb{E}_\ell[n_i(T)]$ with $G_{i,\ell}$,

$$\begin{aligned} \mathbb{E}_\ell[n_i(T)] &\leq \mathbb{E}_\ell[n_i(T) \mid G_{i,\ell}] + T\mathbb{P}(G_{i,\ell}^c \mid E_{i,\ell}) \\ &= \mathbb{E}_\ell[n_i(T) \mid G_{i,\ell}] + T\mathbb{P}(G_{i,\ell}^c) \quad (\text{since } G_{i,\ell} \text{ is defined } \forall t \in [T] \text{ agnostic to episode}). \end{aligned} \quad (38)$$

Conditioned on $G_{i,\ell}$, first we determine a limit τ_i^{BAI} on the samples of arm a_i due to the BAI-filter. Once $\text{UCB}_i < \text{LCB}_{i^*}$, arm i will accrue no further samples during episode ℓ . Therefore,

$$\begin{aligned} \tau_i^{\text{BAI}} &= \min\{n_i \mid \text{UCB}_i < \text{LCB}_{i^*}\} \\ &\leq \min\{n_i \mid \mu_i + 2\beta_i(t, \delta) < \mu^* - 2\beta_{i^*}(t, \delta)\} \\ &\leq \min\{n_i \mid 4\beta_i(t, \delta) < \mu^* - \mu_i\} \quad (\text{since } n_i(t) \leq n_{i^*}(t) \implies \beta_i(t, \delta) > \beta_{i^*}(t, \delta)) \\ &= \min\left\{n_i \mid \sqrt{\frac{1}{2n_i} \log(1/\delta)} < \frac{\Delta_i}{4}\right\} \quad (\text{using definition of } \beta_i(t, \delta)) \\ &= \frac{8 \log(1/\delta)}{\Delta_i^2} \quad (\text{rearranging the constraint to be in } n_i). \end{aligned} \quad (39)$$

Next we determine a limit on the number of samples of a_i due to episode ℓ terminating. From Algorithm 1, arm a_ℓ is deemed infeasible once $\prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell}(t) < \delta$ (infeasibility criteria). We bound the number of samples needed for the infeasibility criteria to hold by upper bounding the product $\prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell}$, and determining the sample count that is sufficient for the upper bound to be less than δ .

The formula for $\epsilon_{k,\ell}(t)$ ³ is given by,

$$\epsilon_{k,\ell} = \begin{cases} \exp\left(-2n_k \left(\hat{\mu}_k - \frac{\text{UCB}_\ell}{(1-\alpha)}\right)^2\right), & \text{if } \hat{\mu}_k > \frac{\text{UCB}_\ell}{(1-\alpha)}. \\ 1, & \text{otherwise.} \end{cases} \quad (40)$$

Let $\hat{\Delta}_{k,\ell} := (1-\alpha)\hat{\mu}_k - \text{UCB}_\ell$, and $\hat{\Delta}_{k,\ell}^+ = \max\{0, \hat{\Delta}_{k,\ell}\}$. Then,

$$\begin{aligned} \prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell} &= \exp\left(\frac{-2}{(1-\alpha)} \sum_{a_k \in \mathcal{A}} n_k (\hat{\Delta}_{k,\ell}^+)^2\right) \\ &\leq \exp\left(-2 \sum_{a_k \in \mathcal{A}^\ell} n_k (\hat{\Delta}_{k,\ell}^+)^2\right). \end{aligned}$$

Where we drop $(1-\alpha)$, and take sum over $\mathcal{A}^\ell$ instead since $\epsilon_{k,\ell} \leq 1$. Arms in $\mathcal{A}^\ell \subset \mathcal{A}$ (Definition D.5) have the feature of having $(1-\alpha)\mu_k > \mu_\ell$. Next⁴, to upper bound $\prod \epsilon_{k,\ell}$, we lower bound each $\hat{\Delta}_{k,\ell}^+$ by leveraging clauses B and C of $G_{i,\ell}$.

$$\begin{aligned} (1-\alpha)\hat{\mu}_k &\geq (1-\alpha)(\mu_k - \beta_k) \quad \forall a_k \in \mathcal{A}^\ell, \\ \text{UCB}_\ell &\leq \mu_\ell + 2\beta_\ell \\ \implies \hat{\Delta}_{k,\ell} &= (1-\alpha)\hat{\mu}_k - \text{UCB}_\ell \\ &\geq (1-\alpha)\mu_k - (1-\alpha)\beta_k - \mu_\ell - 2\beta_\ell \\ &\geq \Delta_{k,\ell} - \beta_k - 2\beta_\ell \quad (\text{using Definition D.5 for } \Delta_{k,\ell}) \\ \implies (\hat{\Delta}_{k,\ell})^+ &\geq \max\{\Delta_{k,\ell} - \beta_k - 2\beta_\ell, 0\} = (\Delta_{k,\ell} - \beta_k - 2\beta_\ell)^+. \end{aligned}$$

³Going forward we drop explicit time parameterization (t) for ϵ , $\hat{\mu}$, UCB , and LCB unless necessary

⁴We start denoting $\beta_k(t, \delta) = \sqrt{\frac{1}{2n} \log(\frac{1}{\delta})}$ by just β_k for any arbitrary arm $a_k \in \mathcal{A}$ unless the parameters are needed

Therefore, we can continue to bound $\prod \epsilon_{k,\ell}$ as,

$$\prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell} \leq \exp \left(-2 \sum_{a_k \in \mathcal{A}^\ell} n_k \left((\Delta_{k,\ell} - \beta_k - 2\beta_\ell)^+ \right)^2 \right).$$

For any arm $a_k \in \mathcal{A}^\ell$, conditioned on the event $G_{i,\ell}$, we have,

$$\text{UCB}_k = \hat{\mu}_k + \beta_k \leq \mu_k + 2\beta_k \quad \forall t \in [T]. \quad (41)$$

Combining 37 and 41 we can see that due to the BAI-filter, a_k can have at most $\frac{8 \log(1/\delta)}{\Delta_k^2}$ samples (analogous to the derivation of $\tau_{i,\ell}^{\text{BAI}}$). Let n represent the running count of the sampling rounds completed during episode ℓ . In each sampling round the unfiltered arms of $\mathcal{G}_\ell$ are sampled. Then for a_ℓ we will have $n_\ell \geq n$. Similar to a_ℓ , for each $a_k \in \mathcal{A}^\ell$ that was not filtered $n_k \geq n$. Further, let $\beta = \sqrt{\frac{1}{2n} \log(\frac{1}{\delta})}$ be the confidence radius corresponding to n and error tolerance δ . Then $\beta \geq \max\{\beta_\ell, \beta_k\}$, and,

$$(\Delta_{k,\ell} - \beta_k - 2\beta_\ell)^+ = \max(\Delta_{k,\ell} - \beta_k - 2\beta_\ell, 0) \geq \max(\Delta_{k,\ell} - 3\beta, 0) = (\Delta_{k,\ell} - 3\beta)^+.$$

We can use this fact to further upper bound $\prod \epsilon_{k,\ell}$.

$$\prod_{a_k \in \mathcal{A}} \epsilon_{k,\ell} \leq \exp \left(-2n \sum_{a_k \in \mathcal{A}^\ell} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,\ell} - 3\beta)^+ \right)^2 \right).$$

To get our desired bound we wish to find the smallest integer τ_ℓ such that (conditioned on $G_{i,\ell}$), a_ℓ is deemed infeasible $\forall n \geq \tau_\ell$.

$$\begin{aligned} \tau_\ell(\delta) &= \min \left\{ n \mid \exp \left(-2n \sum_{a_k \in \mathcal{A}^\ell} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,\ell} - 3\beta)^+ \right)^2 \right) \leq \delta \right\} \\ &= \min \left\{ n \mid \sum_{a_k \in \mathcal{A}^\ell} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,\ell} - 3\beta)^+ \right)^2 \geq \beta^2 \right\} \quad (\text{using definition of } \beta \text{ and rearranging}). \end{aligned} \quad (42)$$

We define $\mathcal{A}^\ell(p)$ as the collection of the top p reward arms in $\mathcal{A}^\ell$. Reintroducing the reward ordered indexing scheme $\phi(i)$ first introduced in the statement of Lemma D.8,

$$\mathcal{A}^\ell(p) := \{a_{\phi(1)}, a_{\phi(2)}, \dots, a_{\phi(p)}\}.$$

Where as in the statement of Lemma D.8, $a_{\phi(i)}$ is the i^{th} highest reward arm in $\mathcal{A}^\ell$. To solve the optimization problem of 42, we define a candidate solution $\tau_{\ell,p}$ for τ_ℓ as,

$$\tau_{\ell,p} := \min \left\{ n \mid \sum_{a_k \in \mathcal{A}^\ell(p)} (\Delta_{k,\ell} - 3\beta)^2 \geq \beta^2 \right\}.$$

We say that candidate solution $\tau_{\ell,p}$ is feasible when,

$$\sum_{a_k \in \mathcal{A}^\ell(p)} (\Delta_{k,\ell} - 3\beta(\tau_{\ell,p}, \delta))^2 = \sum_{a_k \in \mathcal{A}^\ell} \mathbb{I} \left\{ \tau_{\ell,p} \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \cdot \left((\Delta_{k,\ell} - 3\beta(\tau_{\ell,p}, \delta))^+ \right)^2.$$

We can get the desired τ_ℓ as the minimum over the feasible solutions,

$$\tau_\ell(\delta) = \min \{ \tau_{\ell,p}(\delta) \mid p = 1, \dots, |\mathcal{A}^\ell|, \text{ and } \tau_{\ell,p} \text{ is feasible} \}.$$

Using Lemma D.7 we can then say that each candidate $\tau_{\ell,p}$ is given by,

$$\tau_{\ell,p}(\delta) = \frac{(3\sqrt{p} + 1)^2}{2} \frac{\log(1/\delta)}{\sum_{i=1}^p \Delta_{\phi(i),\ell}^2}.$$

In general the best feasible $\tau_{\ell,p}$ will depend on the bandit instance. If we define A such that $\tau_{\ell,A}$ is the smallest feasible solution, then,

$$\tau_{\ell}(\delta) = \frac{(3\sqrt{A} + 1)^2}{2} \frac{\log(1/\delta)}{\sum_{i=1}^A \Delta_{\phi(i),\ell}^2}.$$

Conditioned on $G_{i,\ell}$, overall the expected number of samples are bound as,

$$\begin{aligned} \mathbb{E}_{\ell}[n_i(T) \mid G_{i,\ell}] &\leq \min\{\tau_i^{\text{BAI}}, \tau_{\ell}\} \\ &\leq \min\left\{\frac{8\log(1/\delta)}{\Delta_i^2}, \frac{(3\sqrt{A} + 1)^2}{2} \frac{\log(1/\delta)}{\sum_{i=1}^A \Delta_{\phi(i),\ell}^2}\right\}. \end{aligned}$$

To complete the bound on $\mathbb{E}_{\ell}[n_i(T)]$ from Equation 38 we must also bound $\mathbb{P}(G_{i,\ell}^c)$. We start by combining the definition of $G_{i,\ell}$ with De-morgan's rule.

$$\begin{aligned} \mathbb{P}(G_{i,\ell}^c) &= \mathbb{P}\left(\bigcup_{t=1}^T \{|\hat{\mu}_i - \mu_i| \geq \beta_i(t, \delta)\} \cup \bigcup_{t=1}^T \{|\hat{\mu}_{\ell} - \mu_{\ell}| \geq \beta_{\ell}(t, \delta)\} \cup \bigcup_{t=1}^T \left\{ \bigcap_{a_k \in \mathcal{A}^{\ell}} \{|\hat{\mu}_k - \mu_k| \geq \beta_k(t, \delta)\} \right\}\right) \\ &\leq \sum_{t=1}^T \mathbb{P}(|\hat{\mu}_i - \mu_i| \geq \beta_i(t, \delta)) + \sum_{t=1}^T \mathbb{P}(|\hat{\mu}_{\ell} - \mu_{\ell}| \geq \beta_{\ell}(t, \delta)) + \sum_{t=1}^T \sum_{a_k \in \mathcal{A}^{\ell}} \mathbb{P}(|\hat{\mu}_k - \mu_k| \geq \beta_k(t, \delta)) \\ &\leq KT\delta. \end{aligned}$$

Plugging back in this bound on $\mathbb{P}(G_{i,\ell}^c)$ and the earlier bound on $\mathbb{E}_{\ell}[n_i(T) \mid G_{i,\ell}]$ into Equation 38 we get the bound stated in Lemma D.8. $\square$

Remark D.9 (Role of gating filter). Line 3 of Algorithm 1 resets the gating arms at the start of every iteration of COF and in effect acts like a gating filter. This remark addresses why the gating filter does not play a role in the upper bound on the expected number of samples of an expensive arm during episodes $\ell < a^*$.

For episodes $\ell < a^*$, $\mu_{\ell} < (1 - \alpha)\mu^*$. If for some arm a_i , $(1 - \alpha)\mu_i > \mu_{\ell}$ then it is a legitimate member of the gating set $\mathcal{G}_{\ell}$ and not at risk from being removed from sampling by the gating filter. In the other case $\mu_{\ell} > (1 - \alpha)\mu_i$ and we can study the gap $\mu_{\ell} - (1 - \alpha)\mu_i$ to understand the filtering impact.

$$\begin{aligned} \mu_{\ell} - (1 - \alpha)\mu_i &< (1 - \alpha)\mu^* - \mu_i \\ &< \mu^* - \mu_i = \Delta_i \end{aligned}$$

Therefore from the perspective of sample complexity analysis of an arbitrary expensive arm a_i , BAI-filter will always kick in before gating set filter.

However in the case of expected samples of a_i during episode a^* gating filter does play a role as shown in Lemma D.10.

Lemma D.10 (Bound on number of samples of an expensive arm in episode a^*). *Under COF the expected number of samples of expensive arm $a_i \in \mathcal{A}^+$, conditioned on its final sample being drawn in episode a^* , and on the event G_{a^*} (defined in Equation 31), over horizon T is upper bounded as,*

$$\mathbb{E}[n_i(T) \mid E_{i,a^*}, G_{a^*}] \leq \frac{8\log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} + T^2\delta.$$

Proof. Corresponding to episode a^* for the samples $n_i(T)$ we wish to bound, $\mathbb{E}[n_i(T) \mid E_{i,a^*}, G_{a^*}]$. In addition to the normativity secured by conditioning on G_{a^*} , we must also condition on an additional event to secure the normativity around arm i 's elimination by arm a^* . Hence we define G_{i,a^*} as,

$$\begin{aligned} G_{i,a^*} &= \{|\hat{\mu}_i - \mu_i| < \beta_i(t, \delta) \mid \forall t \in [T]\} \\ &= \bigcap_{t=1}^T \{|\hat{\mu}_i - \mu_i| < \beta_i(t, \delta)\}. \end{aligned}$$

We bound $\mathbb{E}[n_i(T) \mid E_{i,a^*}, G_{a^*}]$ by further conditioning on G_{i,a^*} using Lemma B.13,

$$\begin{aligned}\mathbb{E}_{a^*}[n_i(T) \mid G_{a^*}] &\leq \mathbb{E}_{a^*}[n_i(T) \mid G_{a^*}, G_{i,a^*}] + T \cdot \mathbb{P}(G_{i,a^*}^c \mid E_{i,a^*}, G_{a^*}) \\ &= \mathbb{E}_{a^*}[n_i(T) \mid G_{a^*}, G_{i,a^*}] + T \cdot \mathbb{P}(G_{i,a^*}^c) \quad (\text{arms are independent, } G_{i,a^*} \text{ is agnostic of episode}) \\ &\leq \mathbb{E}_{a^*}[n_i(T) \mid G_{a^*}, G_{i,a^*}] + T^2 \delta \quad (\text{using union bound and Lemma B.10}).\end{aligned}\tag{43}$$

Where we have introduced the $\mathbb{E}_{a^*}[\cdot] := \mathbb{E}[\cdot \mid E_{i,a^*}]$ operator for brevity. Next we show that conditioned on G_{i,a^*}, G_{a^*} , the number of samples of a_i during episode a^* can be at most a certain bandit instance dependent quantity τ_{i,a^*} .

When G_{i,a^*} and G_{a^*} hold, we can write,

$$\text{LCB}_{a^*} = \hat{\mu}_{a^*} - \beta_{a^*}(t, \delta) \geq \mu_{a^*} - 2\beta_{a^*}(t, \delta)\tag{44}$$

$$\text{UCB}_i = \hat{\mu}_i + \beta_i(t, \delta) \leq \mu_i + 2\beta_i(t, \delta).\tag{45}$$

Since the lower bound on LCB_{a^*} defined in Equation 44 and the upper bound on UCB_i defined in equation 45 are strictly increasing and strictly decreasing functions of time t (or equivalently number of samples $n_{a^*}(t)$ and $n_i(t)$ respectively), once a strict separation is achieved between the lower bound from Equation 44 and the upper bound from Equation 45, arm a_i will be eliminated from the set of gating arms for episode a^* and will not be sampled again during that episode.

Hence conditioned on G_{a^*}, G_{i,a^*} we can find the maximum possible samples by plugging in the filtering condition on $\mathcal{G}_{a^*}$ implemented by COF (Line 3, Algorithm 1).

$$(1 - \alpha)\text{UCB}_i \leq (1 - \alpha)(\mu_i + 2\beta_i(t, \delta)) < \mu_{a^*} - 2\beta_{a^*}(t, \delta) \leq \text{LCB}_{a^*}.$$

Rearranging terms we have,

$$2(1 - \alpha)\beta_i(t, \delta) + 2\beta_{a^*}(t, \delta) \leq \mu_{a^*} - (1 - \alpha)\mu_i.$$

Since during episode a^* , $n_{a^*}(t) \geq n_i(t)$, $\beta_i(t, \delta) > \beta_{a^*}(t, \delta)$, the inequality will be satisfied when,

$$2(2 - \alpha)\beta_i(t, \delta) < \mu_{a^*} - (1 - \alpha)\mu_i.$$

Hence the required value of $\beta_i(t, \delta)$ is governed by,

$$\beta_i(t, \delta) < \frac{\mu_{a^*} - (1 - \alpha)\mu_i}{2(2 - \alpha)}.$$

Hence, the maximum number of samples of n_i is governed by,

$$\sqrt{\frac{1}{2n_i} \log\left(\frac{1}{\delta}\right)} < \frac{\mu_{a^*} - (1 - \alpha)\mu_i}{2(2 - \alpha)}.$$

We can choose the least value that will satisfy the above condition by setting equality,

$$\begin{aligned}\tau_{i,a^*} &= \frac{2(2 - \alpha)^2 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \\ &\leq \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2}.\end{aligned}$$

Hence we can further bound the RHS of Equation 43 as,

$$\mathbb{E}_{a^*}[n_i(T) \mid G_{a^*}] \leq \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} + T^2 \delta.\tag{46}$$

□

We are now close to ready to bound the overall number of samples of an expensive arm. Before we can, we need a final intermediate result to obtain an ordering between the number of sampling rounds required to disqualify the best reward cheap arm $a_{\dagger}$ and some other arbitrary cheap arm a_ℓ , $\ell \neq \dagger$.

Lemma D.11 (Disqualifying $a_{\dagger}$ requires the most samples). *Let $a_{\dagger}$ denote the highest reward cheap arm. In the notation of Lemma D.8 let $\tau_\ell(\delta)$ denote the least number of samples from each participating arm (sampling rounds) needed to deem a_ℓ infeasible with an error tolerance δ . Then,*

$$\tau_{\dagger}(\delta) = \max_{\ell < a^*} \tau_\ell(\delta).$$

Proof. By definition $\tau_{\dagger}$ satisfies Equation 42,

$$\tau_{\dagger}(\delta) = \min \left\{ n \mid \sum_{a_k \in \mathcal{A}^{\dagger}} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,j} - 3\beta)^+ \right)^2 \geq \beta^2 \right\}.$$

To prove Lemma D.11, we rely on two observations,

1. As remarked in D.6, $\mathcal{A}^{\dagger} \subseteq \mathcal{A}^{\ell}$. That is, any arm included in $\mathcal{A}^{\dagger}$ is also a part of other $\mathcal{A}^{\ell}$ for any cheap a_{ℓ} .
2. Since $(\Delta_{k,\ell} - 3\beta)^+ = \max\{\Delta_{k,\ell} - 3\beta, 0\}$, and $\Delta_{k,\ell} = (1 - \alpha)\mu_k - \mu_{\ell}$, we have $\Delta_{k,\dagger} \leq \Delta_{k,\ell} \forall \ell < a^*$.

Starting with the general feasibility condition for τ_{ℓ} and applying our two observations,

$$\sum_{a_k \in \mathcal{A}^{\ell}} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,\ell} - 3\beta)^+ \right)^2 \geq \sum_{a_k \in \mathcal{A}^{\dagger}} \mathbb{I} \left\{ n \leq \frac{8 \log(1/\delta)}{\Delta_k^2} \right\} \left((\Delta_{k,\dagger} - 3\beta)^+ \right)^2.$$

Which in turn means that $\tau_{\ell}(\delta)$ as defined in Equation 42 is not more than $\tau_{\dagger}(\delta)$. $\square$

Lemma D.12 (Bound on number of samples of an expensive arm). *When the error tolerance $\delta = T^{-2}$, the expected number of samples of expensive arm $a_i \in \mathcal{A}^+$ under COF over horizon T is upper bounded as,*

$$\mathbb{E}[n_i(T)] \leq \max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + K.$$

Where $a_{\dagger} = \arg \max_{a_k \in \mathcal{A}^-} \mu_k$ is the best reward cheap arm, and $\tau_{\dagger}$ is as defined in Lemma D.8.

Proof. Plugging in the bounds from Lemmas D.8 (for episodes $\ell < a^*$) and D.10 (for episode a^*) into the max over episodes bound from Lemma D.3 we get,

$$\begin{aligned} \mathbb{E}[n_i(T)] &\leq \max \left\{ \max_{\ell < a^*} \left\{ \min \left\{ \frac{8 \log(1/\delta)}{\Delta_i^2}, \tau_{\ell}(\delta) \right\} + KT^2\delta \right\}, \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} + 3T^2\delta \right\} \\ &\leq \max \left\{ \max_{\ell < a^*} \left\{ \min \left\{ \frac{8 \log(1/\delta)}{\Delta_i^2}, \tau_{\ell}(\delta) \right\} \right\}, \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + KT^2\delta \quad (\text{pulling out larger error term}) \\ &\leq \max \left\{ \min \left\{ \max_{\ell < a^*} \{ \tau_{\ell}(\delta) \}, \frac{8 \log(1/\delta)}{\Delta_i^2} \right\}, \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + KT^2\delta \quad (\text{using Lemma B.15}). \\ &= \max \left\{ \min \left\{ \tau_{\dagger}(\delta), \frac{8 \log(1/\delta)}{\Delta_i^2} \right\}, \frac{8 \log(1/\delta)}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + KT^2\delta \quad (\text{using Lemma D.11}). \end{aligned}$$

Plugging in the error tolerance $\delta = T^{-2}$, we obtain the expression stated in Lemma D.12. $\square$

D.3 Bound Cost and Quality Regret

We put together the bound on the samples of cheap and expensive arms together to prove the upper bound on cost and quality regret stated in Theorem 3.2.

Proof of Theorem 3.2. Using the regret decomposition (Lemma B.16) we can express and bound the expected cumulative cost regret, and the expected cumulative quality regret as,

$$\begin{aligned} \mathbb{E}[\text{Cost_Reg}(T, \nu)] &= \sum_{a_i \in \mathcal{A}^+} \Delta_{C,i}^+ \mathbb{E}[n_i(T)], \quad (\Delta_{C,i}^+ = 0 \text{ for cheap arms}) \\ \mathbb{E}[\text{Quality_Reg}(T, \nu)] &= \sum_{a_i \in \mathcal{A}^-} \Delta_{Q,i}^+ \mathbb{E}[n_i(T)] + \sum_{a_i \in \mathcal{A}^+} \Delta_{Q,i}^+ \mathbb{E}[n_i(T)], \end{aligned}$$

respectively. To bound $\mathbb{E}[n_i(T)]$ for cheap arms ($a_i \in \mathcal{A}^-$) we invoke Lemma D.2, and for expensive arms ($a_i \in \mathcal{A}^+$) we use Lemma D.12.

$$\begin{aligned}\mathbb{E}[\text{Cost_Reg}(T, \nu)] &\leq \sum_{a_i \in \mathcal{A}^+} \Delta_{C,i}^+ \left(\max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + K \right) \\ &= \sum_{a_i \in \mathcal{A}^+} \Delta_{C,i}^+ \left(\max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} \right) + K \sum_{a_i \in \mathcal{A}^+} \Delta_{C,i}^+.\end{aligned}$$

Similarly, for quality regret,

$$\begin{aligned}\mathbb{E}[\text{Quality_Reg}(T, \nu)] &\leq \sum_{a_i \in \mathcal{A}^-} \Delta_{Q,i}^+ \left(\frac{16 \log T}{\Delta_{Q,i}^2} + 2 \right) \\ &\quad + \sum_{a_i \in \mathcal{A}^+} \Delta_{Q,i}^+ \left(\max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} + K \right) \\ &= \sum_{a_i \in \mathcal{A}^-} \frac{16 \log T}{\Delta_{Q,i}^+} + \sum_{a_i \in \mathcal{A}^+} \Delta_{Q,i}^+ \left(\max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} \right) \\ &\quad + 2 \sum_{a_i \in \mathcal{A}^-} \Delta_{Q,i}^+ + K \sum_{a_i \in \mathcal{A}^+} \Delta_{Q,i}^+ \\ &\leq \sum_{a_i \in \mathcal{A}^-} \frac{16 \log T}{\Delta_{Q,i}^+} + \sum_{a_i \in \mathcal{A}^+} \Delta_{Q,i}^+ \left(\max \left\{ \min \left\{ \tau_{\dagger}(T^{-2}), \frac{16 \log T}{\Delta_i^2} \right\}, \frac{16 \log T}{(\mu_{a^*} - (1 - \alpha)\mu_i)^2} \right\} \right) \\ &\quad + K \sum_{a_i \in \mathcal{A}} \Delta_{Q,i}^+.\end{aligned}$$

To obtain the bound stated in Theorem 3.2 we substitute $\tau_{\dagger}(T^{-2}) = \frac{(3\sqrt{A} + 1)^2 \log T}{\sum_{i=1}^A \Delta_{\phi(i), \dagger}^2}$. □