# PAIRWISE ELIMINATION WITH INSTANCE-DEPENDENT GUARANTEES FOR BANDITS WITH COST SUBSIDY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-armed bandits (MAB) are commonly used in sequential online decision-making when the reward of each decision is an unknown random variable. In practice, however, the typical goal of maximizing total reward may be less important than minimizing the total cost of the decisions taken, subject to a reward constraint. For example, we may seek to make decisions that have at least the reward of a reference "default" decision. This problem was recently introduced in the Multi-Armed Bandits with Cost Subsidy (MAB-CS) framework. MAB-CS is broadly applicable to problem domains where a primary metric (cost) is constrained by a secondary metric (reward), and there is an inability to explicitly determine the trade-off between these metrics. In our work, we first introduce the Pairwise-Elimination algorithm for a simplified variant of the cost subsidy problem with a known reference arm. We then generalize PE to PE-CS to solve the MAB-CS problem in the setting where the reference arm is the un-identified optimal arm. Next, we analyze the performance of both PE and PE-CS on the dual metrics of Cost and Quality Regret. Our instance-dependent analysis of PE and PE-CS reveals that both algorithms have an order-wise logarithmic upper bound on Cost and Quality Regret, making our policy the first with such a guarantee. Finally, experiments are conducted using the MovieLens 25M dataset for both PE and PE-CS and using a synthetic toy experiment for PE-CS revealing that our method invariably outperforms the ETC-CS baseline from the literature.

## 1 INTRODUCTION

Online sequential decision-making problems capture many applications where decisions must be made without knowing their outcomes in advance. After each decision, the resulting outcome or reward is observed, and an internal model is updated to improve future decisions. In clinical trials for example, the goal is to compare the therapeutic value of various drugs against an ailment. The decisions in this case represent administering a certain drug and the rewards are the apriori unknown efficacies of the candidate drugs. Communication networks are another example. Here decisions must be made about the communication channel to be employed. In this scenario the reward represents the success or lack thereof of communicating over a chosen channel. Multi-Armed Bandits (MABs) (Lattimore & Szepesvári, 2020) are a framework for *stateless* sequential decision making where the available decisions are abstracted as arms of a MAB problem instance. The stateless assumption implies that the distribution of rewards associated with an arm is not affected by the choices of past arms. The setting within MABs we work with is that of *stationary stochastic bandits* where the distribution of arm rewards does not evolve with time.

The generality of the assumptions imposed by the stationary stochastic bandits setting provides a wide net to capture a range of problem domains. However, in real-world applications there are often several competing objectives that go beyond the limited goal of maximizing reward. For instance consider the problem faced by a marketing agency where there are several communication modalities available to communicate the agency's advertising message. Blindly maximizing the overall success rate (reward) in this case would be naive. Since such can approach would ignore the drastically different costs of using these modalities. Our example reveals that costs being associated with the sampling of any particular arm is a structure that appears quite naturally in applications.

In the marketing agency problem we know that the various communication modalities shall have unknown success rates for any brand new ad campaign. However, the cost of employing any modality will typically be known. These known arm sampling costs might manifest themselves in the form of a prescribed cost budget (Badanidiyuru et al., 2018), or as a metric whose cumulative value is to be minimized (Sinha et al., 2021), to specify two among many possible cost structures. We work with the latter among these settings. In particular our paper works with variants of the MAB with Cost-Subsidy (MAB-CS) framework introduced recently in Sinha et al. (2021).

### 1.1 Multi-Armed Bandits with Cost Subsidy (MAB-CS) Setting

What makes the MAB-CS setting so interesting is that it requires the bandit policy minimize cumulative costs while obtaining cumulative reward that is *satisfactory* and not necessarily maximal. The dual objectives of minimizing costs while maintaining satisfactory reward are captured by the cumulative cost and quality regret objectives defined in Equations 1 and 2 which we seek to minimize. To complete these definitions we define the best action $a^*$ in Equation 3 and re-formulate the cumulative regret expressions in terms of cost gaps $\Delta_{C,i} = (c_i - c_{a^*})^+$ and quality gaps $\Delta_{Q,i} = (\mu_{\text{CS}} - \mu_i)$ using a standard regret decomposition result.

$$\text{Cost\_Reg}\left(T, \nu\right) = \sum_{t=1}^{T} \mathbb{E}_{\pi}\left[\left(c_{k_t} - c_{a^*}\right)^+\right] = \sum_{i=1}^{K} \Delta_{C,i} \mathbb{E}\left[n_i\left(T\right)\right] \tag{1}$$

$$\text{Quality\_Reg}\left(T, \nu\right) = \sum_{t=1}^{T} \mathbb{E}_{\pi}\left[\left(\mu_{\text{CS}} - \mu_{k_t}\right)^+\right] = \sum_{i=1}^{K} \Delta^+_{Q,i} \mathbb{E}\left[n_i\left(T\right)\right] \tag{2}$$

$$a^* = \arg\min_{i \in S} c_i, \text{ where, } S = \left\{i : \mu_i \geq \mu_{\text{CS}}\right\}. \tag{3}$$

The regret definitions are cumulative over problem horizon $T$. Just like a conventional stationary stochastic bandit instance, specifying a $K$-armed MAB-CS instance $\nu$ entails providing expected returns $\mu_i, i \in [K]$ corresponding to the stationary arm reward distributions. Additionally for an MAB-CS instance we require arm-costs $c_i$ associated with each arm $i$, and a reward threshold $\mu_{\text{CS}}$. The arm $a^*$ is the least cost arm that satisfies the constraint implicit in the specification of reward threshold $\mu_{\text{CS}}$, and is known as the *best action*. Lastly, $\pi$ is the bandit policy, and the expectation in the definitions is over the choice of the arm $k_t$ made by policy $\pi$ during time slot $t$.

In conventional bandits, it is typical to define sub-optimality gaps as the difference in reward between the largest expected return $\mu^*$ of optimal arm $i^* = \arg\max_i \mu_i$ and sub-optimal arm $i$ as $\Delta_i = \mu^* - \mu_i$. $x^+$ denotes $\max\{0, x\}$, i.e. the zero-clipped $x$. Zero-clipping the per round incremental regret in Equations 1 and 2 serves a critical purpose for the applications we work with. The zero-clipping ensures that a stellar performance in one time-slot cannot compensate for a poor performance in another time-slot. In other words, unlike some prior works, in our setting, bad decisions in some time-slots cannot make up for stellar decisions in other time-slots. In advertising, for example, a good advertising decision for one product cannot make up for a poor advertising of another product.

To conclude the presentation of the MAB-CS setting we motivate its variants using our running example of a marketing agency. Consider that there are three modalities available for the agency to deliver their message. These methods are: (1) very expensive personalized door-step solicitation, (2) moderately expensive automated phone call, and (3) inexpensive email. Given these modalities, the agency's goal may be to achieve a prescribed sales rate with the minimum possible cost. Or, the goal may be that sales be at least a prescribed fraction of the sales of a certain communication modality. Finally, we may not have a reference mode in mind, and we may just desire a conversation rate that is (say) 80% as much as the modality with the highest sales rate that is unknown. The first and second settings are captured by our novel contributions of the fixed threshold and known reference arm settings of MAB-CS. The third setting was introduced in prior work and we refer to in our paper as the full cost subsidy setting.

### 1.2 Key Contributions

Our first contribution is to extend the MAB-CS framework to include two new settings. (1) The known threshold $\mu_0$ MAB-CS setting with $\mu_{\text{CS}} = \mu_0$. (2) When $\mu_{\text{CS}} = (1 - \alpha)\mu_\ell$ is the subidized but unknown return of a known reference arm $\ell$. We then present an original regret minimization

algorithm, called Pairwise Elimination (PE) for the latter setting. Under PE non-reference arms are pit against the reference arm in the ascending order of their costs. PE uses a principled elimination based regret minimization algorithm called Improved-UCB (Auer & Ortner, 2010) to determine whether an arm provides satisfactory rewards. Moreover PE intelligently re-uses samples for downstream comparisons. Further we present an asymmetric variant of PE that leverages accrued up samples of the reference arm $\ell$ to require fewer samples from arms further downstream and show an improved empirical performance using this variant.

We show that our PE has an instance dependent upper bound on both expected cost and quality regret that is $\mathcal{O}(\log T)$, and that PE only samples arms more expensive then the best action at most a constant number of times under expectation. Next, we develop a generalization of PE for the full cost subsidy setting called PE-CS. We show that PE-CS too admits an $\mathcal{O}(\log T)$ instance dependent upper bound on both cost and quality regret that involves both notions of conventional sub-optimality gaps and quality gaps.

Not only is PE-CS the first algorithm for the full cost subsidy setting with instance dependent upper bounds on cost and quality regret, is also offers an improvement in the guarantee over the only other algorithm that has one, namely ETC-CS. While ETC-CS admits an $\mathcal{O}\left(T^{2/3}\right)$ upper bound, PE-CS admits an $\mathcal{O}(\log T)$ one on summed cost and quality regret. Further, through experiments based on data both real and synthetic, we demonstrate that PE-CS offers the best balance between performance and reliability when compared with baselines.

## 2 Related Work

**Structured Bandits:** There have been numerous works that impose additional structure onto the stationary stochastic bandits problem with the goal of better addressing specific application domains. This structure can sometimes come in the form of relationships imposed on the rewards of arms. These reward-relationships may be known (Kleinberg et al., 2008) or unknown (Gupta et al., 2021). Adding constraints that depend on the risks associated with sampling the rewards of an arm as in Wu et al. (2016) or Chen et al. (2022) is another form of the structured problem.

**Bandits with Costs:** We contextualize the core contributions of this paper by comparing and contrasting our setting and methods with related ones from the Literature. We build on the MAB-CS setting introduced in Sinha et al. (2021). A core component of the MAB-CS setting is that there is a known cost associated with sampling any arm that is specified as part of the problem instance. There have been numerous works within the MAB literature that include the notion that a price has to be paid for sampling an arm. Notably the Bandits with Knapsacks (Badanidiyuru et al., 2018) line of work also considers a setting with known costs. However, in Badanidiyuru et al. (2018) there is a limited cost-budget and reward must be optimized while satisfying strict budget constraints. In MAB-CS and its variants on the other hand, the goal is to minimize cumulative costs without there being any explicit constraints on cost. In MAB-CS, the constraints are in fact on reward, and are referred to as quality constraints.

**Bandit with Constraints:** The quality constraints in our work closely resemble the constraints on expected rewards that are imposed in the work on Conservative Bandits (Wu et al., 2016). In both our work and in Wu et al. (2016), there is a constraint that requires the accumulated reward to exceed a $(1-\alpha)$ discounted version of the reward of a reference arm. The conservative bandits setting only considers the cases where either the return of the reference arm is a known constant $\mu_0$, or the case where the reference arm is known, but its return is unknown. The primary difference in our work is that in addition to satisfying a quality constraint, in the MAB-CS setting, we must work to minimize the cumulative cost. The notion of costs are completely absent in Wu et al. (2016), moreover, in addition to the cases with a known reward threshold $\mu_0$, and a known reference arm with an unknown return, which we consider as novel extensions to the MAB-CS framework, we also address the problem of the original MAB-CS framework where the reference arm is the unknown optimal arm. Another key difference is that Wu et al. (2016) imposes the reward constraints in a cumulative anytime manner, whereas we impose it at every time-step.

**BAI and Improved UCB** In our paper, we work with the notions of Cost Regret and Quality Regret which are identical to the ones introduced by Sinha et al. (2021), however unlike the setting in Sinha et al. (2021) which only considers the case where the reference reward comes from the so-

far unidentified optimal arm, we consider the additional cases (1) Where there is a fixed known threshold to be exceeded (which we call the known threshold setting), and (2) When there is a known reference arm $\ell$ whose reward $\mu_\ell$ has to be exceeded however $\mu_\ell$ itself is unknown. In addition to our novel PE-CS algorithm for the setting from Sinha et al. (2021) we present novel algorithms for our new settings (1) and (2) as well. In Sinha et al. (2021) the authors present three novel algorithms for the MAB-CS setting, the former two among which construct a set of empirically satisfactory arms by interleaving exploration and exploitation. We build up our approach to optimizing for the regret objectives by first solving the known reference arm $\ell$ with unknown reward $\mu_\ell$ setting using a successive elimination style algorithm that compares candidate arms one at a time against the reference arm to see if they are satisfactory. We call this approach Pairwise Elimination (PE), and we adapt the elimination based regret minimization algorithm Improved-UCB (Auer & Ortner, 2010) to develop it. Then we generalize PE to the case where the reference arm is the unknown optimal arm by prepending PE with a Best-Arm-Identification (BAI) stage (also based on Auer & Ortner (2010)). We call this latter algorithm PE-CS.

## 3 Algorithms and Analysis

As discussed in Section 1, we introduce novel settings called: (1) Fixed threshold MAB-CS with $\mu_{\text{CS}} = \mu_0$ and (2) known reference $\ell$ MAB-CS with $\mu_{\text{CS}} = (1-\alpha)\mu_\ell$. In interest of building up to our presentation on PE-CS we start with the known reference $\ell$ setting and relegate the known threshold setting to Appendix D. For the known reference $\ell$ setting, we present our novel Pairwise-Elimination algorithm in Section 3.1. Our PE algorithm builds upon Improved-UCB (Figure 1 in Auer & Ortner (2010)), a regret minimization algorithm for the stationary stochastic bandits setting. We choose to build on Improved-UCB since its successive elimination approach to regret minimization intuitively adapts to our insight that arms be evaluated in the order of their costs. By adapting Improved-UCB, we compare cheaper arms to arm $\ell$. We eliminate these cheaper arms if they are unsatisfactory or we declare them the best action $a^*$ if they are able to eliminate arm $\ell$.

One of the core features of Improved-UCB is that the cadence of sampling and elimination is governed by **rounds**. We inherit the use of these rounds and associated formulas from Improved-UCB. Moreover, we use the elimination checks prescribed in the Improved-UCB algorithm but adapt them to better leverage the advanced round number and number of samples available for the reference arm as we describe in more detail in Section 3.1.

Finally, our third setting has the reward threshold $\mu_{\text{CS}} = (1-\alpha)\mu^*$. This full cost-subsidy setting is strictly more challenging than the known reference arm $\ell$ setting since the optimal arm itself is unknown. To solve the Full Cost-Subsidy setting, we extend the PE algorithm by pre-pending it with a Best-Arm-Identification (BAI) stage to develop the PE-CS algorithm. The details of the Full Cost-Subsidy setting and PE-CS are presented in Section 3.2.

### 3.1 Pairwise-Elimination for Known Reference Arm Setting

Under the known reference arm setting, the quality regret is calibrated against the expected reward of the reference arm $\ell$ which we denote $\mu_\ell$. For jointly optimizing cost and quality regret in the known reference arm $\ell$ setting, we take an approach where the suitability of an arm with respect to the quality constraint implicit in the quality regret definition (Equation 2) is evaluated in the order of the costs of the arms: cheapest first. This insight motivates a pairwise-comparison between the non-reference candidate arms and the known reference arm $\ell$, where the candidate arms are considered in the ascending order of their known costs. As alluded to earlier we adapt Improved-UCB to facilitate this pairwise comparison. Improved UCB operates in two phases: An **exploration phase** where a set of active arms is progressively refined, and an **exploitation phase** where the last surviving arm is sampled for the remainder of the budget. We adapt this algorithm to our problem setting by (1) Assigning a separate **episode** to each candidate arm, where the candidate arms are being assessed in the ascending order of their costs. (2) Initializing the set of active arms at the start of each episode with only the candidate arm associated with that episode and the reference arm. This initialization of the active set facilitates the pairwise comparison. (3) In our Pairwise Elimination (PE) algorithm, the exploitation phase is only entered once the reference arm is bested by the candidate arm in an episode. This episode is said to be the final episode, and under the nominal operation of the algorithm, this shall be episode $a^*$ that evaluates the candidacy of the arm $a^*$ that is the best action.

Finally, we remark that since any arm that is more expensive than the reference arm $\ell$ is necessarily sub-optimal, these arms are pruned away from the bandit instance and are never sampled as part of this problem formulation.

**Function 1:** Pairwise Elimination Function PE()

**Function** *PE( $\hat{\boldsymbol{\mu}}$: Empirical Means, $\ell$: Reference Arm, $\boldsymbol{n}$: Sample Vector, $T$: Horizon, $\boldsymbol{\omega}$: Round Numbers, $i$: Episode, $\alpha$: Subsidy Factor )*:

$\tilde{\Delta} \leftarrow 2^{-\omega_i}$
$\tau \leftarrow \left\lceil \frac{2\log(T\tilde{\Delta}^2)}{\tilde{\Delta}^2} \right\rceil$
**for** $k \in \{i, \ell\}$ **do**
  **if** $n_k < \tau$ **then**
    **return** $k, \boldsymbol{\omega}, i$ `// k_t = k, round numbers ω and episode i unchanged`
$\beta \leftarrow \sqrt{\frac{\log(T\tilde{\Delta}^2)}{2\tau}}$
**if** $(1-\alpha)(\hat{\mu}_\ell + \beta) < \hat{\mu}_i - \beta$ **then**
  **return** $i, \boldsymbol{\omega}, \textit{None}$ `// Declare i as winner, further episodes are None`
**else if** $\hat{\mu}_i + \beta < (1-\alpha)(\hat{\mu}_\ell - \beta)$ **then**
  **return** $i+1, \boldsymbol{\omega}, i+1$ `// Sample next candidate arm, Rounds ω unchanged, Update episode to that of next candidate arm`
**else**
  $\omega_i \leftarrow \omega_i + 1$ `// Increment round only for arm i being evaluated`
  **return** $i, \boldsymbol{\omega}, i$ `// Move to next round within same episode`

**Algorithm 1:** Pairwise Elimination (PE) for a known reference arm $\ell$

**Inputs:** Bandit Instance $\nu$, Horizon $T$, Reference Arm $\ell$.
**Initialize:** Samples $n_k = 0$, Empirical Means $\hat{\mu}_k = 0$, Current Rounds $\omega_k = 0$, $\forall\, k \in [K]$, PE Episode $i = 1$.

$\nu \leftarrow$ `reorder_per_cost`$(\nu)$ `// reorder into ascending cost order`
**while** $t \leq T$ **do**
  **if** $i \notin \{\textit{None}, \ell\}$ **then**
    $k_t, \boldsymbol{\omega}, i \leftarrow \text{PE}(\hat{\boldsymbol{\mu}}, \ell, \boldsymbol{n}, T, \boldsymbol{\omega}, i, \alpha)$ `// receive arm to be sampled, updated round numbers, and updated episode number`
  **else**
    $k_t \leftarrow k_{t-1}$ `// sample winning arm for remaining budget`
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow \text{sample\_and\_update}(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$

This scheme described for PE is made precise in Algorithm 1. For this novel PE algorithm, we were able to show the guarantees on cumulative cost and quality regret stated in Theorem 3.1. The proof for Theorem 3.1 is available in Appendix E.

**Practical Extensions of PE.** We highlight that although PE makes comparisons between the candidate arms and the reference arm in a pairwise manner, samples of the reference arm are re-used across episodes. This sample reuse is a key feature of the PE algorithm and endows it with good sample efficiency, since the samples accrued during most episodes shall be limited to the ones of the candidate arm undergoing evaluation for its return exceeding $\mu_{\text{CS}}$. In PE as presented in Algorithm 1, during an arbitrary episode evaluating the candidacy of arm $i$, the reference arm $\ell$ shall only ever have to be sampled if the number of samples of arm $i$ exceeds the samples reached by the reference arm $\ell$ in the episodes through $i-1$ preceding episode $i$. In practice, we can implement another version of PE called asymmetric-PE. Asymmetric-PE allows for a mismatch between the number of samples for the arm $i$ under evaluation and the reference arm $\ell$. The details of asymmetric-PE and an example comparing performance to PE are available in Appendix B.

**Theorem 3.1** (Instance dependent upper bound on Cumulative Cost and Quality Regret for Pairwise Elimination)**.** *For bandit instance $\nu$, over horizon $T$, the expected cumulative cost and quality regret of the PE algorithm are upper bounded as*

$$\mathbb{E}\left[\textit{Cost\_Reg}\left(T,\nu\right)\right] < \left(1+\max_{i\leq a^*}\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}\right)\Delta_{C,\ell}+\left(\sum_{i=1}^{a^*}\frac{43}{\Delta_{Q,i}^2}\right)\Delta_{C,\ell} + \frac{43}{\Delta_{Q,a^*}^2}\max_{i>a^*}\Delta_{C,i}.$$

$$\mathbb{E}\left[\textit{Quality\_Reg}\left(T,\nu\right)\right] < \sum_{i=1}^{a^*-1}\left(\Delta_{Q,i}+\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}\right)+\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}}+\frac{43}{\Delta_{Q,a^*}^2}\max_{i>a^*}\Delta_{Q,i}^{+}.$$

Thus, PE achieves *logarithmic* expected cost and quality regret in $T$. We call $a^*$ the best action since it is the only action sampling which leads to accumulation of neither cost nor quality regret.

### 3.2 PE-CS and Full Cost Subsidy Setting

**Algorithm 2:** Pairwise Elimination for Cost Subsidy Problem (PE-CS).

**Inputs:** Bandit Instance $\nu$, Horizon $T$, Subsidy Factor $\alpha$.
**Initialize:** Samples $n_k = 0$, Empirical Means $\hat{\mu}_k = 0$, Current Rounds $\omega_k = 0$, $\forall\, k \in [K]$, BAI Candidates (Active Arms) $\mathcal{A} = [K]$, Arm $\ell =$ None, PE Episode $i = 1$.

$\nu \leftarrow$ `reorder_per_cost`$(\nu)$ `// reorder indices into ascending cost order`
**while** $t \leq T$ **do**
  **if** *len*$(\mathcal{A}) > 1$ **then**
    $k_t, \boldsymbol{\omega}, \mathcal{A} \leftarrow \text{BAI}(\hat{\boldsymbol{\mu}}, \boldsymbol{n}, T, \boldsymbol{\omega}, \mathcal{A})$ `// receive arm to be sampled, updated round numbers, and updated active arms`
    **if** *len*$(\mathcal{A}) = 1$ **then**
      $\ell \leftarrow \mathcal{A}[0]$ `// set` $\ell$ `to be identified best arm`
      **continue** `// ignore sample recommendation` $k_t$
  **else if** $i \notin \{\textit{None}, \ell\}$ **then**
    $k_t, \boldsymbol{\omega}, i \leftarrow \text{PE}(\hat{\boldsymbol{\mu}}, \ell, \boldsymbol{n}, T, \boldsymbol{\omega}, i, \alpha)$ `// receive arm to be sampled, updated round numbers, and updated episode number`
  **else**
    $k_t \leftarrow k_{t-1}$ `// sample winning arm for remaining budget`
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow \text{sample\_and\_update}(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$

When solving the MAB-CS problem in the full cost subsidy setting we face the additional challenge that the arm whose $(1-\alpha)$ subsidized reward we must calibrate against is unidentified. To solve this problem, we present the novel PE-CS algorithm that comprises two stages, the **BAI stage** and the **PE stage**. Importantly, information about the highest number of samples reached in each stage is shared through the vector $\omega$. This information sharing allows for maximal sample re-usage and prevents idle elimination checks.

To solve the full cost-subsidy setting using our Pairwise-Elimination style approach, we must first identify the optimal arm $i^*$. We achieve this using a BAI stage also implemented using Improved UCB and is specified in Function 2 in Appendix A. The difference between our BAI() method and Improved-UCB is that BAI() has no exploit phase. Under BAI() once the set of active arms collapses to a single arm, sampling decisions are passed over to the PE stage. In passing over control from the BAI stage to the PE stage in PE-CS, a key role is played by the $\omega$ vector. Used in both Algorithms 1 and2, omega is responsible for tracking the point up to which any bandit arm has been sampled at any point in time. Once the time comes for passing over control from BAI to PE, the PE stage is able to pick up sampling and elimination checks in any of its pairwise comparison episodes right where the BAI stage left-off sampling.

The PE stage in the PE-CS algorithm works identically to the PE algorithm described in Algorithm 1. In fact, due to the modular and phased nature of PE-CS in the description of PE-CS in Algorithm 2 we use precisely the same function block for PE (Function 2). For the PE-CS algorithm, there is all the more reason to work to make best use of the accumulated samples of the reference arm since not only does the accumulation occur in the various episodes of the PE-stage, but also it occurs in the BAI-stage where the reference arm is by construction the last surviving and therefore the most sampled arm.

We analyze the performance of PE-CS and prove upper bounds on its expected cumulative cost and quality regret in Theorem 3.2. Our modular analysis allows us to sequester the outcomes of the BAI stage where best arm is identified incorrectly and condition on the event that the identified arm is indeed correct. The contribution to expected cumulative cost and quality regret from an incorrect identification of arm $i^*$ is shown to be a constant. Moreover, conditioned on $i^*$ being identified correctly, we are able to analyze the PE stage of PE-CS in a manner that closely parallels the analysis of PE.

**Theorem 3.2** (Instance dependent upper bound on Cumulative Cost and Quality Regret for PE-CS)**.**

$$\begin{aligned}
\mathbb{E}\left[\textit{Cost_Reg}\left(T,\nu\right)\right] < \Delta_{C,i^*}\left(1+\left\{\frac{32\log\left(T\Delta_{\min}^2\right)}{\Delta_{\min}^2}\right\}\right) &+ \sum_{i>a^*,i\neq i^*}\Delta_{C,i}\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\right)\\
&+ \Delta_{C,\max}\left(\frac{11}{\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{\Delta_j^2}\right)+\max_{i>a^*}\Delta_{C,i}\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\\
&+ \Delta_{C,i^*}\left(\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}+\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\right).\\
\mathbb{E}\left[\textit{Quality_Reg}\left(T,\nu\right)\right] < \sum_{i=1}^{a^*-1}\left(\Delta_{Q,i}+\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}\right) &+ \sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^{+}\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\right)\\
&+ \Delta_{Q,\max}^{+}\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)+\max_{i>a^*}\Delta_{Q,i}^{+}\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\\
&+ \sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}}\\
&= \mathcal{O}\left(K+a^*\right)\log(T).
\end{aligned}$$

*Where $a^*$ is the index of the best action and $i^* \geq a^*$ is the index of the best arm. $\Delta_{\min} = \min\left\{\{\Delta_i\}_{i\neq i^*},\{\Delta_{Q,i}\}_{i\leq a^*}\right\}$ is the smallest gap in reward that PE-CS has to contend with.*

As with Theorem 3.1, we see that both the expected cost and quality regret are bounded by a quantity logarithmic in $T$.

## 4 Experiments

While in Section 3 we presented the PE and PE-CS algorithms to be the ideal choice for the MAB-CS setting with known and unknown reference arms respectively, the validation for our approach so far has been based exclusively on theoretical upper bounds on cumulative expected cost and quality regret. In this section, we complement our theoretical analysis with a study of the empirical performance of our methods. In particular, we compare PE and PE-CS with baselines from Sinha et al. (2021) on a problem instance derived from a real-world dataset. The real-world dataset we use is the MovieLens 25M dataset (Harper & Konstan, 2015). Next we describe how we make use of this dataset for our experiments.

The MovieLens 25M dataset consists of 25 million ratings for 62,000 movies rated by 162,000 users. It is a popular dataset for studying the performance of recommendation systems. The movie ratings

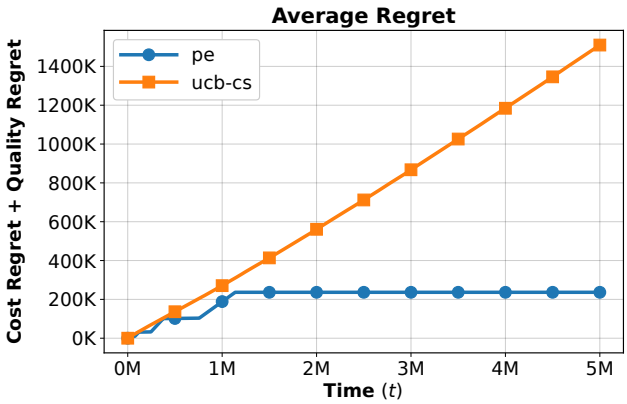


Figure 1: Fixed $\ell$ MovieLens Experiment. In the experiment: $\ell = 11$, $\mu_\ell = 0.703$, $\mu_{\text{CS}} = 0.668$.

in the dataset are from numerous users and on a 5 point scale. Moreover, every movie for which ratings are available in the dataset are tagged with one or more genres. Our MovieLens experiment simulates a scenario where a movie streaming website such as Netflix attempts to recommend movies to its users. The available movie selection is organized into genres and the movie streaming website deploys a recommendation system to decide which genre of movie should be served to its user. Whenever a genre is chosen by the recommendation system, a random movie tagged with that genre is drawn with replacement. Under these conditions, bandit arms become a natural abstraction for movie genres. The cost associated with pulling the arm corresponding to a certain genre is simply the average of the royalties that must be paid to the movie producer for every new streaming of a movie. Since royalty data is unavailable as part of MovieLens 25M, we sample the costs associated with sampling any bandit arm (genre) to lie uniformly at random between 0 and 1.

For every genre we first obtain the mean 5-point scale rating of all movies tagged with that genre and then divide this rating by 5 so that it lies between 0 and 1. We then treat this fractional rating as the expected reward return from that genre. Through this process we end up with a bandit instance consisting of 20 arms corresponding to the 20 distinct genres the details of which are available in Table 1. In all the experiments discussed in Section 4 we plot the summed together values of the cost regret and the quality regret. Looking at the summed regret allows us to view the performance of our method as a whole. The only way for an algorithm to perform well on summed regret is if it locks on satisfactorily to the best action $a^*$.

| **Number of Arms** | **Reward Spread** | **Cost Spread** | **Subsidy Factor** $\alpha$ |
|---|---|---|---|
| 20 | 0.659 - 0.785 | 0.02 - 0.964 | 0.05 |

Table 1: Information about the MovieLens Bandit Instance

### 4.1 Evaluating Our Pairwise Elimination (PE) Algorithm

To understand the effectiveness of PE empirically, we compare PE to a natural variant of the UCB-CS algorithm from Sinha et al. (2021). In the specification of UCB-CS (Algorithm 11 defined in Appendix A), the target reference arm (whose reward determines $\mu_{\text{CS}}$) is the optimal arm $i^*$. UCB-CS estimates the index of arm $i^*$ as the arm with the largest UCB-index in any time-slot. To develop a comparison with PE for the known reference arm $\ell$ setting, we simply replace this estimate with the true known fixed index of the reference arm while keeping the rest of the Algorithm the same. We call this variant of UCB-CS as UCB-CS Known $\ell$ and compare it to PE on the MovieLens Bandit Instance described earlier.

In Figure 1 we plot the performance of our novel PE algorithm and compare it against UCB-CS Known $\ell$. Since there is no notion of a reference arm inherent to the MovieLens dataset, we arrange the 20 arms in the MovieLens bandit instance in ascending cost order and assign the 11th arm in the sequence to be the reference arm. To achieve sub-linear summed regret, an algorithm must

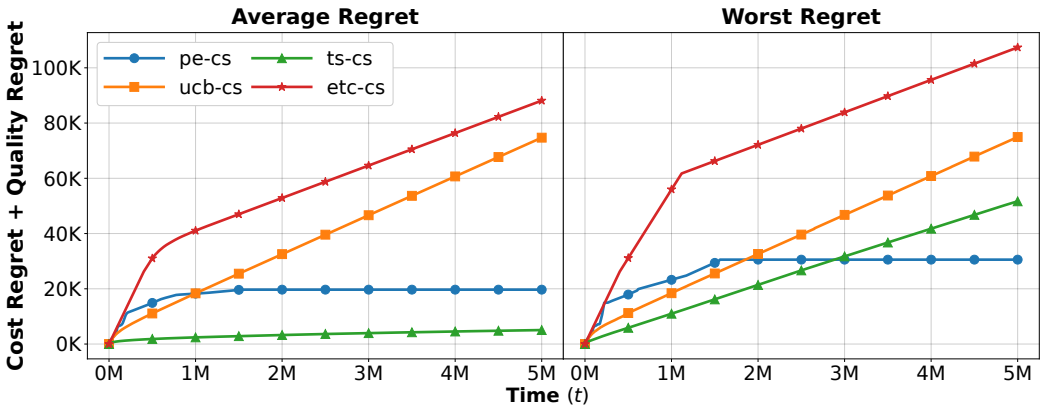


Figure 2: Unknown $i^*$ MovieLens Experiment. In the experiment: $\mu^* = 0.785$, $\mu_{\text{CS}} = 0.746$.

primarily sample the best action and sample the other actions at most a sub-linear in horizon $T$ number of times. Through the results of Figure 1 we find that while PE, through its cost-ordered arm-elimination scheme achieves this goal, UCB-CS known-$\ell$ is unsuccessful at it.

From the results in Section 4.1 we see that an interleaved exploration and exploitation approach such as UCB-CS known $\ell$ does not work for the known $\ell$ setting. Next in Section 4.2 we go into the experiments for PE-CS in the full cost-subsidy setting.

### 4.2 Evaluating Our PE-CS (Pairwise Elimination Cost Subsidy) Algorithm

In Section 3, we saw that PE-CS admitted logarithmic instance dependent guarantees on expected cumulative cost and quality regret. The only algorithm for the full cost-subsidy problem from the literature that has an upper bound guarantee on expected cumulative regret is the ETC-CS algorithm from Sinha et al. (2021). Moreover their work also prescribes the UCB-CS and TS-CS algorithms which are approaches to solving the full cost subsidy problem that interleave exploration and exploitation that lack any performance guarantees. The three algorithms ETC-CS, UCB-CS, and TS-CS comprise all the algorithms from the literature and are specified in Appendix A. We compare PE-CS against all three of these approaches on the MovieLens bandit instance of Table 1.

In Figure 2 we have simulated PE-CS, UCB-CS, TS-CS, and ETC-CS on the MovieLens bandit instance for a horizon of 5 Million samples and we plot the summed cost and quality regret for an average over the 100 independent runs on the left, and the worst performing case among the 100 runs for each of the algorithm on the right. Here the notion of worst is in terms of the summed terminal regret. We find that PE-CS significantly outperforms ETC-CS which is the only other algorithm with an upper bound on regret. As the only other algorithm with a guarantee on summed regret, ETC-CS is our primary competitor. As mentioned earlier ETC-CS has $\mathcal{O}\left(T^{2/3}\right)$ guarantee on summed cost plus quality regret while PE-CS has a $\mathcal{O}\left(\log T\right)$. We see this difference reflected in the performance results of Figure 2.

Moreover PE-CS also outperforms the UCB-CS baseline and the latter algorithm has a linear regret trend arising from a persistent mis-identification of the best action. Although we find that the average of the regret over the 100 independent runs for TS-CS is lower than PE-CS, a closer examination of the regret trend reveals the problem with the performance of TS-CS. While initially it takes PE-CS more exploration to lock onto the best action, it does so in a consistent and reliable way and once it does, there is no further incremental regret. This is seen from both the average and worst-case regret trends in Figure 2 and from our upper bounds in Theorem 3.2. Whereas for TS-CS, while interleaving exploration and exploitation leads to lower regret at the outset, there is a distinct slow-but steady upward trend in regret observed for the method. The worst case summed regret trend reveals that a consistent failure to identify and exploit the best-action occurs for the worst performing case of TS-CS exhibiting its unreliability. The worst and other similar traces of TS-CS contributing linear regret disproportionately contribute to the upward regret trend for TS-CS.

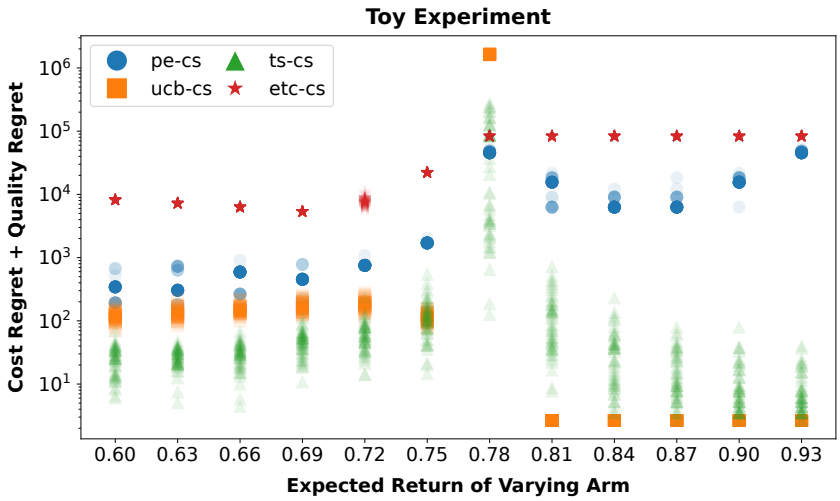


Figure 3: Toy Experiment performed using a family of four armed bandit instances. Expected rewards: $\mu_1 = 0.6, \ldots, 0.93, \mu_2 = 0.81, \mu_3 = 0.95, \mu_4 = 0.8$, and costs: $c_1 = 0.05, c_2 = 0.9, c_3 = 0.9, c_4 = 1.0$, subsidy factor: $\alpha = 0.2$.

A close examination of the Algorithm descriptions for UCB-CS and TS-CS reveals the source of their unreliability that we see in Figure 2. Both these algorithms work by constructing a set of empirically satisfactory arms and choose to sample the cheapest arm in this empirical set. This approach is vulnerable to a satisfactory arm being consistently excluded as a result of an initially poor performance. In our PE-CS on the other hand there is a systematic comparison between the arms where they are evaluated in the order of their costs and eliminated only when they have been sampled to be excluded with sufficient confidence.

To take a closer look at the sensitivity of PE-CS and all three baselines, we perform an additional synthetic experiment in the known reference arm setting which we call the toy experiment. For the toy experiment, we create a family of 12 full cost subsidy problem instances each with four arms. Then we run all four algorithms over 50 independent runs of each instance. The expected reward of the first arm in the instance varies uniformly in the range 0.6 through 0.93 over the 12 instances in the family. Since the optimal return in all instances is $\mu^* = 0.95$ and the subsidy factor is $\alpha = 0.2$, the reward threshold $\mu_{\text{CS}}$ for all the instances is $0.8 \times 0.95 = 0.76$.

In Figure 3 we plot the results from the toy experiment on a scatter plot. On the y-axis is the summed terminal cost and quality regret (in log scale) and on the x-axis is the value of the varying expected return of the first arm of the instance family. Firstly, we find that on almost all instances, PE-CS performs either similar to or better than our primary comparator ETC-CS. Among the 12 instances tested here, the ones where the return of an arm is close to $\mu_{\text{CS}} = 0.76$ are the most challenging. We find that most runs of UCB-CS and several runs of TS-CS are unsuccessful at satisfactorily solving the $\mu_1 = 0.78$ case. Moreover, from Figures 2 and 3 we conclude that among the four algorithms tested here, PE-CS arguably offers the best balance between performance and reliability.

## REFERENCES

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.

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

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`.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pp. 3123–3148. PMLR, 2022.

Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, and Osman Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 67(10):6711–6732, 2021.

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.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690, 2008.

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

Deeksha Sinha, Karthik Abinav Sankararaman, Abbas Kazerouni, and Vashist Avadhanula. Multi-armed bandits with cost subsidy. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3016–3024. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/sinha21a.html.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari. Conservative bandits. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/wu16.html.

# A Appendix Algorithms

In this section of the Appendix. We provide a precise specification of algorithms from prior work that we compare our methods against. In particular these are the ETC-CS, TS-CS, and UCB-CS Algorithms introduced by Sinha et al. (2021) and specified in Algorithms 12, 15, 11 respectively. For the upper bound on Regret Guarantee provided in Sinha et al. (2021) to hold, we require the exploration budget $\tau$ to satisfy $\tau = c\left(\frac{T}{K}\right)^{\frac{2}{3}}$ where $c$ is some unspecified proportionality constant. Based on a few trial runs and examples we pick $c = 5$ since in the average case, the above seemed to give the best performance overall for the ETC-CS approach.

**Algorithm 3:** Cost-Subsidized Explore-Then-Commit (ETC-CS)

**Inputs:** Bandit instance $\nu$, Cost Vector $\mathbf{c}$, Horizon $T$, Exploration Budget $\tau$.
**Initialize:** Empirical means $\hat{\mu}_k = 0$, Number of Samples $n_k = 0$, $\forall\, k \in [K]$.

**while** $t \leq K\tau$ **do**
  $k_t \leftarrow t \mod K$
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow$ sample_and_update$(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$
**while** $K\tau < t \leq T$ **do**
  **for** $i \in [K]$ **do**
    $\beta_i(t) \leftarrow \sqrt{\frac{2\log T}{n_i(t)}}$
    $\mu_i^{\text{UCB}} \leftarrow \min\{\hat{\mu}_i(t) + \beta_i(t), 1\}$
    $\mu_i^{\text{LCB}} \leftarrow \max\{\hat{\mu}_i(t) - \beta_i(t), 0\}$
  $\ell_t \leftarrow \arg\max_{i\in[K]} \mu_i^{\text{LCB}}(t)$
  $\text{Feas}(t) \leftarrow \left\{i : \mu_i^{\text{UCB}}(t) \geq (1-\alpha)\mu_{\ell_t}^{\text{LCB}}(t)\right\}$
  $k_t \leftarrow \arg\min_{i\in\text{Feas}(t)} c_i$
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow$ sample_and_update$(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$

**Algorithm 4:** Cost-Subsidized Thompson Sampling with Beta Priors (TS-CS)

**Inputs:** Bandit Instance $\nu$, Cost Vector $\mathbf{c}$, Subsidy Factor $\alpha$, Beta Priors and Binomial Likelihood (Bernoulli Rewards).
**Initialize:** Successes $S_k = 0$, Failures $F_k = 0$ $\forall\, k \in [K]$.

**while** $t \leq K$ **do**
  $k_t \leftarrow t$
  $r_t \leftarrow \texttt{sample}(k_t)$
  $S_{k_t}(t+1) \leftarrow S_{k_t}(t) + r_t$
  $F_{k_t}(t+1) \leftarrow F_{k_t}(t) + r_t$
  $t \leftarrow t+1$
**while** $K < t \leq T$ **do**
  **for** $i \in [K]$ **do**
    $\theta_i(t) \sim \text{Beta}\left(S_i(t)+1, F_i(t)+1\right)$
  $\ell_t \leftarrow \arg\max_{i\in[K]} \theta_i(t)$
  $\text{Feas}(t) \leftarrow \{i : \theta_i(t) \geq (1-\alpha)\theta_{\ell_t}(t)\}$
  $k_t \leftarrow \arg\min_{i\in\text{Feas}(t)} c_i$
  $r_t \leftarrow \texttt{sample}(k_t)$
  $S_{k_t}(t+1) \leftarrow S_{k_t}(t) + r_t$
  $F_{k_t}(t+1) \leftarrow F_{k_t}(t) + (1-r_t)$

We also provide here the specification of the BAI stage of the PE-CS algorithm introduced in Section 3.

**Algorithm 5:** Cost-Subsidized UCB (UCB-CS)

**Inputs:** Bandit Instance $\nu$, Cost Vector $\mathbf{c}$, Horizon $T$, Subsidy factor $\alpha$.
**Initialize:** Empirical means $\hat{\mu}_k = 0$, Number of Samples $n_k = 0 \ \forall k \in [K]$.

**while** $t \leq K$ **do**
  $k_t \leftarrow t$
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow$ sample_and_update$(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$
**while** $K < t \leq T$ **do**
  **for** $i \in [K]$ **do**
    $\beta_i(t) \leftarrow \sqrt{\frac{2 \log T}{n_i(t)}}$
    $\mu_i^{\text{UCB}} \leftarrow \min\{\hat{\mu}_i(t) + \beta_i(t), 1\}$
  $\ell_t \leftarrow \arg\max_{i \in [K]} \mu_i^{\text{UCB}}(t)$
  $\text{Feas}(t) \leftarrow \left\{i \in [K] : \mu_i^{\text{UCB}} \geq (1-\alpha) \times \mu_{\ell_t}^{\text{UCB}}\right\}$
  $k_t \leftarrow \arg\min_{i \in \text{Feas}(t)} c_i$
  $\hat{\boldsymbol{\mu}}(t+1), \boldsymbol{n}(t+1), t \leftarrow$ sample_and_update$(k_t, \hat{\boldsymbol{\mu}}(t), \boldsymbol{n}(t), t)$

**Function 2:** Best Arm Identification BAI()

**Function** *BAI( $\hat{\boldsymbol{\mu}}$: Empirical Means, $\boldsymbol{n}$: Sample Vector, $T$: Horizon, $\boldsymbol{\omega}$: Round Numbers, $\mathcal{A}$: Active Arms )*:

  $\tilde{\Delta} \leftarrow 2^{-\max_{i \in [K]} \omega_i}$
  $\tau \leftarrow \left\lceil \frac{2 \log\left(T \tilde{\Delta}^2\right)}{\tilde{\Delta}^2} \right\rceil$
  **for** $k \in \mathcal{A}$ **do**
    **if** $n_k < \tau$ **then**
      **return** $k, \boldsymbol{\omega}, \mathcal{A}$ `// next arm to be sampled, unchanged ω and A`
  $\beta \leftarrow \sqrt{\frac{\log\left(T \tilde{\Delta}^2\right)}{2\tau}}$
  **for** $i \in \mathcal{A}$ **do**
    $\mu_i^{\text{UCB}}, \mu_i^{\text{LCB}} \leftarrow \hat{\mu}_i + \beta, \hat{\mu}_i - \beta$
  $\mathcal{A}^+ \leftarrow \{i \in \mathcal{A} : \mu_i^{\text{UCB}} \geq \max_{j \in \mathcal{A}} \mu_j^{\text{LCB}}\}$ `// update set of active arms`
  $k_t \leftarrow \text{Uniform}\left(\mathcal{A}^+\right)$ `// tentatively, the next arm to be sampled`
  **for** $i \in \mathcal{A}^+$ **do**
    $\omega_i \leftarrow \omega_i + 1$ `// increment round number for still active arms`
  **return** $k_t, \boldsymbol{\omega}, \mathcal{A}^+$

## B Asymmetric Pairwise Elimination

Algorithms 1 and 2 as described in Section 3 use Function 1 PE() as a sub-routine. While in PE() the round number $\omega_i$ is used to determine the stipulated number of samples for both the candidate arm $i$ and the reference arm $\ell$, this does not have to be the case. By the time we commence episode $i$ to evaluate the candidacy of arm $i$, we would have already accrued numerous samples of arm $\ell$. In particular, we denote the number of samples of arm $\ell$ as $n_\ell(t) = \tau_{\omega_\ell}$, where the vector $\boldsymbol{\omega}$, first introduced in Section 3, is a vector recording highest round up to which the samples of each arm have evolved. Consequently, $\omega_\ell$ is the highest round number reached for the samples of reference arm $\ell$.

Based on this observation about the greater progressed round number $\omega_\ell$ we create a variant of PE called Asymmetric-PE in Function 3 that replaces Function 1 in Algorithm 1.

Asymmetric-PE described in Function 3 has all the same inputs that conventional PE did. In addition, it has an input $\kappa$ called the Maximum round deviation. While we have no bound on how much

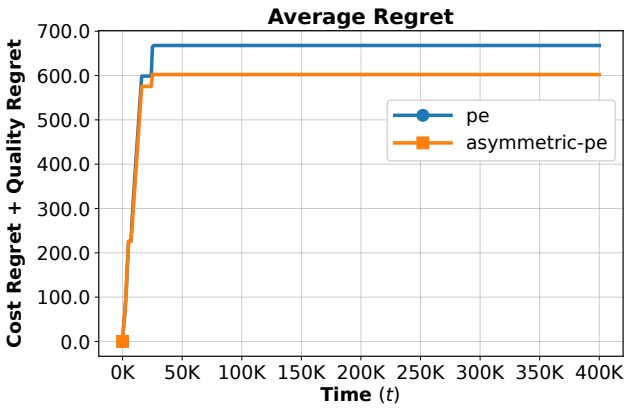


Figure 4: An experiment that illustrates potential savings in regret from the asymmetric-pe optimization. Bandit instance with Expected rewards: $\mu_1 = 0.74, \mu_2 = 0.5, \mu_3 = 0.8, \mu_4 = 0.75$, and costs: $c_1 = 0.15, c_2 = 0.2, c_3 = 0.21, c_4 = 0.25$, subsidy factor: $\alpha = 0.0$, $\ell = 3$.

**Function 3:** Asymmetric Pairwise Elimination Function Asymmetric_PE()

**Function** *Asymmetric_PE( $\hat{\boldsymbol{\mu}}$: Empirical Means, $\ell$: Reference Arm, $\boldsymbol{n}$: Sample Vector, $T$: Horizon, $\boldsymbol{\omega}$: Round Numbers, $i$: Episode, $\alpha$: Subsidy Factor, $\kappa$: Max Round Deviation )*:

1 **for** $k \in \{i, \ell\}$ **do**
2 $\quad \tilde{\Delta}_k \leftarrow 2^{-\max\{\omega_i+\kappa, \omega_k\}}$
3 $\quad \tau_k \leftarrow \left\lceil \frac{2\log(T\tilde{\Delta}_k^2)}{\tilde{\Delta}_k^2} \right\rceil$
4 $\quad$ **if** $n_k < \tau_k$ **then**
5 $\quad\quad$ **return** $k, \boldsymbol{\omega}, i$ `// $k_t = k$, round numbers $\omega$ and episode $i$ unchanged`
6 $\quad \beta_k \leftarrow \sqrt{\frac{\log(T\tilde{\Delta}_k^2)}{2\tau_k}}$
7 **if** $(1-\alpha)(\hat{\mu}_\ell + \beta_\ell) < \hat{\mu}_i - \beta_i$ **then**
8 $\quad$ **return** $i, \boldsymbol{\omega}, None$ `// Declare $i$ as winner, further episodes are None`
9 **else if** $\hat{\mu}_i + \beta_i < (1-\alpha)(\hat{\mu}_\ell - \beta_\ell)$ **then**
10 $\quad$ **return** $i+1, \boldsymbol{\omega}, i+1$ `// Sample next candidate arm, Rounds $\omega$ unchanged, Update episode to that of next candidate arm`
11 **else**
12 $\quad \omega_i \leftarrow \omega_i + 1$ `// Increment round only for arm $i$ being evaluated`
13 $\quad$ **return** $i, \boldsymbol{\omega}, i$ `// Move to next round within same episode`

larger $\omega_\ell$ is compared to $\omega_i$, when it comes to inferring the gap $\tilde{\Delta}_\ell$ corresponding to arm $\ell$ in Line 2 of Function 3, we restrict the round number we use to be at most $\kappa$ larger than the round $\omega_i$.

Effectively, the use of the further advanced round number $\omega_\ell$ trades performance for tightness of the upper bound (as we shall see in the analysis of Section E). We get an improvement in performance since $\beta_\ell \leq \beta_i$ potentially leading to a resolution in lines 7 or 9 of Function 3 with a smaller value of $\beta_i$ thereby requiring fewer samples of candidate arm $i$.

# C PRELIMINARIES

In this section of the appendix, we collate the preliminary results required for the analysis of our cost-subsidy framework algorithms that follow in the forthcoming sections. The results stated without proof are standard results from the MAB literature.

**Definition C.1** (Subgaussian Random Variable)**.** *We say that $X$ is $\sigma$-subgaussian if for any $\epsilon \geq 0$,*

$$\Pr\left(X - \mathbb{E}\left[X\right] \geq \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right). \tag{4}$$

**Lemma C.1** (Bounded random variables are Subgaussian, Example 5.6(c) in Lattimore & Szepesvári (2020))**.** *If Random Variable $X \in [a, b]$ almost surely, then $X$ is $\frac{b-a}{2}$ subgaussian.*

**Lemma C.2** (Hoeffding Bound, Section 5.4 in Lattimore & Szepesvári (2020))**.** *Let $X_1, X_2, \ldots, X_n$ 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,*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{5}$$

*Then Hoeffding's inequality states that,*

$$\Pr\left(\bar{X} - \mathbb{E}\left[X\right] \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right). \tag{6}$$

**Lemma C.3** (Iterated expectation over mutually exclusive and exhaustive events)**.** *Let $X$ be any integrable random variable over probability space $(\Omega, \mathcal{F}, \Pr)$, 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 the following identity holds,*

$$\mathbb{E}\left[X\right] = \sum_{i=1}^{n} \mathbb{E}\left[X \mid E_i\right] \Pr\left(E_i\right). \tag{7}$$

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

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[X \mid E\right] \Pr\left(E\right) + \mathbb{E}\left[X \mid E^c\right] \Pr\left(E^c\right). \tag{8}$$

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

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right]\right] \text{ (Because } \mathcal{G} \subset \mathcal{F}) \tag{9}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[X \mid E_i\right] \Pr\left(E_i\right). \tag{10}$$

□

**Lemma C.4** (Expectation is at most equal to larger of the conditioned expectations)**.** *Let $X$ be any integrable random variable over probability space $(\Omega, \mathcal{F}, \Pr)$, 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,*

$$\mathbb{E}\left[X\right] \leq \max_{i \in [n]} \left\{\mathbb{E}\left[X \mid E_i\right]\right\}. \tag{11}$$

*Proof.* Lemma C.4 can be considered a Corollary to Lemma C.3 as is illustrated by the following proof,

$$\mathbb{E}\left[X\right] = \sum_{i=1}^{n} \mathbb{E}\left[X \mid E_i\right] \Pr\left(E_i\right) \text{ (From the proof of Lemma C.3).} \tag{12}$$

$$\leq \left(\sum_{i=1}^{n} \Pr\left(E_i\right)\right) \cdot \left(\max_{i \in [n]} \left\{\mathbb{E}\left[X \mid E_i\right]\right\}\right) \tag{13}$$

$$= \max_{i \in [n]} \left\{\mathbb{E}\left[X \mid E_i\right]\right\}. \tag{14}$$

□

**Lemma C.5** (Regret Decomposition Lemma, Lemma 4.5 in Lattimore & Szepesvári (2020))**.** *For any policy $\pi$ and stochastic bandit environment $\nu$ with $K$ arms, for horizon $T$, the Expected Cumulative Regret $Reg_\pi(T, \nu)$ of policy $\pi$ in $\nu$ satisfies,*

$$\mathbb{E}\left[Reg_\pi(T, \nu)\right] = \sum_{i \in [K]} \Delta_i \mathbb{E}\left[n_i(T)\right]. \tag{15}$$

*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,*

$$\mathbb{E}\left[Reg_\pi^X(T, \nu)\right] = \sum_{i \in [K]} \Delta_{X,i} \mathbb{E}\left[n_i(T)\right]. \tag{16}$$

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

$$\mathbb{E}\left[Cost\_Reg(T, \nu)\right] = \sum_{i \in [K]} \Delta_{C,i} \mathbb{E}\left[n_i(T)\right] \tag{17}$$

$$\mathbb{E}\left[Quality\_Reg(T, \nu)\right] = \sum_{i \in [K]} \Delta_{Q,i}^{+} \mathbb{E}\left[n_i(T)\right]. \tag{18}$$

# D ANALYSIS FOR MINIMUM TOLERATED REWARD SETTING

**Algorithm 1:** MINIMUM TOLERATED REWARD (MTR) UCB

**Input:** Number of arms $K$, Known costs $c_k$ for each arm $k \in [K]$, Minimum Tolerated Reward $\mu_0$.
**Initialize:** Empirical Means $\hat{\mu}_k = 0$, Pulls of arm $k : n_k = 0,\ \forall k \in [K]$, Time step $t = 0$.
**for** *Each Round* $t$ **do**
  **if** $t \leq K$ **then**
    $k_t \leftarrow t$
  **else**
    $C_t \leftarrow \left\{k : \hat{\mu}_k(t) + \sqrt{\frac{2 \log t}{n_k(t)}} \geq \mu_0\right\}$
    **if** $C_t \neq \phi$ **then**
      $k_t \leftarrow \arg\min_{i \in C_t} c_i$
    **else**
      $k_t \leftarrow \text{Uniform}(K)$
  $X_{k_t}(t) \leftarrow \text{Ber}(\mu_{k_t})$
  $\hat{\mu}_{k_t}(t) \leftarrow \frac{\hat{\mu}_{k_t}(t-1) n_{k_t}(t-1) + X_{k_t}(t)}{n_{k_t}(t-1)+1}$
  $n_{k_t}(t) \leftarrow n_{k_t}(t-1) + 1$
  $t \leftarrow t + 1$

## D.1 GUARANTEES FOR MTR-UCB

Combining the results from the previous two subsections and the formula for Quality and Cost regrets the upper bound for Algorithm 1, is stated in Theorem D.1.

**Theorem D.1** (Overall Regret Upper Bound for MTR-UCB)**.**

$$\textit{Quality\_Reg}\,(n, \phi, \mu_0) = \sum_{i \in C^-} \mathbb{E}\left[T_i(n)\right] \Delta_i^\mu + \sum_{j \in C^+} \mathbb{E}\left[T_j(n)\right] \Delta_j^\mu \tag{19}$$

$$\leq \sum_{i \in C^-} \frac{8 \log n}{\Delta_i^\mu} + \left(1 + \frac{\pi^2}{3}\right) \sum_{j \in [K], j \neq i^*} \Delta_j^\mu \tag{20}$$

$$\textit{Cost\_Reg}\,(n, \phi, \mu_0) = \sum_{j \in C^+} \mathbb{E}\left[T_j(n)\right] \Delta_j^c \tag{21}$$

$$\leq \left(1 + \frac{\pi^2}{3}\right) \sum_{j \in C^+} \Delta_j^c \tag{22}$$

### D.1.1 ANALYSIS FOR MTR-UCB

Under this setting, the quality gaps are defined as the 0-clipped version of the gap between the known minimum tolerated reward $\mu_{\text{CS}} = \mu_0$ and the expected return of an arm $j$ : $\Delta_j^q = \max\{0, \mu_{\text{CS}} - \mu_j\}$.

For a policy $\pi$, we have, Cumulative Regret,

$$\text{Quality\_Reg}\,(T, \phi, \mu_0) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \max\left(\mu_0 - \mu_{\pi_t}, 0\right)\right] \tag{23}$$

$$\text{Cost\_Reg}\,(T, \phi, \mu_0) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \max\left(c_{\pi_t} - c_{i^*}, 0\right)\right] \tag{24}$$

$$\tag{25}$$

For a policy $\pi$, we have, Incremental Regret,

$$\text{Quality\_Reg}_{\pi}^{\text{incr}}(t, \phi, \mu_0) = \mathbb{E}_{\pi}\left[\max\left\{\mu_0 - \mu_{k_t}, 0\right\}\right] \tag{26}$$

$$\text{Cost\_Reg}_{\pi}^{\text{incr}}(t, \phi, \mu_0) = \mathbb{E}_{\pi}\left[\max\left\{c_{k_t} - c_{i^*}, 0\right\}\right] \tag{27}$$

Also define gaps,

$$\Delta_i^{\mu} = \max\left\{\mu_0 - \mu_i, 0\right\} \tag{28}$$

$$\Delta_i^{c} = \max\left\{c_i - c_{i^*}, 0\right\} \tag{29}$$

Note from 10/25: A key takeaway from the adivising meeting today was that the uncertainty of the decision making of a policy can be in the construction of the filtered set of arms or in the final decision taken about arm $k_t$ (either or, or both).

Moreover, under the sorted costs structure, for a Bandit Instance $\phi(\boldsymbol{\mu}, \boldsymbol{c})$, we have,

$$\text{Quality\_Reg}(n, \phi, \mu_0) = \sum_{i=1}^{i^*-1} \mathbb{E}\left[T_i(n)\right] \Delta_i^{\mu} + \sum_{j=i^*+1}^{K} \mathbb{E}\left[T_j(n)\right] \Delta_j^{\mu} \tag{30}$$

$$\text{Cost\_Reg}(n, \phi, \mu_0) = \sum_{j=i^*+1}^{K} \mathbb{E}\left[T_j(n)\right] \Delta_j^{c} \tag{31}$$

To bound quality regret,

1. $n_i(T)$ and $T_i(n)$ are used interchangeably to represent the number of pulls of arm $i$ up to a horizon
2. Bound $\Pr\left(k \in [i^* - 1] \text{ entering } C_t\right)$. This is enough since the contribution from the beyond $i^*$ arms will be a constant term as we shall see soon.
3. $\Pr\left(\hat{\mu}_k(t) + \sqrt{\frac{2\log t}{n_k(t)}} \geq \mu_0\right)$
4. Where, $\mu_0 > \mu_i$, $\mu_0 - \mu_i = \Delta_i^{\mu}$

$$\Pr\left(\left\{\hat{\mu}_k(t) + \sqrt{\frac{2\log t}{n_k(t)}} \geq \mu_0\right\}\right) \tag{32}$$

$$\Pr\left(\left\{\hat{\mu}_k(t) + \sqrt{\frac{2\log t}{n_k(t)}} \geq \mu_i + \Delta_i^{\mu}\right\}\right) \tag{33}$$

$$\Pr\left(\left\{\hat{\mu}_k(t) - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{n_k(t)}}\right\}\right) \tag{34}$$

To deal with the R.V. $\sqrt{\frac{2\log t}{n_k(t)}}$ ($n_k(t)$ in denominator), we do the sum/max/min trick in the original UCB proof in the Auer et al. paper.

Goal: Bound the probability of the event,

$$\hat{\mu}_k(t) + \sqrt{\frac{2\log t}{n_k(t)}} \geq \mu_0, \tag{35}$$

using the Auer et al. paper tricks.

Let $i \in [i^* - 1]$ be any arm with $\Delta_i^\mu > 0$, then,

$$T_i(n) = 1 + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i\} \tag{36}$$

$$\leq \ell + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i, n_i(t-1) \geq \ell\} \tag{37}$$

$$\leq \ell + \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t-1)}{n_k(t-1)}} \geq \mu_0, n_i(t-1) \geq \ell\right\} \tag{38}$$

At this point we do not have precise knowledge of the number of pulls $n_i(t-1)$, so we just sum over all the possibilities and modify the indexing $t^{'} = t-1$. Moreover, we update the notation $\hat{\mu}_i(s)$ to represent the mean return for arm $i$ over $s$ samples.

$$T_i(n) \leq \ell + \sum_{t=1}^{\infty}\sum_{s=\ell}^{t-1} \mathbb{I}\left\{\hat{\mu}_i(s) + \sqrt{\frac{2\log t}{s}} \geq \mu_0\right\} \tag{39}$$

$\hat{\mu}_i + \sqrt{\frac{2\log t}{s}} \geq \mu_0 \iff \hat{\mu}_i(t) - \mu_i \geq \Delta_i^\mu - \sqrt{\frac{2\log t}{s}}$.

Applying Lemma C.2 to equation 39 we get,

$$\Pr\left(\hat{\mu}_i(s) - \mu_i \geq \Delta_i^\mu - \sqrt{\frac{2\log t}{s}}\right) \leq \exp\left(-2s\left(\Delta_i^\mu - \sqrt{\frac{2\log t}{s}}\right)^2\right) \tag{40}$$

$$\leq \exp\left(-2s\left((\Delta_i^\mu)^2 + \frac{2\log t}{s} - 2\Delta_i^\mu\sqrt{\frac{2\log t}{s}}\right)\right) \tag{41}$$

$$\leq \exp\left(-2s\left(-4\log t + 4\Delta_i^\mu\sqrt{2s\log t}\right)\right) \tag{42}$$

$$= \frac{\exp\left(4\Delta_i^\mu\sqrt{2s\log t}\right)}{t^4} \tag{43}$$

- We want to make the thing inside $\exp(-2s(\cdot))$ smaller for upper bound.
- Make $\sqrt{\frac{2\log t}{s}}$ term largest $\implies$ plug in $s = \left\lceil \frac{8\log n}{(\Delta_i^\mu)^2} \right\rceil$

Refine the steps from Friday 10/27:

(a) Bound on $\mathbb{E}[T_i(n)]$ for $i < i^*$
(b) Bound on $\mathbb{E}[T_j(n)]$ for $j > i^*$

#### D.1.2 Bound on (a)

For $i < i^*$, we have $\Delta_i^\mu > 0$.

$$T_i(n) = 1 + \sum_{K+1}^{n} \mathbb{I}\{k_t = i\} \tag{44}$$

$$\leq \ell + \sum_{K+1}^{n} \mathbb{I}\{k_t = i, T_i(t-1) \geq \ell\} \tag{45}$$

$$\leq \ell + \sum_{K+1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t-1)}{T_i(t-1)}} \geq \mu_0, T_i(t-1) \geq \ell\right\}. \tag{46}$$

Here the number of pulls $s$ at time $t$ is implicit in the notation. In the next step when we sum over it, we make it explicit.

Since there is no precise knowledge of the number of pulls $T_i(n)$, we resort to summing over all possibilities and for convenience, we change the indexing of time to be $t^{'} = t - 1$, and take the sum to $\infty$.

$$\leq \sum_{t=1}^{\infty} \sum_{s=\ell}^{t-1} \mathbb{I}\left\{\hat{\mu}_i(s) + \sqrt{\frac{2\log t}{s}} \geq \mu_0\right\}. \tag{47}$$

From previous notes, eventually, on applying Lemma C.2 we get,

$$\Pr\left(\hat{\mu}_i(s) - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{s}}\right) \tag{48}$$

$$\leq \exp\left(-2s\left((\Delta_i^{\mu})^2 + \frac{2\log t}{s} - 2\Delta_i^{\mu}\sqrt{\frac{2\log t}{s}}\right)\right) \tag{49}$$

$$\tag{50}$$

Plug in $s = \frac{8\log n}{(\Delta_i^{\mu})^2}$ for the red $s$, and leave the remaining $s$ untouched to work towards the bound.

Key steps from the analysis from the morning session are recapped for continuity,

$$T_i(n) = 1 + \sum_{K+1}^{n} \mathbb{I}\{k_t = i\} \tag{51}$$

$$\leq \ell + \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t-1)}{T_i(t-1)}} \geq \mu_0, T_i(t-1) \geq \ell\right\} \tag{52}$$

$$\leq \ell + \sum_{t=1}^{n-1} \sum_{s_i=\ell}^{t-1} \mathbb{I}\left\{\hat{\mu}_i(t) + \sqrt{\frac{2\log t}{s_i}} \geq \mu_0\right\} \tag{53}$$

Take expectations on both sides,

$$\mathbb{E}[T_i(n)] \leq \ell + \sum_{t=1}^{n-1} \sum_{s_i=\ell}^{t-1} \Pr\left(\hat{\mu}_i(t) - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right) \tag{54}$$

Applying Hoeffding Inequality, we have,

$$\Pr\left(\hat{\mu_i} - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right) \leq \exp\left(-2s_i\left(\Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right)^2\right) \tag{55}$$

$$\exp\left(-2s_i\left((\Delta_i^{\mu})^2 + \frac{2\log t}{s_i} - 2\Delta_i^{\mu}\sqrt{\frac{2\log t}{s_i}}\right)\right) \tag{56}$$

Plugging into Expected number of pulls expression,

$$\mathbb{E}[T_i(n)] \leq \ell + \sum_{t=1}^{n-1} \sum_{s_i=\ell}^{t-1} \exp\left(-2s_i \left(\frac{2\log t}{s_i} + (\Delta_i^\mu)^2 - 2\Delta_i^\mu \sqrt{\frac{2\log t}{s_i}}\right)\right) \tag{57}$$

We pick $\ell = \left\lceil \frac{8\log n}{(\Delta_i^\mu)^2} \right\rceil$ as usual. (58)

$$\leq \ell + \sum_{t=1}^{n-1} \sum_{s_i=\ell}^{t-1} \exp\left(-4\log t + 2s_i(\Delta_i^\mu)^2 \left(\sqrt{\frac{\log t}{\log n}} - 1\right)\right) \tag{59}$$

since $t < n$, we have, (60)

$$\mathbb{E}\left[T_i(n)\right] \leq \frac{8\log n}{(\Delta_i^\mu)^2} + 1 + \sum_{t=1}^{n-1} \sum_{s_i=\ell}^{t-1} \exp\left(-4\log t\right) \tag{61}$$

$$\leq \frac{8\log n}{(\Delta_i^\mu)^2} + 1 + \sum_{t=1}^{\infty} \frac{1}{t^3} \tag{62}$$

$$\leq \frac{8\log n}{(\Delta_i^\mu)^2} + 1 + \frac{\pi^2}{6} \tag{63}$$

### D.1.3 Bound for term (B)

Initial Attempt on 10/30/23 corrections added on 11/3/23.

Further notational corrections added on 11/15/23.

Bound $\mathbb{E}\left[T_j(n)\right]$ for arms $j > i^*$, define, $\delta_0 = \mu_{i^*} - \mu_0, \delta_0 \geq 0$, and note that $\Delta_{i^*}^\mu = 0$

$$T_j(n) \leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = j\} \tag{64}$$

Just as a note and an aside, these arms $j$ can satisfy, $\mu_j - \mu_0 \geq 0$ or $\mu_j - \mu_0 < 0$. (65)

The former incurs no quality regret but the latter does. (66)

$$\leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\{i^* \notin C_t\} \quad \text{otherwise arm } i^* \text{ would have been pulled and not arm } j \tag{67}$$

$$\leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_{i^*}(t-1) + \sqrt{\frac{2\log(t-1)}{n_{i^*}(t-1)}} < \mu_0\right\} \tag{68}$$

$$= 1 + \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_{i^*}(t-1) - \mu_{i^*} < -\delta_0 - \sqrt{\frac{2\log(t-1)}{n_{i^*}(t-1)}}\right\} \tag{69}$$

Since no direct handle on $s$ is available, we sum over all its possible values (70)

Also NB: We update the notation $\hat{\mu}_i(s)$ to represent (71)

the mean return for arm $i$ over $s$ samples. (72)

$$T_j(n) \leq 1 + \sum_{t=K+1}^{n} \sum_{s=1}^{t-1} \mathbb{I}\left\{\hat{\mu}_{i^*}(s) - \mu_{i^*} < -\delta_0 - \sqrt{\frac{2\log(t-1)}{s}}\right\} \tag{73}$$

Taking expectations, and perform a shift of 1 in time-indexing, we have, (74)

$$\mathbb{E}\left[T_j(n)\right] \leq 1 + \sum_{t=1}^{n-1} \sum_{s=1}^{t-1} \Pr\left(\hat{\mu}_{i^*}(s) - \mu_{i^*} < -\delta_0 - \sqrt{\frac{2\log(t-1)}{s}}\right) \tag{75}$$

$$\leq 1 + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \exp\left(-2s\frac{2\log t}{s}\right) \tag{76}$$

$$= 1 + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \frac{1}{t^4} \tag{77}$$

$$\leq 1 + \frac{\pi^2}{6} \tag{78}$$

#### D.1.4 Final Upper Bound Analysis

MAB instance: $\phi\left(\boldsymbol{\mu}, \boldsymbol{c}\right)$

For this instance, without loss of generality, we assume,

$$\boldsymbol{c} = [c_1, c_2, \ldots, c_K] \tag{79}$$

$$\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K], \text{where,} \tag{80}$$

$$c_1 \leq c_2 \leq c_3 \leq \ldots \leq c_{i^*-1} \leq c_{i^*} \leq c_{i^*+1} \leq \ldots \leq c_K \tag{81}$$

In the MAB-CS with minimum tolerated reward setting, the optimal arm $i^*$ is the least cost arm that has expected reward more than $\mu_0$. We also define $C^*$ to be set of feasible arms per the following,

$$C^* = \{k : \mu_k \geq \mu_0, k \in [K]\} \tag{82}$$

$$i^* = \arg\min_{i \in C^*} c_i. \tag{83}$$

We assume $C^*$ is non-empty, i.e. at least one arm has expected return higher than the minimum-tolerated-reward $\mu_0$.

Further, for the analysis, define,

$$C^- = \{k : c_k < c_{i^*}, k \in [K]\} \tag{84}$$

$$C^+ = \{k : c_k \geq c_{i^*}, k \in [K]\} \tag{85}$$

$$\text{Quality\_Reg}\,(n, \phi, \mu_0) = \sum_{i \in C^-} \mathbb{E}\left[T_i(n)\right] \Delta_i^\mu + \sum_{j \in C^+} \mathbb{E}\left[T_j(n)\right] \Delta_j^\mu \tag{86}$$

$$\text{Cost\_Reg}\,(n, \phi, \mu_0) = \sum_{j \in C^+} \mathbb{E}\left[T_j(n)\right] \Delta_j^c \tag{87}$$

#### D.1.5 Bound on the Expected Number of pulls of High Cost Arms

- An arm $j \in C^+$ contributes to Quality Regret if $\mu_j < \mu_0$, otherwise $\Delta_j^\mu = 0$ and there is no contribution
- An arm $j \in C^+$ always contributes to Cost Regret

$$T_j(n) \leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\left\{k_t = j\right\} \tag{88}$$

$$\leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\left\{k_t = j, C_t \neq \phi\right\} + \sum_{t=K+1}^{n} \mathbb{I}\left\{k_t = j, C_t = \phi\right\} \tag{89}$$

$$\leq 1 + \sum_{t=K+1}^{n} \mathbb{I}\left\{i^* \notin C_t\right\} + \sum_{t=K+1}^{n} \mathbb{I}\left\{i^* \notin C_t\right\} \tag{90}$$

$$\leq 1 + 2 \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_{i^*}\left(T_{i^*}(t-1)\right) + \sqrt{\frac{2\log(t-1)}{T_{i^*}(t-1)}} < \mu_0\right\} \tag{91}$$

$$= 1 + 2 \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_{i^*}\left(T_{i^*}(t-1)\right) - \mu_{i^*} < \mu_0 - \mu_{i^*} - \sqrt{\frac{2\log(t-1)}{T_{i^*}(t-1)}}\right\} \tag{92}$$

$$\leq 1 + 2 \sum_{t=K+1}^{n} \sum_{s=1}^{t-1} \mathbb{I}\left\{\hat{\mu}_{i^*}(s) - \mu_{i^*} < \mu_0 - \mu_{i^*} - \sqrt{\frac{2\log(t-1)}{s}}\right\} \tag{93}$$

Taking expectations, and perform a shift of 1 in time-indexing, we have,

$$\mathbb{E}\left[T_j(n)\right] \leq 1 + 2 \sum_{t=1}^{\infty} \sum_{s=1}^{t} \Pr\left(\hat{\mu}_{i^*}(s) - \mu_{i^*} < \mu_0 - \mu_{i^*} - \sqrt{\frac{2\log t}{s}}\right) \tag{94}$$

$$\leq 1 + 2 \sum_{t=1}^{\infty} \sum_{s=1}^{t} \exp\left(-2s \frac{2\log t}{s}\right) \tag{95}$$

$$= 1 + 2 \sum_{t=1}^{\infty} \sum_{s=1}^{t} \frac{1}{t^4} \tag{96}$$

$$\leq 1 + \frac{\pi^2}{3} \tag{97}$$

Therefore, the expected number of pulls of a high-cost arm under MTR-UCB is given by,

$$\mathbb{E}\left[T_j(n)\right] \leq 1 + \frac{\pi^2}{3} \tag{98}$$

### D.1.6 BOUND ON THE PULLS OF LOW-COST UNSATISFACTORY ARMS

- An arm $i \in C^-$ always contributes to Quality Regret and does not contribute to cost regret

$$T_i(n) = 1 + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i, C_t \neq \phi\} + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i, C_t = \phi\} \tag{99}$$

$$= 1 + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i, C_t \neq \phi\} + \sum_{t=K+1}^{n} \mathbb{I}\{i^* \notin C_t\} \tag{100}$$

$$= \ell + \sum_{t=K+1}^{n} \mathbb{I}\{k_t = i, C_t \neq \phi, T_i(t-1) \geq \ell\} + \sum_{t=K+1}^{n} \mathbb{I}\{i^* \notin C_t\} \tag{101}$$

$$= \ell + \sum_{t=K+1}^{n} \mathbb{I}\left\{\hat{\mu}_i\left(T_i(t-1)\right) + \sqrt{\frac{2\log(t-1)}{T_i(t-1)}} \geq \mu_0, T_i(t-1) \geq \ell\right\} + \sum_{t=K+1}^{n} \mathbb{I}\{i^* \notin C_t\} \tag{102}$$

$$\leq \ell + \sum_{t=1}^{\infty}\sum_{s=\ell}^{t} \mathbb{I}\left\{\hat{\mu}_i(s) + \sqrt{\frac{2\log t}{s}} \geq \mu_0\right\} + \sum_{t=1}^{\infty} \mathbb{I}\{i^* \notin C_t\} \tag{103}$$

Taking expectations on both sides, we have,

$$\mathbb{E}[T_i(n)] \leq \ell + \sum_{t=1}^{\infty}\sum_{s_i=\ell}^{t} \Pr\left(\hat{\mu}_i(s_i) - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right) + \sum_{t=1}^{\infty} \Pr\left(i^* \notin C_t\right) \tag{104}$$

Applying Hoeffding Inequality, we have,

$$\Pr\left(\hat{\mu}_i - \mu_i \geq \Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right) \leq \exp\left(-2s_i\left(\Delta_i^{\mu} - \sqrt{\frac{2\log t}{s_i}}\right)^2\right) \tag{105}$$

$$\leq \exp\left(-2s_i\left((\Delta_i^{\mu})^2 + \frac{2\log t}{s_i} - 2\Delta_i^{\mu}\sqrt{\frac{2\log t}{s_i}}\right)\right) \tag{106}$$

Plugging in the bound in 106 and recognizing that we have already bounded the term $\sum_{t=1}^{\infty} \Pr\left(i^* \notin C_t\right)$ in the previous subsection when we bounded the number of pulls of a high-cost arm, we have,

$$\mathbb{E}[T_i(n)] \leq \ell + \sum_{t=1}^{\infty}\sum_{s_i=\ell}^{t} \exp\left(-2s_i\left(\frac{2\log t}{s_i} + (\Delta_i^{\mu})^2 - 2\Delta_i^{\mu}\sqrt{\frac{2\log t}{s_i}}\right)\right) + \frac{\pi^2}{6} \tag{107}$$

We pick $\ell = \left\lceil \frac{8\log n}{(\Delta_i^{\mu})^2} \right\rceil$ as in the proof technique in Auer et al. (2002)

$$\mathbb{E}[T_i(n)] \leq \ell + \sum_{t=1}^{\infty}\sum_{s_i=\ell}^{t} \exp\left(-4\log t + 2s_i(\Delta_i^{\mu})^2\left(\sqrt{\frac{\log t}{\log n}} - 1\right)\right) + \frac{\pi^2}{6} \tag{108}$$

since $t < n$ , we have, (109)

$$\mathbb{E}\left[T_i(n)\right] \leq \frac{8\log n}{(\Delta_i^{\mu})^2} + 1 + \sum_{t=1}^{n-1}\sum_{s_i=\ell}^{t-1} \exp\left(-4\log t\right) + \frac{\pi^2}{6} \tag{110}$$

$$\leq \frac{8\log n}{(\Delta_i^{\mu})^2} + 1 + \sum_{t=1}^{\infty} \frac{1}{t^3} + \frac{\pi^2}{6} \tag{111}$$

$$\leq \frac{8\log n}{(\Delta_i^{\mu})^2} + 1 + \frac{\pi^2}{3} \tag{112}$$

Therefore, the expected number of pulls of a low-cost (necessarily unsatisfactory) arm under MTR-UCB is given by,

$$\mathbb{E}[T_i(n)] \leq \frac{8 \log n}{(\Delta_i^{\mu})^2} + 1 + \frac{\pi^2}{3} \tag{113}$$

# E Analysis for PE in the Known Reference Arm Setting

In this Section we build up to the proof of Theorem 3.1 by upper bounding the expected number of samples of low-cost unsatisfactory arms, the reference arm $\ell$, and of high-cost arms under the Asymmetric-PE setting with maximum round-deviation $\kappa$. We then particularize these results to the $\kappa = 0$ case corresponding to conventional PE to obtain an upper bound on the expected regret for Algorithm 1 PE.

## E.1 Definitions and Setup Required for Analysis

As discussed in the main paper, the PE algorithm is inspired by the Improved-UCB successive elimination approach where sampling of arms occurs in un-interrupted batches called rounds. In Improved-UCB, a set of active arms is maintained and at the end of every round, arms in the active set are re-tested for their candidacy using an elimination criteria. Since in Pairwise-Elimination, we inherit the un-interrupted round based sampling scheme and elimination-criteria first used in Improved-UCB, to prove Theorem 3.1 we build on the analysis from Auer & Ortner (2010).

For $i \leq a^*$, define round number $\rho_i$ to be,

$$\rho_i = \min \left\{ \omega_i \,|\, \tilde{\Delta}_{\omega_i} < \frac{|\Delta_{Q,i}|}{2} \right\}, \tag{114}$$

Intuitively, $\rho_i$ is the PE round number during episode $i$ by which we expect to either identify low-cost unsatisfactory arm $i < a^*$ as being unsuitable or identify the best action $a^*$ as being suitable. Moreover from Function 3, we know that for any arm the required number of samples to be drawn by round $\omega_i$ is given by,

$$\tau_{\omega_i} = \left\lceil \frac{2 \log \left( T \tilde{\Delta}^2_{\omega_i} \right)}{\tilde{\Delta}^2_{\omega_i}} \right\rceil. \tag{115}$$

Table 2: Probabilistic Events Descriptions and Symbols for PE Analysis

| Symbol | Event Description |
|---|---|
| $G_{1,i}, \forall\, i \leq a^*$ | Episode $i$ is executed to evaluate arm $i$ |
| $G_{2,i}, \forall\, i < a^*$ | Arm $i$ is eliminated by arm $\ell$ by when round $\omega_i = \rho_i$, during episode $i$ |
| $G_{2,a^*}$ | Arm $\ell$ is eliminated by arm $a^*$ by when $\omega_{a^*} = \rho_{a^*}$, during episode $a^*$ |
| $G_{3,i}, \forall\, i < a^*$ | Arm $\ell$ is not eliminated by arm $i$ by when round $\omega_i = \rho_i - 1$, during episode $i$ |
| $G_{3,a^*}$ | Arm $a^*$ is not eliminated by arm $\ell$ by when round $\omega_{a^*} = \rho_{a^*} - 1$, during episode $a^*$ |
| $E_i, \forall\, i \leq a^*$ | Available samples ran out during episode $i$ before the sampling for round $\rho_i$ could conclude and before an arm could be eliminated |

To lay the groundwork for the forthcoming analysis we introduce notation $n_i\,(t_1, t_2)$ for the random variable denoting the number of samples of arm $i$ accrued between time steps $t_1$ and $t_2$ both inclusive. Further, we use $t_i$ to denote the final time-step in episode $i$. Thereby, the variable $n_\ell(1, t_i)$ denotes the number of samples of reference arm $\ell$ accrued by the end of episode $i$.

At the outset of our analysis, we define a large collection of probabilistic events needed for developing Theorem 3.1's intermediate results in Table 2. Each event in Table 2 is a subset of the sample space $\Omega$ associated with a run of PE. In the descriptions of events in Table 2, when we say that an elimination event occurs *by a round*, we are including the round being mentioned. For example: "Arm $i$ is eliminated by round $\rho_i$" means that the arm $i$ was eliminated in round $\omega_i$ such that $\omega_i \leq \rho_i$.

**Remark E.1.** *In the event descriptions, whenever we refer to the reference arm $\ell$ available to the algorithm, we are really referring to the action whose expected return is $(1 - \alpha)\mu_\ell$, where $\alpha$ is the subsidy factor.*

Arm $\ell$ not being eliminated by arm $i$ during round $\rho_i$ is covered by $G_{2,i}$, hence the rounds range up to $\rho_i - 1$ in the definition of $G_{3,i}$. Unlike the episodes $i < a^*$ where the nominal outcome is for the reference arm to win over the candidate arm, for episode $a^*$, nominally the arm $a^*$ shall be the winner and therefore the events $G_{2,a^*}, G_{3,a^*}$ are defined separately to account for this reality.

Lastly, we define compound event $G_i$ using events in Table 2 as,

$$G_i = G_{1,i} \cap ((G_{2,i} \cap G_{3,i}) \cup E_i)\,. \tag{116}$$

In English the event $G_i, i \leq a^*$ is the event that episode $i$ occurs (event $G_{1,i}$) and that either an accurate and timely elimination of one arm by another is made (event $G_{2,i} \cap G_{3,i}$), or we run out of samples before a decision could be made and prior to the conclusion of round $\rho_i$ (event $E_i$). Conditioning the samples $n_i(T)$ for PE on event $G_i$ gives us the following key result,

$$\Pr\left(n_i(T) \leq \tau_{\rho_i} \mid G_i\right) = 1. \tag{117}$$

We leverage Equation 117 in conjunction with Lemma C.3 to prove Theorem 3.1. To use this procedure, we require Lemma E.1 that partitions the probability space $\Omega$ into mutually exclusive and exhaustive events including $G_i$.

**Lemma E.1** (Partition of $\Omega$ with $G_i$)**.** *The events $G_i$, $B_{1,i} = G_{1,i}^c$, and $B_i = G_{1,i} \cap \left(G_{2,i}^c \cup G_{3,i}^c\right) \cap E_i^c$ are exhaustive and pairwise exclusive $\forall\, i \leq a^*$.*

*Proof.* We can prove that the events stated in Lemma E.1 are mutually exclusive and exhaustive by showing that $G_i^c = B_{1,i} \cup B_i$. Since $G_i$ and $G_i^c$ are exhaustive showing so will show that the three events are exhaustive. Moreover, since $G_i$ and $G_i^c$ are mutually exclusive, and since $B_{1,i}$ and $B_i$ are mutually exclusive by construction, we would also have all the events being pairwise mutually exclusive in addition to being exhaustive.

$$\begin{aligned}
B_{1,i} \cup B_i &= G_{1,i}^c \cup \left(G_{1,i} \cap \left(G_{2,i}^c \cup G_{3,i}^c\right) \cap E_i^c\right) && (118)\\
&= \left(G_{1,i}^c \cup G_{1,i}\right) \cap \left(G_{1,i}^c \cup \left(\left(G_{2,i}^c \cup G_{3,i}^c\right) \cap E_i^c\right)\right) \quad (\cup \text{ distributes over } \cap) && (119)\\
&= G_{1,i}^c \cup \left(\left(G_{2,i}^c \cup G_{3,i}^c\right) \cap E_i^c\right) && (120)\\
&= G_i^c. && (121)
\end{aligned}$$

Where Equation 121 follows from the expression obtained using De Morgan's laws for $G_i^c$ using the definition of $G_i$ in Equation 116. □

BOUND SAMPLES FOR THE CASE $i < a^*$

**Lemma E.2** (Bound on the number of samples of a low-cost unsatisfactory arm under Pairwise-Elimination)**.** *When the maximum round deviation $\kappa = 0$, the expected number of samples of a low-cost arm with index $i < a^*$ over horizon $T$ is upper bounded by,*

$$\mathbb{E}\left[n_i\left(T\right)\right] < 1 + \frac{32 \log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2} + \frac{43}{\Delta_{Q,i}^2}.$$

*Proof.* The expected number of pulls $\mathbb{E}\left[n_i\left(T\right)\right]$ are bound by using the Iterated Expectation Lemma C.3 and conditioning on the event collection $G_i, B_{1,i}, B_i$ which are mutually exclusive and exhaustive per Lemma E.1.

$$\begin{aligned}
\mathbb{E}\left[n_i\left(T\right)\right] &= \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] \cdot \Pr\left(G_i\right) + \mathbb{E}\left[n_i\left(T\right) \mid B_{1,i}\right] \cdot \Pr\left(B_{1,i}\right)\\
&\quad + \mathbb{E}\left[n_i\left(T\right) \mid B_i\right] \cdot \Pr\left(B_i\right) && (122)\\
&\leq \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \Pr\left(B_i\right) && (123)\\
&= \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \Pr\left(G_{1,i} \cap \left(G_{2,i}^c \cup G_{3,i}^c\right) \cap E_i^c\right) \quad (B_i \text{ from Lemma E.1})\\
& && (124)\\
&\leq \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \Pr\left(G_{1,i} \cap \left(G_{2,i}^c \cup G_{3,i}^c\right)\right) && (125)\\
&= \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \Pr\left(\left(G_{1,i} \cap G_{2,i}^c\right) \cup \left(G_{1,i} \cap G_{3,i}^c\right)\right) \text{ (distributivity of } \cap) && (126)\\
&\leq \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \Pr\left(B_{2,i} \cup B_{3,i}\right) \quad \text{(simplifying notation)} && (127)\\
&= \mathbb{E}\left[n_i\left(T\right) \mid G_i\right] + T \cdot \left(\Pr\left(B_{2,i}\right) + \Pr\left(B_{3,i}\right)\right) \quad \text{(using the union bound).} && (128)
\end{aligned}$$

Where Equation 123 follows from the fact that $n_i(T) \mid B_{1,i} = 0$ since there can be no samples of arm $i$ if episode $i$ never occurs, and we define $G_{1,i} \cap G_{2,i}^c$ and $G_{1,i} \cap G_{3,i}^c$ as $B_{2,i}$ and $B_{3,i}$ respectively for notational simplicity.

First we bound the number of samples of arm $i$ conditioned on the good event $G_i$. Since during episode $i < a^*$, we either make the correct decision of eliminating arm $i$ by episode $\omega_i = \rho_i$ as captured by $G_{2,i} \cap G_{3,i}$. Alternatively, under $G_i$ we run out of samples as captured by $E_i$. In either case we will not have more than $\tau_{\rho_i}$ samples of arm $i$, where $\tau_{\rho_i}$ is given by,

$$\tau_{\rho_i} = \left\lceil \frac{2\log(T\tilde{\Delta}_{\rho_i}^2)}{\tilde{\Delta}_{\rho_i}^2} \right\rceil. \tag{129}$$

By construction of the round $\rho_i$, for all $i \leq a^*$, we have,

$$\frac{|\Delta_{Q,i}|}{4} \leq \tilde{\Delta}_{\rho_i} < \frac{|\Delta_{Q,i}|}{2}. \tag{130}$$

Plugging in the bounds in Equation 130,

$$\tau_{\rho_i} \leq \left\lceil \frac{32\log\left(\frac{T\Delta_{Q,i}^2}{4}\right)}{\Delta_{Q,i}^2} \right\rceil \tag{131}$$

$$< 1 + \frac{32\log(T\Delta_{Q,i}^2)}{\Delta_{Q,i}^2}. \tag{132}$$

Therefore, we have,

$$\mathbb{E}[n_i(T) \mid G_i] \leq \tau_{\rho_i} < 1 + \frac{32\log(T\Delta_{Q,i}^2)}{\Delta_{Q,i}^2}. \tag{133}$$

Next we bound $\Pr(B_{2,i})$ and $\Pr(B_{3,i})$ in order. Since $B_{2,i} = G_{1,i} \cap G_{2,i}^c$, from the specification of $G_{2,i}^c$ and the fact that intersecting with $G_{1,i}$ puts us in the sub-space of $\Omega$ where episode $i$ occurs, $B_{2,i} \forall\, i < a^*$ is the event: "Arm $i$ is **not** eliminated by arm $\ell$ by when round $\omega_i = \rho_i$, during episode $i$". Similarly $B_{3,i} \forall\, i < a^*$ is the event that "Arm $\ell$ **is eliminated** by arm $i$ by when round $\omega_i = \rho_i - 1$, during episode $i$"

Along the lines of the proof composition in Auer & Ortner (2010) for the Improved UCB algorithm, we construct three clauses on the empirical returns of arms $i$ and $\ell$ in Equations 134, 135, and 136.

$$\hat{\mu}_i \leq \mu_i + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}} \tag{134}$$

$$\hat{\mu}_\ell \geq \mu_\ell - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}} \tag{135}$$

$$\hat{\mu}_\ell \geq \mu_\ell - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_\ell}^2\right)}{2\tau_{\omega_\ell}}}. \tag{136}$$

Clauses 134 and 135 holding when $\omega_i = \rho_i$ lead to the elimination of arm $i$ by arm $\ell$ as is shown in the work that follows,

$$\sqrt{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)/2\tau_{\rho_i}} \leq \tilde{\Delta}_{\rho_i}/2 < \Delta_{Q,i}/4. \tag{137}$$

Therefore,

$$\hat{\mu}_i + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)}{2\tau_{\rho_i}}} \leq \mu_i + 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)}{2\tau_{\rho_i}}} \text{ (From clause 134, and } \omega_i = \rho_i \text{ )} \tag{138}$$

$$< \mu_i + \Delta_{Q,i} - 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)}{2\tau_{\rho_i}}} \text{ (From the ordering 137)} \tag{139}$$

$$= (1-\alpha)\mu_\ell - 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)}{2\tau_{\rho_i}}} \tag{140}$$

$$\leq (1-\alpha)\hat{\mu}_\ell - (1-\alpha)\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_i}^2\right)}{2\tau_{\rho_i}}} \text{ (From clause 135, and } \omega_i = \rho_i \text{ )} \tag{141}$$

$$\leq (1-\alpha)\hat{\mu}_\ell - (1-\alpha)\sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_\ell}^2\right)}{2\tau_{\omega_\ell}}} \text{ (Since } \omega_\ell \geq \omega_i). \tag{142}$$

Here, Equation 142 is the criteria for arm $i$ being eliminated by arm $\ell$ in PE. We upper bound the probability of the arm $i$ not being eliminated by union bounding the probability of the complements of the Clauses 134 and 135 using Lemma C.2 (Hoeffding's Inequality). In addition, we include the bound on the complement of Clause 136 which shall be useful later in bounding $\Pr(B_{3,i})$.

$$\Pr\left(\hat{\mu}_i > \mu_i + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}}\right) \leq \frac{1}{T\tilde{\Delta}_{\omega_i}^2} \tag{143}$$

$$\Pr\left(\hat{\mu}_\ell < \mu_\ell - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}}\right) \leq \exp\left(-\frac{\tau_{\omega_\ell}}{\tau_{\omega_i}}\log\left(T\tilde{\Delta}_{\omega_i}^2\right)\right) \leq \frac{1}{T\tilde{\Delta}_{\omega_i}^2} \text{ (Since } \tau_{\omega_\ell} \geq \tau_{\omega_i}) \tag{144}$$

$$\Pr\left(\hat{\mu}_\ell < \mu_\ell - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_\ell}^2\right)}{2\tau_{\omega_\ell}}}\right) \leq \frac{1}{T\tilde{\Delta}_{\omega_\ell}^2} \leq \frac{4^\kappa}{T\tilde{\Delta}_{\omega_i}^2} \text{ (Since } \omega_\ell \leq \omega_i + \kappa, \text{ and } \tilde{\Delta}_m = 2^{-m}). \tag{145}$$

If either of the two clauses 134 or 135 are violated, then elimination of arm $i$ will not occur. Therefore, we can bound,

$$\Pr(B_{2,i}) \leq \Pr\left(\hat{\mu}_i > \mu_i + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}}\right) + \Pr\left(\hat{\mu}_\ell < \mu_\ell - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_i}^2\right)}{2\tau_{\omega_i}}}\right) \tag{146}$$

$$\leq \frac{2}{T\tilde{\Delta}_{\omega_i}^2}. \tag{147}$$

Plugging in round number $\omega_i = \rho_i$ in Equation 147, and then plugging in the lower bound on $\tilde{\Delta}_{\rho_i}$ from Ordering 130, we have,

$$\Pr(B_{2,i}) \leq \frac{2}{T\tilde{\Delta}_{\rho_i}^2} \leq \frac{32}{T\Delta_{Q,i}^2}. \tag{148}$$

Finally, we wish to bound $\Pr(B_{3,i})$ . Say that the actual elimination of arm $\ell$ by arm $i$ occurs in some round $\omega_i = \rho < \rho_i$. To bound the probability of this clause of the event $G_i^c$, we note that the

clauses in Equations 134 and 136 holding simultaneously preclude arm $\ell$ from being removed by arm $i$ regardless of the round number $\rho$ in question. Therefore, using the results in Equations 143 and 145, the probability of a round $\rho$, where $\ell$ is removed, existing, can be found by plugging in $\omega_i = \rho$, and is upper bounded by $\frac{4^\kappa+1}{T\tilde{\Delta}^2_\rho}$[1]. While there is no definitive round number associated with $\rho$, from the clause itself we know that we must have $\rho < \rho_i$.

$$\Pr\left(B_{3,i}\right) \leq \sum_{\rho=0}^{\rho_i-1} \frac{4^\kappa+1}{T\tilde{\Delta}^2_\rho} = \sum_{\rho=0}^{\rho_i-1} \frac{(4^\kappa+1)\cdot 4^\rho}{T} \text{ (Using } \tilde{\Delta}_m = 2^{-m}) \tag{149}$$

$$< \frac{4^\kappa+1}{3T} \cdot 4^{\rho_i} \text{ (Using the formula for the sum of a Geometric Series)} \tag{150}$$

$$= \frac{4^\kappa+1}{3T\tilde{\Delta}^2_{\rho_i}} \text{ (Since } \tilde{\Delta}_m = 2^{-m}) \tag{151}$$

$$= \frac{16\left(4^\kappa+1\right)}{3T\Delta^2_{Q,i}} \text{ (Because } \tilde{\Delta}_{\rho_i} \geq \frac{\Delta_{Q,i}}{4}) \tag{152}$$

$$< \frac{11}{T\Delta^2_{Q,i}} \text{ (Since we impose in Lemma E.2).} \tag{153}$$

Plugging in the bounds in Expressions 133, 148 and 153 into Equation 200 we get the overall bound on the number of samples stated in Lemma E.2. □

### BOUNDING SAMPLES OF REFERENCE ARM $\ell$

**Lemma E.3** (Bound on Samples of Reference arm $\ell$ under Pairwise-Elimination)**.** *Samples of reference arm $\ell$ emanate from the episodes of candidate arms being compared to arm $\ell$. When $\kappa = 0$, we show the bound,*

$$\mathbb{E}\left[n_\ell(T)\right] < 1 + \max_{i \leq a^*} \frac{32\log\left(T\Delta^2_{Q,i}\right)}{\Delta^2_{Q,i}} + \frac{43}{\Delta^2_{Q,a^*}} + \sum_{i=1}^{a^*} \frac{43}{\Delta^2_{Q,i}}.$$

*Proof.* Under Pairwise-Elimination the nominal outcome is for episodes $i = 1, \ldots, a^* - 1$ to result in the candidate arm $i$ being eliminated by arm $\ell$, followed arm $a^*$ eliminating arm $\ell$ during episode $a^*$. To prove Lemma E.3 we condition on this nominal sequence and upper bound the probability of the outcome deviating from this sequence by a factor proportional to $\frac{1}{T}$. Throughout the episodes $i = 1, \ldots a^*$ the number of samples $n_\ell\left(T\right)$ are equal to the number of samples of the most sampled candidate arm $i \leq a^*$. This is because in PE we re-use samples of arm $\ell$ across episodes and only further sample $\ell$ to keep up with the samples of a candidate arm. Motivated by this fact about $n_\ell$ we begin by defining a compound high-probability good event $G$.

The compound good event $G$ is the event that for each episode $i \leq a^*$ that was executed, the episode satisfied the episode-wise good event $G_i$. Let the random variable $Z$ denote the final episode in the run of PE. Then $\{Z = z\}$ constitutes a measurable event in the sample space $\Omega$. Mathematically we define,

$$G = \bigcup_{z=1}^{a^*} \left( \{Z = z\} \cap \bigcap_{i=1}^{z} G_i \right). \tag{154}$$

[1]Because for any events $A, B$, and $C$, $\Pr\left(A \cap B\right) \leq \Pr\left(C^c\right) \implies \Pr\left(C\right) \leq \Pr\left(A^c \cup B^c\right)$.

The complement of $G$, namely $G^c$ can be written out using the definition of $G$ and De Morgan's laws as,

$$G^c = \bigcap_{z=1}^{a^*} \left( \{Z = z\}^c \cup \bigcup_{i=1}^{z} G_i^c \right) \tag{155}$$

$$\subseteq \bigcup_{i=1}^{a^*} G_i^c \cup \{Z \neq a^*\} \quad \text{(picking } z = a^* \text{ from the iterated intersection)} \tag{156}$$

$$= \bigcup_{i=1}^{a^*} G_i^c \cup \{Z < a^*\} \cup \{Z > a^*\} \tag{157}$$

$$= \bigcup_{i=1}^{a^*} (B_{1,i} \cup B_i) \cup \bigcup_{i=1}^{a^*} B_{1,i} \cup \{Z > a^*\} \tag{158}$$

$$= \bigcup_{i=1}^{a^*} (B_{1,i} \cup B_i) \cup \{Z > a^*\} \tag{159}$$

$$= \bigcup_{i=1}^{a^*} G_i^c \cup \{Z > a^*\}. \tag{160}$$

Where the first term in Equation 158 follows from the equivalence between $G_i^c$ and $B_{1,i} \cup B_i$ shown in the proof of Lemma E.1. The second term in Equation 158 is based on the equivalence between the event $\{Z < a^*\}$, meaning that the final episode precedes $a^*$, and the event $\bigcup_{i=1}^{a^*} B_{1,i}$ which means that some episode $i = 1, \ldots, a^*$ was not executed.

We leverage event $G$ to bound the expected number of arm $\ell$,

$$\mathbb{E}\left[n_\ell(T)\right] = \mathbb{E}\left[n_\ell(T) \mid G\right] \Pr\left(G\right) + \mathbb{E}\left[n_\ell(T) \mid G^c\right] \Pr\left(G^c\right) \tag{161}$$

$$\leq \mathbb{E}\left[n_\ell(T) \mid G\right] + T \cdot \Pr\left(G^c\right). \tag{162}$$

Since samples of arm $\ell$ are reused between episodes with further sampling of arm $\ell$ only occurring to match the demand from a higher round number, we have,

$$\mathbb{E}\left[n_\ell\left(T\right) \mid G\right] = \mathbb{E}\left[\left(\max_{i \leq Z} n_i\left(1, t_Z\right)\right) \mid G\right] \tag{163}$$

$$\leq \mathbb{E}\left[\max_{i \leq a^*} n_i\left(1, t_{a^*}\right) \mid G\right] \quad (\because Z \mid G \leq a^*) \tag{164}$$

$$\leq \max_{i \leq a^*} \tau_{\rho_i} \quad \text{(using Equation 117, and by construction of } G \text{ as } \cap G_i) \tag{165}$$

$$= \max_{i \leq a^*} \left\lceil \frac{2 \log\left(T \tilde{\Delta}_{\rho_i}^2\right)}{\tilde{\Delta}_{\rho_i}^2} \right\rceil \tag{166}$$

$$< 1 + \max_{i \leq a^*} \frac{2 \log\left(T \tilde{\Delta}_{\rho_i}^2\right)}{\tilde{\Delta}_{\rho_i}^2} \tag{167}$$

$$\leq 1 + \max_{i \leq a^*} \frac{32 \log\left(T \Delta_{Q,i}^2/4\right)}{\Delta_{Q,i}^2} \text{ (Using the ordering in Relation 130)} \tag{168}$$

$$< 1 + \max_{i \leq a^*} \frac{32 \log\left(T \Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}. \tag{169}$$

To complete the bound $\mathbb{E}\left[n_\ell(T)\right]$, we develop a bound on the term $\Pr\left(G^c\right)$ in Lemma E.6. However, to prove Lemma E.6 we first require two intermediate results in the form of Lemmas E.4 and E.5.

**Lemma E.4.** *In all the outcomes contained in $G^c$ ending in some episode $Z = i \leq a^*$, some $B_j, j \leq i$ must have held.*

$$G^c \cap \{Z = i\} \subseteq \bigcup_{j=1}^{i} B_j \quad \forall\, i \leq a^*. \tag{170}$$

*Proof.*

$$\begin{aligned}
G^c \cap \{Z = i\} &= ((G_i^c \cup G_i) \cap \{Z = i\}) \cap G^c && (171)\\
&= ((G_i^c \cap \{Z = i\}) \cup (G_i \cap \{Z = i\})) \cap G^c \quad (\cap \text{ distributes over } \cup) && (172)\\
&= (G_i^c \cap \{Z = i\}) \cup (G_i \cap \{Z = i\} \cap G^c) \quad (\because G_i^c \cap \{Z = i\} \subseteq G^c) && (173)\\
&= ((B_{1,i} \cup B_i) \cap \{Z = i\}) \cup (G_i \cap \{Z = i\} \cap G^c) \quad (\because G_i^c = B_{1,i} \cup B_i) && (174)\\
&\subseteq B_i \cup (G_i \cap \{Z = i\} \cap G^c) \quad (\because B_{1,i} \cap \{Z = i\} = \phi). && (175)
\end{aligned}$$

Now consider just the event $G_i \cap \{Z = i\} \cap G^c$ from Equation 175. We can find an event it is subsumed within in the following way,

$$\begin{aligned}
G_i \cap \{Z = i\} \cap G^c &= G_i \cap \left( \bigcap_{j=1}^{i-1} G_j \cup \left( \bigcap_{j=1}^{i-1} G_j \right)^c \right) \cap \{Z = i\} \cap G^c && (176)\\
&= \left( \bigcap_{j=1}^{i} G_j \cap \{Z = i\} \cap G^c \right) \cup \left( G_i \cap \bigcup_{j=1}^{i-1} G_j^c \cap \{Z = i\} \cap G^c \right) && (177)\\
&= G_i \cap \bigcup_{j=1}^{i-1} \left( G_j^c \cap \{Z = i\} \right) \cap G^c \quad (\bigcap_{j=1}^{i} G_j \cap \{Z = i\} \subseteq G, \forall\, i \leq a^*) && (178)\\
&= G_i \cap \bigcup_{j=1}^{i-1} ((B_{1,j} \cup B_j) \cap \{Z = i\}) \cap G^c \quad (\text{from Lemma E.1}) && (179)\\
&= G_i \cap \bigcup_{j=1}^{i-1} (B_j \cap \{Z = i\}) \cap G^c \quad (\because B_{1,j} \cap \{Z = i\} = \phi, \forall\, j \leq i) && (180)\\
&\subseteq \bigcup_{j=1}^{i-1} B_j. && (181)
\end{aligned}$$

Plugging in Equation 181 into Equation 175 gives us the result stated in Lemma E.4. □

**Lemma E.5.** *Within the space of events that constitute $G^c$, if an episode $i \leq a^*$ does not occur, then there exists $j < i$ such that the event $B_j$ occurred. Mathematically,*

$$B_{1,i} \cap G^c \subseteq \bigcup_{j=1}^{i-1} B_j \,,\, \forall\, i \leq a^*. \tag{182}$$

*Proof.* We can prove the result by induction. First note that $B_{1,1} = \phi$ since Episode 1 always occurs. Let $i = 2$ represent the base case. Then,

$$\begin{aligned}
B_{1,2} \cap G^c &= \{Z = 1\} \cap G^c && (183)\\
&\subseteq B_1 \quad \text{(Using Lemma E.4).} && (184)
\end{aligned}$$

Now say that the statement in Lemma E.5 holds true for some $i = k < a^*$. That is,

$$B_{1,k} \cap G^c \subseteq \bigcup_{j=1}^{k-1} B_j. \tag{185}$$

To prove Lemma E.5 we need to show this result for $i = k + 1$.

$$B_{1,k+1} \cap G^c = (B_{1,k} \cup \{Z = k\}) \cap G^c \tag{186}$$
$$= (B_{1,k} \cap G^c) \cup (\{Z = k\} \cap G^c) \quad \cap \text{ over } \cup \tag{187}$$
$$\subseteq \bigcup_{j=1}^{k-1} B_j \cup \bigcup_{j=1}^{k} B_j \quad \text{(using the induction hypothesis and Lemma E.4).} \tag{188}$$
$$= \bigcup_{j=1}^{k} B_j. \tag{189}$$

Where Equation 186 uses the identity $B_{1,k+1} = B_{1,k} \cup \{Z = k\}$ which breaks down the event of episode $k + 1$ not occurring into the event that episode $k$ did not occur, or the event that episode $k$ was the final episode $Z$. Equation 189 is the required result from the statement of Lemma E.5. □

**Lemma E.6** (Bound on $\Pr(G^c)$)**.** *We can upper bound the probability of the compound good event $G$ not occurring as,*

$$\Pr(G^c) \leq \sum_{i=1}^{a^*} \Pr(B_j) + \Pr(B_{a^*}) \tag{190}$$
$$\leq \sum_{i=1}^{a^*-1} \frac{43}{\Delta_{Q,i}^2} + 2\Pr(B_{a^*}) \quad \textit{(bound on } \Pr(B_i) \textit{ shown in the proof of Lemma E.2).} \tag{191}$$

*Where $B_i$, $\forall\, i \leq a^*$ is as defined in Lemma E.1.*

*Proof.* We begin by introducing the expanded expression for $G^c$ developed in Equation 160.

$$\Pr(G^c) = \Pr(G^c \cap G^c) \tag{192}$$
$$\leq \Pr\left(\left(\bigcup_{i=1}^{a^*} G_i^c \cup \{Z > a^*\}\right) \cap G^c\right) \quad \text{(from Equation 160)} \tag{193}$$
$$\leq \Pr\left(\left(\bigcup_{i=1}^{a^*} G_i^c \cap G^c\right) \cup (\{Z > a^*\} \cap G^c)\right) \quad (\cap \text{ distributes over } \cup) \tag{194}$$
$$\leq \Pr\left(\bigcup_{i=1}^{a^*} (G_i^c \cap G^c)\right) + \Pr(Z > a^*) \quad \text{(union bound and } \Pr(A, B) \leq \Pr(A)) \tag{195}$$
$$= \Pr\left(\bigcup_{i=1}^{a^*} ((B_{1,i} \cup B_i) \cap G^c)\right) + \Pr(Z > a^*) \quad \text{(by definition of } G_i) \tag{196}$$
$$= \Pr\left(\bigcup_{i=1}^{a^*} ((B_{1,i} \cap G^c) \cup (B_i \cap G^c))\right) + \Pr(Z > a^*) \quad (\cap \text{ distributes over } \cup) \tag{197}$$
$$\leq \Pr\left(\bigcup_{i=1}^{a^*} \left(B_i \cup \bigcup_{j=1}^{i-1} B_j\right)\right) + \Pr(Z > a^*) \quad \text{(using Lemma E.5)} \tag{198}$$
$$= \Pr\left(\bigcup_{i=1}^{a^*} B_i\right) + \Pr(Z > a^*) \tag{199}$$
$$\leq \sum_{i=1}^{a^*} \Pr(B_j) + \Pr(B_{a^*}) \quad \text{(Union bound, and } \{Z > a^*\} \implies B_{a^*}\text{).} \tag{200}$$

□

To complete the upper bound on the expected number of samples $\mathbb{E}\left[n_\ell\left(T\right)\right]$ from Equation 162, we upper bound $\Pr\left(B_{a^*}\right)$ along the same lines as $\Pr\left(B_i\right), i < a^*$ in the proof of Lemma E.2. The key difference being that the roles of candidate arm $a^*$ and reference arm $\ell$ are reversed as compared to the earlier procedure since $\mu_{a^*} \geq \mu_\ell$. Moreover, just like the earlier proof we can define $B_{2,a^*} = G_{1,a^*} \cap G_{2,i}^c$ and $B_{3,a^*} = G_{1,a^*} \cap G_{3,i}^c$. In words $B_{2,a^*}$ is the event that "Arm $\ell$ is **not** eliminated by arm $a^*$ by when $\omega_{a^*} = \rho_{a^*}$, during episode $a^*$". Similarly, $B_{3,a^*}$ is the event "Arm $a^*$ **is eliminated** by arm $\ell$ by when round $\omega_{a^*} = \rho_{a^*} - 1$, during episode $a^*$".

We bound $\Pr\left(B_{2,a^*}\right)$ and $\Pr\left(B_{3,a^*}\right)$ separately by constructing clauses on $\hat{\mu}_\ell$, and $\hat{\mu}_{a^*}$ as before.

$$\hat{\mu}_\ell \leq \mu_\ell + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_{a^*}}^2\right)}{2\tau_{\omega_{a^*}}}} \tag{201}$$

$$\hat{\mu}_\ell \leq \mu_\ell + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_\ell}^2\right)}{2\tau_{\omega_\ell}}} \tag{202}$$

$$\hat{\mu}_{a^*} \geq \mu_{a^*} - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_{a^*}}^2\right)}{2\tau_{\omega_{a^*}}}}. \tag{203}$$

Clauses 201 and 203 holding when $\omega_{a^*} = \rho_{a^*}$ lead to the elimination of arm $\ell$ by arm $a^*$ as is shown in the following steps,

$$\sqrt{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)/2\tau_{\rho_{a^*}}} < \tilde{\Delta}_{\rho_{a^*}}/2 < |\Delta_{Q,a^*}|/4. \tag{204}$$

Therefore using $\omega_{a^*} = \rho_{a^*}$ ,

$$(1-\alpha)\left(\hat{\mu}_\ell + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\omega_\ell}^2\right)}{2\tau_{\omega_\ell}}}\right) \leq (1-\alpha)\left(\hat{\mu}_\ell + \sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}}\right) \tag{205}$$

$$\leq (1-\alpha)\left(\mu_\ell + 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}}\right) \text{ (Equation 201)} \tag{206}$$

$$\leq (1-\alpha)\mu_\ell + 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}} \tag{207}$$

$$< (1-\alpha)\mu_\ell + |\Delta_{Q,a^*}| - 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}} \text{ (From 204)} \tag{208}$$

$$= \mu_{a^*} - 2\sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}} \tag{209}$$

$$\leq \hat{\mu}_{a^*} - \sqrt{\frac{\log\left(T\tilde{\Delta}_{\rho_{a^*}}^2\right)}{2\tau_{\rho_{a^*}}}} \text{ (From Equation 203) .} \tag{210}$$

Here Equation 210 is the criteria for arm $\ell$ being eliminated by arm $a^*$ in PE. Since Clause 201 and Clause 203 being true and applicable at round $\omega_i = a^*$ imply arm $\ell$ being eliminated, we can upper

bound $\Pr\left(B_{2,a^*}\right)$ as,

$$\Pr\left(B_{2,a^*}\right) \leq \Pr\left(\hat{\mu}_\ell > \mu_\ell + \sqrt{\frac{\log\left(T\tilde{\Delta}^2_{\omega_{a^*}}\right)}{2\tau_{\omega_{a^*}}}}\right) + \Pr\left(\hat{\mu}_{a^*} < \mu_{a^*} - \sqrt{\frac{\log\left(T\tilde{\Delta}^2_{\omega_{a^*}}\right)}{2\tau_{\omega_{a^*}}}}\right) \tag{211}$$

$$\leq \frac{2}{T\tilde{\Delta}^2_{\omega_{a^*}}} \text{ (Similar to the bounds in 143 and 144 ).} \tag{212}$$

Plugging in round number $\omega_{a^*} = \rho_{a^*}$ in Equation 212, and then plugging in the lower bound on $\tilde{\Delta}_{\rho_{a^*}}$ from Ordering 130, we have,

$$\Pr\left(B_{2,a^*}\right) \leq \frac{2}{T\tilde{\Delta}^2_{\rho_{a^*}}} \leq \frac{32}{T\Delta^2_{Q,a^*}}. \tag{213}$$

To complete the bound on $\Pr\left(B_{a^*}\right)$ we must bound $\Pr\left(B_{3,a^*}\right)$ which is the the probability of arm $a^*$ being eliminated by arm $\ell$ by round $\omega_{a^*} = \rho_{a^*}$. Similar to the arguments in the Proof of Lemma E.2, the clauses in Equations 201 and 203 holding simultaneously preclude arm $a^*$ from being removed by arm $\ell$ regardless of the round number, and we shall have,

$$\Pr\left(B_{3,a^*}\right) < \frac{16\left(4^\kappa + 1\right)}{3T\Delta^2_{Q,a^*}} \tag{214}$$

$$< \frac{11}{T\Delta^2_{Q,a^*}} \text{ (When we impose } \kappa = 0). \tag{215}$$

Using steps identical to those that lead up to Equation 128 we shall have,

$$\Pr\left(B_{a^*}\right) \leq \Pr\left(B_{2,a^*}\right) + \Pr\left(B_{3,a^*}\right) \tag{216}$$

$$< \frac{43}{T\Delta^2_{Q,a^*}} \quad \text{(from upper bounds in Equations 213 and 215).} \tag{217}$$

Combining the upper bound on the expected number of samples of arm $\ell$ in Equation 169, with the bound on $\Pr\left(G^c\right)$ in Lemma E.6, and the bound on $\Pr\left(B_{a^*}\right)$ in Equation 217 we reach the upper bound stated in Lemma E.3. □

### BOUNDING SAMPLES FOR HIGH COST ARMS $a^* < i < \ell$

**Lemma E.7** (Bound on the expected number of samples of all high cost arms under Pairwise-Elimination)**.** *When $\kappa = 0$ the expected number of samples of all the higher cost, non-reference arms, that is, arms with index in the range $a^* < i < \ell$ is upper bounded by a constant given by,*

$$\sum_{i=a^*+1}^{\ell} \mathbb{E}\left[n_i(T)\right] < \frac{43}{\Delta^2_{Q,a^*}}.$$

*Proof.* We can upper bound the expected number of samples $\sum_{a^*<i<\ell} \mathbb{E}\left[n_i\left(T\right)\right]$ by conditioning on the event $\{Z \leq a^*\}$ and its complement.

$$\sum_{i=a^*+1}^{\ell} \mathbb{E}\left[n_i\left(T\right)\right] = \mathbb{E}\left[\sum_{i=a^*+1}^{\ell} n_i\left(T\right)\right] \quad \text{(from the linearity of Expectation operator)} \tag{218}$$

$$= \mathbb{E}\left[\sum_{i=a^*+1}^{\ell} n_i\left(T\right) \mid \{Z \leq a^*\}\right] \cdot \Pr\left(\{Z \leq a^*\}\right)$$

$$+ \mathbb{E}\left[\sum_{i=a^*+1}^{\ell} n_i\left(T\right) \mid \{Z > a^*\}\right] \cdot \Pr\left(\{Z > a^*\}\right) \quad \text{(from Lemma C.3)} \tag{219}$$

$$\leq T \cdot \Pr\left(B_{a^*}\right) \quad \left(\because \{Z > a^*\} \implies B_{a^*}\right) \tag{220}$$

$$< \frac{43}{\Delta^2_{Q,a^*}} \quad \text{(from Equation 217).} \tag{221}$$

Where $\mathbb{E}\left[\sum_{i=a^*+1}^{\ell} n_i(T) \mid \{Z \leq a^*\}\right] = 0$ follows from the fact that there cannot be any samples of an arm $i > a^*$ in the case when the final episode $Z$ is less than $a^*$. □

Finally, we are in a position to Prove Theorem 3.1.

*Proof of Theorem 3.1.* First, we apply Regret decomposition in Lemma C.5 to Cost Regret,

$$\mathbb{E}\left[\text{Cost\_Reg}(T,\nu)\right] \leq \sum_{i=a^*+1}^{\ell} \Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\right] \quad \text{(because } \Delta_{C,i} \leq 0 \text{ for } i \leq a^*\text{)} \tag{222}$$

$$= \Delta_{C,\ell}\,\mathbb{E}\left[n_\ell\left(T\right)\right] + \sum_{i=a^*+1}^{\ell-1} \Delta_{C,i}\,\mathbb{E}\left[n_i\left(T\right)\right] \tag{223}$$

$$< \left(1 + \max_{i \leq a^*} \frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2} + \sum_{i=1}^{a^*} \frac{43}{\Delta_{Q,i}^2}\right)\Delta_{C,\ell} + \frac{43}{\Delta_{Q,a^*}^2}\left(\max_{a^*<i<\ell} \Delta_{C,i} + \Delta_{C,\ell}\right). \tag{224}$$

Similarly for Quality Regret we have,

$$\mathbb{E}\left[\text{Quality\_Reg}(T,\nu)\right] = \sum_{i=1}^{\ell} \Delta_{Q,i}^{+}\mathbb{E}\left[n_i\left(T\right)\right] \tag{225}$$

$$= \sum_{i=1}^{a^*-1} \Delta_{Q,i}\mathbb{E}\left[n_i\left(T\right)\right] + \sum_{i=a^*+1}^{\ell-1} \Delta_{Q,i}^{+}\mathbb{E}\left[n_i\left(T\right)\right] \tag{226}$$

$$\leq \sum_{i=1}^{a^*-1} \Delta_{Q,i}\mathbb{E}\left[n_i\left(T\right)\right] + \max_{a^*<i<\ell} \Delta_{Q,i}^{+} \sum_{i=a^*+1}^{\ell-1} \mathbb{E}\left[n_i\left(T\right)\right] \tag{227}$$

$$< \sum_{i=1}^{a^*-1}\left(\Delta_{Q,i} + \frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}} + \frac{43}{\Delta_{Q,i}}\right) + \frac{43}{\Delta_{Q,a^*}^2}\max_{a^*<i<\ell} \Delta_{Q,i}^{+}. \tag{228}$$

Which are the bounds stated in Theorem 3.1. □

For an improved understanding of these upper bounds, we provide a description of the terms.

First for Cost Regret,

$$\underbrace{\left(1 + \max_{i \leq a^*} \frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}\right)\Delta_{C,\ell}}_{\substack{\text{Contribution from } \ell \text{ under nominal} \\ \text{termination in PE episode } a^*}} + \underbrace{\left(\sum_{i=1}^{a^*} \frac{43}{\Delta_{Q,i}^2}\right)\Delta_{C,\ell}}_{\substack{\text{Contribution from } \ell \text{ under} \\ \text{mis-termination in PE episode} \leq a^*}} + \underbrace{\frac{43}{\Delta_{Q,a^*}^2}\max_{i>a^*}\Delta_{C,i}}_{\substack{\text{Contribution from episodes} > a^* \\ \text{in case of mis-termination during ep } a^*}}.$$

Next for Quality Regret,

$$\underbrace{\sum_{i=1}^{a^*-1}\left(\Delta_{Q,i} + \frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}\right)}_{\substack{\text{Contribution from } i < a^* \\ \text{under nominal termination in PE episode } a^*}} + \underbrace{\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}}}_{\substack{\text{Contribution from } i < a^* \text{ under} \\ \text{mis-termination in PE episode} \leq a^*}} + \underbrace{\frac{43}{\Delta_{Q,a^*}^2}\max_{i>a^*}\Delta_{Q,i}^{+}}_{\substack{\text{Contribution from episodes} > a^* \\ \text{in case of mis-termination during ep } a^*}}.$$

# F ANALYSIS FOR PE-CS IN THE FULL COST-SUBSIDY SETTING

We now turn towards proving Theorem 3.2 that establishes an upper bound on Expected Cumulative Cost and Quality regret for PE-CS. The PE-CS algorithm operates in the unknown reference arm setting, which we also refer to from time-to-time to be the Full Cost-Subsidy Setting Sinha et al. (2021). We have already shown upper bounds on Cost and Quality regret for Pairwise-Elimination (PE) for its operation in the known reference arm setting. The principle hurdle in generalizing the PE analysis to an analysis for PE-CS is working through the uncertainty associated with the identification of the best arm in the BAI stage of PE-CS. To perform the PE-CS analysis, not only do we need the definitions, notation, and setup from the analysis of PE, but also we require some additional constructs spelled out in the following Section.

## F.1 PE-CS AND UNKNOWN REFERENCE ARM SETTING SPECIFIC DEFINITIONS

As described in Algorithm 2, PE-CS operates in two phases, a Best-Arm-Identification (BAI) phase and a Pairwise Elimination (PE) phase. As discussed in Section 3, the BAI phase of PE-CS is the Improved UCB algorithm Auer & Ortner (2010) terminated once there is a single active arm remaining. The single remaining arm is assigned to be the empirical reference arm denoted by $\ell$. As we show in the subsequent work, by the manner in which PE-CS is setup, the event that the identified reference arm gets arm $i^*$ in the line $\ell \leftarrow \mathcal{A}[0]$ (Line 6, Algorithm 2) is a high probability event. The core idea behind the analysis is to condition the expected number of samples on this desirable and likely outcome occurring during the BAI stage.

In addition to the notation defined in the main paper, we define more constructs that are specific to the analysis of PE-CS. For arm $i$, define round number $\sigma_i$ within the Best-Arm-Identification (BAI) phase of PE-CS to be,

$$\sigma_i = \min\left\{m \,|\, \tilde{\Delta}_m < \frac{\Delta_i}{2}\right\}. \tag{229}$$

Intuitively, round $\sigma_i$ is the round number by which we expect arm $i$ to be eliminated by arm $i^*$ in the BAI stage of PE-CS.

We let the final round during which arm $i$ was sampled during the BAI stage be denoted by the random variable $\Sigma_i$. To apportion the contributions of the BAI stage and the PE stage to the total number of samples $n_i(T)$, we introduce the variable $t_{\text{BAI}}$ to denote the final time-step $t$ in the BAI stage. As a consequence of these two definitions $n_i(1, t_{\text{BAI}}) \leq \tau_{\Sigma_i}$. The highest round number reached during the BAI stage overall for any and all active arms,

$$\Sigma_f = \max_{i \neq i^*} \Sigma_i. \tag{230}$$

And to denote the set of active arms at the last time-step of the BAI stage we use,

$$\mathcal{A}_f = \mathcal{A}(t_{\text{BAI}}). \tag{231}$$

Next, we define a collection of useful events in Table 3 for the analysis that follows. These events are contingent on outcomes occurring during the BAI stage alone.

Table 3: Probabilistic Events Descriptions and Symbols for PE-CS Analysis

| Symbol | Event Description |
|---|---|
| $\Gamma_i,\ \forall\, i \neq i^*$ | Arm $i$ is eliminated by when round $m = \sigma_i$ during the BAI-stage |
| $\beta_{1,i},\ \forall\, i \neq i^*$ | Arm $i$ is not eliminated by when round $m = \sigma_i$ and Arm $i^* \in \mathcal{A}_{\sigma_i}$ during the BAI-stage |
| $\beta_{2,i},\ \forall\, i \neq i^*$ | Arm $i$ is not eliminated by when round $m = \sigma_i$ and Arm $i^* \notin \mathcal{A}_{\sigma_i}$ during the BAI-stage |
| $F_i,\ \forall\, i \neq i^*$ | Final BAI round $\Sigma_f < \sigma_i$ and Arm $i \in \mathcal{A}_f$ |

**Remark F.1** (The event $F_i$)**.** *The event $F_i$ in words is the event that samples run out before the validity of the event $\Gamma_i$ could be checked and before the arm $i$ could be eliminated. Equivalent to the description in Table 3, we can write $F_i = \{\Sigma_f < \sigma_i\} \cap \{i \in \mathcal{A}_f\}$. Using De Morgan's rules, its complement is given by $F_i^c = \{\Sigma_f \geq \sigma_i\} \cup \{i \notin \mathcal{A}_f\}$. Consequently, $\Gamma_i \subseteq F_i^c$, and $\Gamma_i = \Gamma_i \cap F_i^c$. Since event $F_i$ requires that $i \in \mathcal{A}_f$, the event $\{t_{BAI} = T\} \subset F_i \ \forall\, i \neq i^*$.*

**Remark F.2** (Events are proper subsets of $\Omega$)**.** *Similar to the events defined in Table 2 for the analysis of PE, the events in Table 3 are proper subsets of the sample space $\Omega$. From the setup inherited from Improved-UCB, $\Gamma_i$ is a desirable and likely fate for Arm $i$ during the BAI-stage.*

To analyze the outcome of the BAI stage of the PE-CS algorithm we intuit and validate a three way partition of the sample space $\Omega$ into the events $\Gamma$, $\beta$, and $F$. $\Gamma$ is the event conditioning on which ensures that $\ell = i^*$ by requiring that the events $\Gamma_i$ held for each arm $i \neq i^*$. Additionally we require that there are samples remaining at the end of the BAI stage by intersecting with $\{t_{\text{BAI}} < T\}$. This structure makes the downstream analysis of the PE-stage tractable.

$$\Gamma = \{t_{\text{BAI}} < T\} \cap \bigcap_{i \neq i^*} \Gamma_i. \tag{232}$$

Let $A$ denote the set of sub-optimal arms, we know that $|A| = K - 1$. We use $\mathcal{P}(A)$ to denote the power set of $A$, that is the collection of all the possible sub-sets of $A$. Let $S \in \mathcal{P}(A)$ denote an arbitrary subset of A. We define an event $F(S) \subseteq \Omega$ parameterized by the set $S$ as,

$$F(S) = \{t_{\text{BAI}} = T\} \cap \left( \bigcap_{i \in S} F_i \cap \bigcap_{j \in A \backslash S} \left(\Gamma_j \cap F_j^c\right) \right) \tag{233}$$

Intuitively, the set $S$ consists of arms $i \in S$ for which $F_i$ held thereby making $\Gamma_i$ unverifiable. In contrast the arms $j$ contained in $j \in A \backslash S$ are those for which $\Gamma_i$ held. An inspection of the definition of $F(S)$ in Equation 233 reveals that $F(S_1) \cap F(S_2) = \phi \ \forall\, S_1, S_2 \in \mathcal{P}(A), S_1 \neq S_2$. Taking a union over all possible $F(S)$ we get the compound event $F$,

$$F = \bigcup_{S \in \mathcal{P}(A)} F(S). \tag{234}$$

Finally, the event $\beta$ is the event that $\Gamma_i$ did not hold for some arm $i \neq i^*$ despite it being verifiable ($F_i^c$ holding).

$$\beta = \bigcup_{i \neq i^*} \left(\Gamma_i^c \cap F_i^c\right). \tag{235}$$

**Lemma F.1** (A partition of $\Omega$ using $\Gamma$)**.** *The events $\Gamma$, $F$, and $\beta$ as defined in Equations 232, 234, and 235 respectively, form a mutually exclusive and exhaustive partition of the sample space $\Omega$.*

*Proof.* First we show that the sets are mutually exclusive by showing that their pairwise intersections namely $\Gamma \cap F$, $F \cap \beta$, and $\beta \cap \Gamma$ are all $\phi$. Starting off, it is easy to see that $\Gamma \cap F = \phi$ since,

$$\Gamma \cap F \subseteq \{t_{\text{BAI}} < T\} \cap \{t_{\text{BAI}} = T\} = \phi \text{ (definitions from Equations 232 and 234).} \tag{236}$$

Next, to show that $F \cap \beta = \phi$, it is sufficient to show that $F(S) \cap \beta = \phi$ for arbitrary $S \in \mathcal{P}(A)$.

$$F(S) \cap \beta = \{t_{\text{BAI}} = T\} \cap \left( \bigcap_{i \in S} F_i \cap \bigcap_{j \in A \backslash S} \left(\Gamma_j \cap F_j^c\right) \right) \cap \bigcup_{k \neq i^*} \left(\Gamma_k^c \cap F_k^c\right) \tag{237}$$

Since both $F_i \cap (\Gamma_i^c \cap F_i^c) = \phi$ and $(\Gamma_i \cap F_i^c) \cap (\Gamma_i^c \cap F_i^c) = \phi$, we shall have $F(S) \cap \beta = \phi$.

Lastly, for the pair $\beta \cap \Gamma$ we have,

$$\beta \cap \Gamma = \bigcup_{i \neq i^*} \left(\Gamma_i^c \cap F_i^c\right) \cap \bigcap_{j \neq i^*} \Gamma_j \quad \text{(since the indexing for } \beta \text{ and } \Gamma \text{ need not coincide)} \tag{238}$$

$$= \bigcup_{i \neq i^*} \left(\Gamma_i^c \cap F_i^c\right) \cap \bigcap_{j \neq i^*} \left(\Gamma_j \cap F_j^c\right) \quad \text{(since } \Gamma_j = F_j^c \cap \Gamma_j) \tag{239}$$

$$= \bigcup_{i \neq i^*} \bigcap_{j \neq i^*} \left( \left(\Gamma_i^c \cap F_i^c\right) \cap \left(\Gamma_j \cap F_j^c\right) \right) = \phi. \tag{240}$$

Next we show that $\Gamma \cup F \cup \beta = \Omega$, that is, the event collection considered in Lemma F.1 is exhaustive.

$$\Gamma \cup F \supset \Gamma \cup F(\phi) = \bigcap_{i \neq i^*} (\Gamma_i \cap F_i^c) \quad \text{(plugging } F(S) \text{ when } S = \phi \text{ per Equation 233)} \tag{241}$$

$$\implies \Gamma \cup F \cup \beta \supset \bigcap_{i \neq i^*} (\Gamma_i \cap F_i^c) \cup \bigcup_{j \neq i^*} \left(\Gamma_j^c \cap F_j^c\right) \tag{242}$$

$$= \bigcup_{j \neq i^*} \bigcap_{i \neq i^*} \left(\left(\Gamma_j^c \cap F_j^c\right) \cup (\Gamma_i \cap F_i^c)\right) \tag{243}$$

$$\supseteq \bigcap_{i \neq i^*} ((\Gamma_i^c \cap F_i^c) \cup (\Gamma_i \cap F_i^c)) = \bigcap_{i \neq i^*} F_i^c. \tag{244}$$

We shall now show that $F \cup \beta \supseteq \bigcup_{i \neq i^*} F_i$ which combined with Equation 244 completes the check on the exhaustive criteria. To do this we show that $F_i \subseteq F \cup \beta \;\; \forall\, i \neq i^*$.

PROOF THAT $F_i \subseteq F \cup \beta$:

To prove this result, we start with the event $F(\{i\})$,

$$F(\{i\}) = \{t_{\text{BAI}} = T\} \cap \left( F_i \cap \bigcap_{j \in A, j \neq i} \left(\Gamma_j \cap F_j^c\right) \right) \tag{245}$$

$$= F_i \cap \bigcap_{j \in A, j \neq i} \left(\Gamma_j \cap F_j^c\right) \quad (\because F_i \subset \{t_{\text{BAI}} = T\}). \tag{246}$$

Without loss of generality, let the set of remaining sub-optimal arms $A \setminus \{i\} = \{p, q, \ldots\}$. The idea behind this proof is to identify sub-events $F(\cdot)$ such that iteratively taking their union with one another and with events lying in $\beta$ reveals that $F_i \subseteq F \cup \beta$. Since $\beta = \bigcup_{k \neq i^*} (F_k^c \cap \Gamma_k^c)$ we have,

$$\left(F_p^c \cap \Gamma_p^c\right) \subseteq \beta \implies F_i \cap \left(\Gamma_p^c \cap F_p^c\right) \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq \beta \tag{247}$$

(intersecting with sets keeps us inside $\beta$)

$$\implies F(\{i\}) \cup F_i \cap \left(\Gamma_p^c \cap F_p^c\right) \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta \quad (\because F(\{i\}) \subseteq F) \tag{248}$$

$$\implies F_p^c \cap F_i \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta \tag{249}$$

$(\because \left(\Gamma_p^c \cap F_p^c\right) \cup \left(\Gamma_p \cap F_p^c\right)$ from $\beta, F(\{i\}))$

$$\implies F(\{i,p\}) \cup F_p^c \cap F_i \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta \tag{250}$$

$$\implies F_i \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta. \tag{251}$$

In reaching Equation 251, we have removed the dependence on Arm $p$ for the event on the left. With the next series of equations, we further remove the dependence on Arm $q$.

$$F_i \cap \bigcap_{j \in A \setminus \{i,p\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta \tag{252}$$

$$\implies F_i \cap \left(\Gamma_q \cap F_q^c\right) \cap \bigcap_{j \in A \setminus \{i,p,q\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta. \tag{253}$$

Similar to what we saw in the first iteration of this procedure, we have,

$$F_i \cap \left(\Gamma_q^c \cap F_q^c\right) \cap \bigcap_{j \in A \setminus \{i,p,q\}} \left(\Gamma_j \cap F_j^c\right) \subseteq \beta \tag{254}$$

$$\implies F_i \cap F_q^c \cap \bigcap_{j \in A \setminus \{i,p,q\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta \quad \text{(combining with Equation 253).} \tag{255}$$

Similar to how we reached the result in Equation 251 in starting from $F(\{i\})$, if instead we had started with the set $F(\{i,q\})$, and then eliminated the dependence on $p$, we would have shown,

$$F_i \cap F_q \cap \bigcap_{j \in A \setminus \{i,p,q\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta. \tag{256}$$

Combining Equations 255 and 256, we have,

$$F_i \cap \bigcap_{j \in A \setminus \{i,p,q\}} \left(\Gamma_j \cap F_j^c\right) \subseteq F \cup \beta. \tag{257}$$

Repeating this procedure iteratively, it is clear that the event on the left can be pruned down to simply $F_i$, and therefore,

$$F_i \subseteq F \cup \beta. \tag{258}$$

Since no assumptions were made on the choice of $i$, we have,

$$\bigcup_{i \neq i^*} F_i \subseteq F \cup \beta. \tag{259}$$

As mentioned earlier, we can combine $\Gamma \cup F \cup \beta \supset \bigcap_{i \neq i^*} F_i^c$ from Equation 244 and $F \cup \beta \supseteq \bigcup_{i \neq i^*} F_i$ from Equation 259 to obtain $\Gamma \cup F \cup \beta = \Omega$. □

**Corollary F.1** (Corollary to Lemma F.1)**.** *The events* $\Gamma, \{F(S)\}_{S \in \mathcal{P}(A)}, \beta$ *form a mutually exclusive and exhaustive partition over the sample space* $\Omega$*. This result follows trivially from Lemma F.1 and the fact that* $S_1 \neq S_2 \implies F(S_1) \cap F(S_2) = \phi$*.*

Next we prove an upper bound on the probability of the event $\beta$ which we will need repeatedly in proving subsequent results.

**Lemma F.2** (Bound on $\Pr(\beta)$)**.** *To bound the expected number of samples in all the cases pertinent to PE-CS we show the following bound on* $\Pr(\beta)$*,*

$$\Pr(\beta) < \frac{11}{T\Delta_{\min}^2} + \sum_{i \neq i^*} \frac{32}{T\Delta_i^2}. \tag{260}$$

*Proof.*

$$\Pr(\beta) = \Pr\left(\bigcup_{i \neq i^*} (\Gamma_i^c \cap F_i^c)\right) \tag{261}$$

$$= \Pr\left(\bigcup_{i \neq i^*} (\beta_{1,i} \cup \beta_{2,i})\right) \quad \text{(since } \Gamma_i^c \cap F_i^c = \beta_{1,i} \cup \beta_{2,i}\text{)} \tag{262}$$

$$\leq \Pr\left(\bigcup_{i \neq i^*} \beta_{1,i}\right) + \Pr\left(\bigcup_{i \neq i^*} \beta_{2,i}\right) \quad \text{(Union Bound)} \tag{263}$$

$$\leq \sum_{i \neq i^*} \Pr(\beta_{1,i}) + \Pr\left(\bigcup_{i \neq i^*} \{\text{Arm } i^* \notin \mathcal{A}_{\sigma_i}\}\right) \text{ (Union Bound, Latter clause of } \beta_{2,i}\text{)} \tag{264}$$

$$= \sum_{i \neq i^*} \Pr(\beta_{1,i}) + \Pr(\{\text{ Arm } i^* \notin \mathcal{A}_{\sigma_{\max}} \}). \tag{265}$$

Where $\sigma_{\max} = \max_{i \neq i^*} \sigma_i$, and Equation 265 follows from the fact that $i^* \notin \mathcal{A}_{\sigma_1} \implies i^* \notin \mathcal{A}_{\sigma_2}$ when $\sigma_2 > \sigma_1$.

The event $\beta_{1,i}$ is the event that Arm $i$ is not eliminated by round $\sigma_i$ while Arm $i^*$ is active at the end of the sampling for round $\sigma_i$. The term $\Pr(\beta_{1,i})$ therefore can be bounded in a manner analogous

to the way the probability of low-cost unsatisfactory arm $i$ not being eliminated by reference arm $\ell$ was bound in Equation 147. The difference being that the round number $\rho_i$ is replaced by the round $\sigma_i$, or equivalently, the gap $\Delta_{Q,i}$ is replaced by the gap $\Delta_i$. Therefore,

$$\Pr\left(\beta_{1,i}\right) \leq \frac{32}{T\Delta_i^2}. \tag{266}$$

The problem of analyzing $\Pr\left(\beta_{2,i}\right)$ is analogous to the analysis of $\Pr\left(B_{3,i}\right)$ in the proof of Lemma E.2 for the PE algorithm. By applying the Hoeffding bound (Lemma C.2) to the clauses in Equations 134 and 136 we were able to establish that the probability of the event $\{$Arm $\ell$ eliminated by unsatisfactory arm $i$ after the sampling for an arbitrary round $\rho$ concludes$\}$ is upper bounded by $\frac{4^{\kappa}+1}{T\tilde{\Delta}_{\rho}^2}$. Since for the BAI setting the samples of all the arms are always matched ($\kappa = 0$) here we shall have,

$$\Pr\left(\{\text{ Arm } i^* \notin \mathcal{A}_{\sigma_{\max}}\}\right) \leq \sum_{\rho=0}^{m_{\sigma_{\max}}-1} \frac{2}{T\tilde{\Delta}_{\rho}^2} \leq \frac{2}{T} \sum_{\rho=0}^{m_{\sigma_{\max}}-1} 4^{\rho} \tag{267}$$

$$< \frac{2 \cdot 4^{\sigma_{\max}}}{3T} \tag{268}$$

$$\leq \frac{2}{3T\tilde{\Delta}_{\sigma_{\max}}^2} \tag{269}$$

$$\leq \frac{32}{3T\Delta_{\min}^2} \tag{270}$$

$$< \frac{11}{T\Delta_{\min}^2}. \tag{271}$$

Combining the bounds shown in Equations 266 and 271 we obtain the overall bound stated in Lemma F.2. □

We now move on to analyzing the evolution of samples in the PE stage of PE-CS. The pieces needed from the analysis of the BAI stage are the partition over $\Omega$ from Lemma F.1, the bound on $\Pr\left(\beta\right)$ shown in Lemma F.2, and the Iterated Expectation Lemma C.3. The key difference between the analysis of PE and the PE-stage in PE-CS is the possibility that the round number $\Sigma_i$ to which the samples of an arbitrary arm $i \neq i^*$ advance during the BAI-stage exceeds the round number $\rho_i$ defined in Equation 114. Our modular proof technique sequesters both the pathological (event $F$) and the unlikely (event $\beta$) outcomes of the BAI stage away from the PE stage. In our approach the $\Sigma_i > \rho_i$ case in the analysis of the PE-stage of PE-CS only surfaces for episode $a^*$. Moreover, analyzing samples accrued during episode $a^*$ is only called for when bounding the expected number of samples of the best arm $i^*$.

**Remark F.3.** *Similar to Remark E.1 we note here that Arm $i^*$ during the PE stage of PE-CS really refers to a hypothetical Bandit Arm with expected return $(1-\alpha)\mu_*$.*

Using all the definitions and constructs introduced in this section, we are now in a position to show an upper bound on the expected number of samples of low-cost arms in Lemma F.3, the best arm $i^*$ in Lemma F.4, and high-cost arms in Lemma F.5.

BOUND SAMPLES FOR ARMS $i < a^*$

**Lemma F.3** (Bound on the expected number of samples of a low-cost arm under PE-CS)**.** *For any low cost unsatisfactory arm $i < a^*$, its expected number of samples accrued is upper bounded by,*

$$\mathbb{E}\left[n_i\left(T\right)\right] < 1 + \frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2} + \frac{43}{\Delta_{Q,i}^2} + \mathbb{E}\left[n_i\left(T\right) \mid \beta\right] \cdot \left(\frac{11}{T\Delta_{\min}^2} + \sum_{j \neq i^*} \frac{32}{T\Delta_j^2}\right).$$

*Proof.* We begin the analysis by applying the Iterated Expectation Lemma C.3 to the partition developed in Lemma F.1.

$$\mathbb{E}\left[n_i\left(T\right)\right] = \mathbb{E}\left[n_i\left(T\right) \mid \Gamma\right]\Pr\left(\Gamma\right) + \sum_{S \in \mathcal{P}(A)} \mathbb{E}\left[n_i\left(T\right) \mid F(S)\right]\Pr\left(F(S)\right) + \mathbb{E}\left[n_i\left(T\right) \mid \beta\right]\Pr\left(\beta\right) \tag{272}$$

$$\leq \max\left\{\mathbb{E}\left[n_i\left(T\right) \mid \Gamma\right], \left\{\mathbb{E}\left[n_i\left(T\right) \mid F(S)\right]\right\}_{S \in \mathcal{P}(S)}\right\} + \mathbb{E}\left[n_i\left(T\right) \mid \beta\right]\Pr\left(\beta\right) \tag{273}$$

$$(\because \Pr\left(\Gamma\right) + \textstyle\sum \Pr\left(F(S)\right) < 1)$$

$$\leq \max\left\{\mathbb{E}\left[n_i\left(T\right) \mid \Gamma\right], \tau_{\sigma_i}\right\} + \mathbb{E}\left[n_i\left(T\right) \mid \beta\right]\Pr\left(\beta\right). \tag{274}$$

Where Equation 274 follows from the fact that conditioned on any $F(S)$, the maximum round up to which arm $i$ can be sampled is $\sigma_i$ both in the case when $\Gamma_i$ holds ($i \notin S$), and in the case when $i \in S$, and $F_i = \{\Sigma_f < \sigma_i\} \cap \{i \in \mathcal{A}_f\}$ holds instead. We now proceed by separately bounding the $\mathbb{E}\left[n_i\left(T\right) \mid \Gamma\right]$ term by further conditioning on the cases where $\Sigma_i > \rho_i$ and $\Sigma_i \leq \rho_i$.

$$\begin{aligned}\mathbb{E}\left[n_i(T) \mid \Gamma\right] &= \mathbb{E}\left[n_i(T) \mid \Sigma_i \leq \rho_i, \Gamma\right]\Pr\left(\Sigma_i \leq \rho_i \mid \Gamma\right) \\ &\quad + \mathbb{E}\left[n_i(T) \mid \Sigma_i > \rho_i, \Gamma\right]\Pr\left(\Sigma_i > \rho_i \mid \Gamma\right) \quad \text{(using Lemma C.3 on } \{\Sigma_i \leq \rho_i\}) \end{aligned} \tag{275}$$

$$\leq \mathbb{E}\left[n_i(T) \mid \Sigma_i \leq \rho_i, \Gamma\right] = \mathbb{E}_1\left[n_i(T)\right] \quad \text{(introducing shorthand notation } \mathbb{E}_1). \tag{276}$$

Where $\rho_i$ is as defined in Equation 114. In writing Equation 276 we leverage the fact that for a low cost arm with index $i < a^*$, $\Delta_{Q,i} = (1-\alpha)\mu_* - \mu_i$ is necessarily a smaller gap than $\Delta_i = \mu_* - \mu_i$ since by construction, each of these low cost arms has a return $\mu_i < \mu_{\text{CS}} = (1-\alpha)\mu_*$. It follows that $\Pr\left(\Sigma_i > \rho_i, \Gamma\right) = \Pr\left(\Sigma_i > \rho_i \mid \Gamma\right) = 0$ because the largest value that the random variable $\Sigma_i$ can take under $\Gamma$ is $\sigma_i$, and $\Delta_{Q,i} < \Delta_i \implies \rho_i > \sigma_i$.

BOUND ON $\mathbb{E}_1\left[n_i\left(T\right)\right]$

Since we enter the PE stage of the algorithm with a round number $\Sigma_i \leq \rho_i$, we use the same event construction of the compound event $G_i \, \forall \, i < a^*$, defined and used in the Proof of Lemma E.2. Therefore, just as before we work towards a bound by conditioning on the partition with $G_i$ introduced in Lemma E.1[2]. Let $\Lambda_i$ be the random variable denoting the highest round number corresponding to which sampling was performed for arm $i$ during the run of PE-CS.

$$\mathbb{E}_1\left[n_i(T)\right] = \mathbb{E}_1\left[n_i(T) \mid G_i\right]\Pr\left(G_i\right) + \mathbb{E}_1\left[n_i(T) \mid B_{1,i}\right]\Pr\left(B_{1,i}\right) + \mathbb{E}_1\left[n_i(T) \mid B_i\right]\Pr\left(B_i\right) \tag{277}$$

$$\leq \max\left\{\mathbb{E}_1\left[n_i(T) \mid G_i\right], \mathbb{E}_1\left[n_i(T) \mid B_{1,i}\right]\right\} + T \cdot \Pr\left(B_i\right) \tag{278}$$

$$(\text{Since } \Pr\left(G_i\right) + \Pr\left(B_{1,i}\right) < 1\,)$$

$$= \max\left\{\mathbb{E}_1\left[n_i(T) \mid G_i\right], \mathbb{E}_1\left[n_i(1, t_{\text{BAI}}) + n_i(t_{\text{BAI}} + 1, T) \mid B_{1,i}\right]\right\} + T \cdot \Pr\left(B_i\right) \tag{279}$$

$$\leq \max\left\{\mathbb{E}_1\left[\tau_{\Lambda_i} \mid G_i\right], \mathbb{E}_1\left[\tau_{\Sigma_i} + 0 \mid B_{1,i}\right]\right\} + T \cdot \Pr\left(B_i\right) \tag{280}$$

$$(\tau_m \text{ as defined in Equation 115})$$

$$\leq \max\left\{\tau_{\rho_i}, \tau_{\sigma_i}\right\} + T \cdot \Pr\left(B_i\right) \tag{281}$$

$$= \tau_{\rho_i} + T \cdot \Pr\left(B_i\right) \text{ (since } (1-\alpha)\,\mu_* - \mu_i = \Delta_{Q,i} < \Delta_i = \mu_* - \mu_i \, \forall \, i < a^*) \tag{282}$$

$$< 1 + \frac{32 \log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2} + T \cdot \Pr\left(B_i\right) \quad \text{(Using bound on } \tau_{\rho_i} \text{ in Equation 133).} \tag{283}$$

Where the treatment of the random variable $n_i(t_{\text{BAI}} + 1, T)$ is based on the expression for the additional rounds for which arm $i$ is sampled during the PE stage of PE-CS. In Equation 280 conditioned on $G_i$ there may be more samples, however conditioned on $B_{1,i}$ there shall be no further samples

[2]The probability operator Pr is also for the conditional distribution conditioned on $\{\Sigma_i \leq \rho_i, \Gamma\}$.

since the episode corresponding to $i$ is never initiated. Now to bound $\Pr(B_i \mid \Sigma_i \leq \rho_i, \Gamma)$ we use arguments similar to the ones in Proof of Lemma E.2.

$$\Pr(B_i \mid \Sigma_i \leq \rho_i, \Gamma) = \Pr(B_{1,i} \cup B_{2,i} \mid \Sigma_i \leq \rho_i, \Gamma). \tag{284}$$

Where $B_{1,i} = G_{1,i} \cap G_{2,i}^c$ and $B_{2,i} = G_{1,i} \cap G_{3,i}^c$ as in the proof of Lemma E.2. The Probability $\Pr(B_{1,i})$ in Lemma E.2 was bound by the probability of either Clause 134 or Clause 135 being violated during round $\rho_i$. Due to the parallel nature of the construction here, and the possibility of round $\rho_i$ being conducted, we can upper bound $\Pr(B_{1,i} \mid \Sigma_i \leq \rho_i, \Gamma)$ identically as,

$$\Pr(B_{1,i} \mid \Sigma_i \leq \rho_i, \Gamma) \leq \frac{32}{T\Delta_{Q,i}^2}. \tag{285}$$

Now we move on to bounding $\Pr(B_{2,i} \mid \Sigma_i \leq \rho_i, \Gamma)$ by bounding the probability of arm $\ell$ being eliminated in any round $\omega_i = \rho$ lying in the range $\Sigma_i \leq \rho \leq \rho_i$. Just like in the proof of Lemma E.2, even here Clauses 134 and 136 holding simultaneously preclude arm $\ell$ from being eliminated by arm $i$ regardless of round $\rho$. Therefore using work identical to the one that goes into establishing Equations 152 and 153 we have,

$$\Pr(B_{2,i} \mid \Sigma_i \leq \rho_i, \Gamma) \leq \sum_{\rho=\Sigma_i}^{\rho_i - 1} \frac{4^\kappa + 1}{T\tilde{\Delta}_\rho^2} \tag{286}$$

$$\leq \sum_{\rho=0}^{\rho_i - 1} \frac{4^\kappa + 1}{T\tilde{\Delta}_\rho^2} \text{ (Because } \Sigma_i \geq 0) \tag{287}$$

$$< \frac{11}{T\Delta_{Q,i}^2} \text{ (From Equation 153 in Lemma E.2, } \kappa = 0). \tag{288}$$

Applying the Union bound to Equation 284, and plugging in the bounds in 285 and 288, we have,

$$\Pr(B_i \mid \Sigma_i \leq \rho_i, \Gamma) < \frac{43}{T\Delta_{Q,i}^2}. \tag{289}$$

Substituting the bounds in Equations 289, 283, and Lemma F.2 into Equation 274 we obtain,

$$\mathbb{E}[n_i(T)] \leq \max\{\tau_{\rho_i} + T \cdot \Pr(B_i), \tau_{\sigma_i}\} + \mathbb{E}[n_i(T) \mid \beta] \Pr(\beta) \tag{290}$$

$$< 1 + \frac{32 \log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2} + \frac{43}{\Delta_{Q,i}^2} + \mathbb{E}[n_i(T) \mid \beta] \Pr(\beta). \tag{291}$$

Which when combined with Lemma F.2 is the bound stated in Lemma F.3. □

### BOUND SAMPLES FOR ARM $i^*$

**Lemma F.4** (Bound on the expected number of samples of the best arm)**.** *For the best arm $i^* = \arg\max_{i \in [K]} \mu_i$, the expected number of samples accrued is upper bounded as,*

$$\mathbb{E}[n_{i^*}(T)] < 1 + \max\left\{\frac{32 \log\left(T\Delta_{\min}^2\right)}{\Delta_{\min}^2}, \left\{\frac{32 \log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}\right\}_{i \leq a^*}\right\}$$
$$+ \sum_{i=1}^{a^*-1} \frac{43}{\Delta_{Q,i}^2} + 2\left(\frac{32}{\Delta_{a^*}^2} + \frac{43}{\Delta_{Q,a^*}^2}\right) + \mathbb{E}[n_{i^*}(T) \mid \beta] \cdot \left(\frac{11}{T\Delta_{\min}^2} + \sum_{j \neq i^*} \frac{32}{T\Delta_j^2}\right).$$

*Proof.* Just like in the proof of Lemma F.3 we sequester away outcomes of the BAI stage that make analysis of the PE stage intractable.

$$\mathbb{E}[n_{i^*}(T)] = \mathbb{E}[n_{i^*}(T) \mid \Gamma] \Pr(\Gamma) + \sum_{S \in \mathcal{P}(A)} \mathbb{E}[n_{i^*}(T) \mid F(S)] \Pr(F(S)) + \mathbb{E}[n_{i^*}(T) \mid \beta] \Pr(\beta) \tag{292}$$

$$\leq \max\left\{\mathbb{E}[n_{i^*}(T) \mid \Gamma], \{\mathbb{E}[n_{i^*}(T) \mid F(S)]\}_{S \in \mathcal{P}(S)}\right\} + \mathbb{E}[n_{i^*}(T) \mid \beta] \Pr(\beta) \tag{293}$$

$$(\because \Pr(\Gamma) + \Pr(F) < 1)$$

$$\leq \max\{\mathbb{E}[n_{i^*}(T) \mid \Gamma], \tau_{\sigma_{\max}}\} + \mathbb{E}[n_{i^*}(T) \mid \beta] \Pr(\beta). \tag{294}$$

Equation 294 results from the observation that the maximum round number up till which any arm is sampled during the BAI stage under $F(S)\ \forall S \in \mathcal{P}(A)$ is $\sigma_{\max}$. Conditioned on $\Gamma$, the reference arm $\ell$ is identified correctly to be $i^*$. Consequently, the expected number of samples $\mathbb{E}[n_{i^*}(T) \mid \Gamma]$ can be analyzed in a manner that closely parallels the analysis of $\mathbb{E}[n_\ell(T)]$ in the Proof of Lemma E.3.

The key difference from Lemma E.3 is that we grapple with the case $\Sigma_{a^*} > \rho_{a^*}$. This possibility arises because our bandit instance may have $\Delta_{Q,a^*} < \Delta_{a^*}$, which in turn implies that $\Pr(\rho_{a^*} < \Sigma_{a^*} < \sigma_{a^*} \mid \Gamma) > 0$. The outcome $\Sigma_{a^*} > \rho_{a^*}$ skips the checks associated with the events $G_{2,i}$ and $G_{3,i}$. Mathematically, this means that for the event $G_{a^*}$ as defined in Equation 116 we shall have $\Pr(G_{1,a^*} \cap ((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*}) \mid \Sigma_{a^*} > \rho_{a^*}, \Gamma) = 0$. This motivates us to expand the scope of the good event $G_{a^*}$.

First we define new event $G_{4,a^*} \subset \Omega$ and then we augment the definition of $G_{a^*}$ using this new event.

$G_{4,a^*}$ : {Arm $i^*$ is eliminated by arm $i$ in round $\Sigma_{a^*}$, during episode $a^*$ of the PE stage of PE-CS}

**Remark F.4.** *Since arm $a^*$ enters the PE stage of PE-CS with samples corresponding to $\Sigma_{a^*}$ number of rounds already accrued, checking whether the event $G_{4,a^*}$ holds does not involve any further sampling of arms.*

The definition of $G_{a^*}$ is now changed to the one in Equation 295 which supersedes the prior generic $G_{a^*}$ (Equation 116).

$$G_{a^*} = \underbrace{G_{1,a^*}}_{\text{Ep. } a^* \text{ is executed}} \cap \left( \underbrace{((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*})}_{\text{Prior } G_{a*} \text{ Clause}} \cup \underbrace{(G_{4,a^*} \cap \{\Sigma_{a^*} > \rho_{a^*}\})}_{\text{New Clause}} \right). \tag{295}$$

**Remark F.5** (Implicit event in Prior $G_{a^*}$ clause)**.** *We remark here that the event $\{\Sigma_{a^*} \leq \rho_{a^*}\}$ is implicit in the event $G_{1,a^*} \cap ((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*})$ i.e. $G_{1,a^*} \cap ((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*}) \subseteq \{\Sigma_{a^*} \leq \rho_{a^*}\}$. This is because the event is contingent on correct eliminations happening leading up to and during $\rho_{a^*}$.*

While retaining the definition of $G$ from Equation 154, and the definition of $B_{1,a^*}$ from Lemma E.1 we redefine $B_{a^*}$ so that the relation $G^c_{a^*} = B_{1,a^*} \cup B_{a^*}$ continues to hold.

$$G_{a^*} = G_{1,a^*} \cap (((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*}) \cup (G_{4,a^*} \cap \{\Sigma_{a^*} > \rho_{a^*}\})) \tag{296}$$

$$= G_{1,a^*} \cap ((((G_{2,a^*} \cap G_{3,a^*}) \cup E_{a^*})) \cap \{\Sigma_{a^*} \leq \rho_{a^*}\} \cup (G_{4,a^*} \cap \{\Sigma_{a^*} > \rho_{a^*}\})) \tag{297}$$

(due to Remark F.5)

$$\implies G^c_{a^*} = G^c_{1,a^*} \cup \left( \underbrace{(((G^c_{2,a^*} \cup G^c_{3,a^*}) \cap E^c_{a^*})}_{B^{\text{original}}_{a^*}} \cup \{\Sigma_{a^*} > \rho_{a^*}\}) \cap \left(G^c_{4,a^*} \cup \{\Sigma_{a^*} \leq \rho_{a^*}\}\right) \right) \tag{298}$$

$$= B_{1,a^*} \cup \Big( \Big( \Big( B^{\text{original}}_{a^*} \cap \{\Sigma_{a^*} \leq \rho_{a^*}\} \Big) \cup \{\Sigma_{a^*} > \rho_{a^*}\} \Big) \cap \left(G^c_{4,a^*} \cap \{\Sigma_{a^*} > \rho_{a^*}\}\right) \cup \{\Sigma_{a^*} \leq \rho_{a^*}\}) \Big) \tag{299}$$

$$= B_{1,a^*} \cup \underbrace{\Big( \Big( B^{\text{original}}_{a^*} \cap \{\Sigma_{a^*} \leq \rho_{a^*}\} \Big) \cup \left(G^c_{4,a^*} \cap \{\Sigma_{a^*} > \rho_{a^*}\}\right) \Big)}_{B_{a^*}}. \tag{300}$$

Where Equation 299 follows from $A \cup B = (A \cap B^c) \cup B$ being applied to both the left and right clauses. Equation 300 gives us the updated definition of the event $B_{a^*}$. Armed with the updated

event $B_{a^*}$, we can now develop a bound for $\Pr(G^c \mid \Gamma)$ using the result in Lemma E.6. By trivially generalizing the bound on $\Pr(G^c)$ developed in Lemma E.6 to the scenario for this Proof where we condition on $\Gamma$ we shall have,

$$\Pr(G^c \mid \Gamma) \leq \sum_{i=1}^{a^*} \Pr(B_j \mid \Gamma) + \Pr(B_{a^*} \mid \Gamma) \tag{301}$$

$$\leq \sum_{i=1}^{a^*-1} \frac{43}{T\Delta_{Q,i}^2} + 2 \cdot \Pr(B_{a^*} \mid \Gamma), \tag{302}$$

$\because$ Equation 289, and $\Pr(\Sigma_i > \rho_i \mid \Gamma) = 0 \ \forall i < a^*$. We need now only develop a bound for $\Pr(B_{a^*} \mid \Gamma)$ under its definition in Equation 300.

$$\Pr(B_{a^*} \mid \Gamma) = \Pr\left(\left(B_{a^*}^{\text{original}} \cap \{\Sigma_{a^*} \leq \rho_{a^*}\}\right) \cup \left(G_{4,a^*}^c \cap \{\Sigma_{a^*} > \rho_{a^*}\}\right) \mid \Gamma\right) \tag{303}$$

$$\leq \Pr\left(B_{a^*}^{\text{original}} \cap \{\Sigma_{a^*} \leq \rho_{a^*}\} \mid \Gamma\right) + \Pr\left(G_{4,a^*}^c \cap \{\Sigma_{a^*} > \rho_{a^*}\} \mid \Gamma\right) \tag{304}$$

(Union Bound)

$$\leq \frac{43}{T\Delta_{Q,a^*}^2} + \Pr\left(G_{4,a^*}^c \cap \{\Sigma_{a^*} > \rho_{a^*}\} \mid \Gamma\right) \quad \text{(From Equation 217).} \tag{305}$$

From Equation 212 in the Proof of Lemma E.2, we have the probability of arm $\ell$ not being eliminated by arm $a^*$ during round $\omega_{a^*}$ of episode $a^*$ as being upper bounded by $\frac{2}{T\tilde{\Delta}_{\omega_{a^*}}^2}$. Therefore since elimination under $G_{4,a^*}$ happens during episode $a^*$ in round $\Sigma_{a^*}$,

$$\Pr\left(G_{4,a^*}^c \cap \{\Sigma_{a^*} > \rho_{a^*}\} \mid \Gamma\right) \leq \frac{2}{T\tilde{\Delta}_{\sigma_{a^*}}^2} \tag{306}$$

(since $\Pr(\Sigma_{a^*} > \sigma_{a^*} \mid \Gamma) = 0$, and $\tilde{\Delta}_m$ decreases with $m$).

$$\leq \frac{32}{T\Delta_{a^*}^2} \quad \text{(since by definition of } \sigma_i, \tilde{\Delta}_{\sigma_i} \geq \frac{\Delta_i}{4}\text{).} \tag{307}$$

Combining the bounds in Equations 305 and 307 with the Expression 302 we have,

$$\Pr(G^c \mid \Gamma) \leq \sum_{i=1}^{a^*-1} \frac{43}{T\Delta_{Q,i}^2} + 2\left(\frac{43}{T\Delta_{Q,a^*}^2} + \frac{32}{T\Delta_{a^*}^2}\right). \tag{308}$$

Armed with the result in Equation 308, we are now in a position to bound $\mathbb{E}[n_{i^*}(T) \mid \Gamma]$.

$$\mathbb{E}[n_{i^*}(T) \mid \Gamma] = \mathbb{E}[n_{i^*}(T) \mid G, \Gamma] \Pr(G \mid \Gamma) + \mathbb{E}[n_{i^*}(T) \mid G^c, \Gamma] \Pr(G^c \mid \Gamma) \tag{309}$$

$$\leq \mathbb{E}[n_{i^*}(T) \mid G, \Gamma] + T \cdot \Pr(G^c \mid \Gamma) \tag{310}$$

$$\leq \mathbb{E}[n_{i^*}(T) \mid G, \Gamma] + \sum_{i=1}^{a^*-1} \frac{43}{\Delta_{Q,i}^2} + 2\left(\frac{43}{\Delta_{Q,a^*}^2} + \frac{32}{\Delta_{a^*}^2}\right). \tag{311}$$

As in the proof of Lemma E.3, let $Z$ denote the last PE episode. To bound $\mathbb{E}[n_{i^*}(T) \mid G, \Gamma]$ we leverage the result $\Pr(n_{i^*}(T) > \max_{i \neq i^*} n_i(1, t_Z) \mid G, \Gamma) = 0$. Conditioned on $G, \Gamma$ there can be no further sampling of the best arm $i^*$ beyond time $t_Z$, since under $\Gamma$ the best arm $i^*$ is the reference arm. Consequently,

$$\mathbb{E}[n_{i^*}(T) \mid G, \Gamma] \leq \mathbb{E}\left[\max_{i \neq i^*} n_i(1, t_Z) \mid G, \Gamma\right] \tag{312}$$

$$\leq \mathbb{E}\left[\max_{i \neq i^*} n_i(1, t_{a^*}) \mid G, \Gamma\right] \quad \text{(because } \Pr(Z > a^* \mid G, \Gamma) = 0\text{)} \tag{313}$$

$$= \mathbb{E}\left[\max\left\{\{n_i(1, t_{a^*})\}_{i<a^*}, n_{a^*}(1, t_{a^*}), \{n_i(1, t_{a^*})\}_{i>a^*}\right\} \mid G, \Gamma\right] \tag{314}$$

$$\leq \mathbb{E}\left[\max\left\{\{\tau_{\Lambda_i}\}_{i<a^*}, \tau_{\Lambda_{a^*}}, \{\tau_{\Sigma_i}\}_{i>a^*}\right\} \mid G, \Gamma\right] \tag{315}$$

(using definitions of $\tau$, $\Lambda$, and $\Sigma$)

$$\leq \max\left\{\left\{\tau_{\rho_i}\right\}_{i<a^*}, \max\left\{\tau_{\rho_{a^*}}, \tau_{\sigma_{a^*}}\right\}, \left\{\tau_{\sigma_i}\right\}_{i>a^*}\right\} \tag{316}$$

(because $\Pr\left(\Lambda_{a^*} > \max\left\{\rho_{a^*}, \sigma_{a^*}\right\}\right) = 0$)

$$\leq \max\left\{\left\{\tau_{\rho_i}\right\}_{i\leq a^*}, \tau_{\sigma_{\max}}\right\}. \tag{317}$$

Combining Equations 317 and 311 and using the result in Equation 294 we obtain the upper bound stated in Lemma F.4,

$$\mathbb{E}\left[n_{i^*}\left(T\right)\right] \leq \max\left\{\left\{\tau_{\rho_i}\right\}_{i\leq a^*}, \tau_{\sigma_{\max}}\right\} + \sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2} + 2\left(\frac{43}{\Delta_{Q,a^*}^2} + \frac{32}{\Delta_{a^*}^2}\right) + \mathbb{E}\left[n_{i^*}\left(T\right) \mid \beta\right]\Pr\left(\beta\right) \tag{318}$$

$$< 1 + \max\left\{\frac{32\log\left(T\Delta_{\min}^2\right)}{\Delta_{\min}^2}, \left\{\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}\right\}_{i\leq a^*}\right\} + \sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2} \tag{319}$$

$$+ 2\left(\frac{43}{\Delta_{Q,a^*}^2} + \frac{32}{\Delta_{a^*}^2}\right) + \mathbb{E}\left[n_{i^*}\left(T\right) \mid \beta\right]\left(\frac{11}{T\Delta_{\min}^2} + \sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right). \tag{320}$$

Where Equation 320 follows directly from from Equation 117 and Lemma F.2. □

BOUND SAMPLES FOR ARMS $i > a^*, i \neq i^*$

**Lemma F.5** (Bound on the expected number of samples of high-cost arms)**.** *For any high cost arm with $i > a^*, i \neq i^*$, its expected number of samples are upper bounded as,*

$$\mathbb{E}\left[n_i\left(T\right)\right] < 1 + \frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2} + \mathbb{E}\left[n_i\left(T\right) \mid \left\{Z > a^*\right\}, \Gamma\right]\cdot\left(\frac{32}{T\Delta_{a^*}^2} + \frac{43}{T\Delta_{Q,a^*}^2}\right) + \mathbb{E}\left[n_i\left(T\right) \mid \beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2} + \sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right).$$

*Proof.* To prove the bound stated in Lemma F.5 we proceed with initial steps identical to the ones that go into proving Lemmas F.3 and F.4.

$$\mathbb{E}\left[n_i(T)\right] = \mathbb{E}\left[n_i(T) \mid \Gamma\right]\Pr\left(\Gamma\right) + \sum_{S\in\mathcal{P}(A)}\mathbb{E}\left[n_i(T) \mid F(S)\right]\Pr\left(F(S)\right) + \mathbb{E}\left[n_i(T) \mid \beta\right]\Pr\left(\beta\right) \tag{321}$$

$$\leq \max\left\{\mathbb{E}\left[n_i(T) \mid \Gamma\right], \max_{S\in\mathcal{P}(A)}\mathbb{E}\left[n_i(T) \mid F(S)\right]\right\} + \mathbb{E}\left[n_i(T) \mid \beta\right]\Pr\left(\beta\right) \tag{322}$$

($\because \Pr\left(\Gamma\right) + \Pr\left(F\right) < 1$)

$$\leq \max\left\{\mathbb{E}\left[n_i(T) \mid \Gamma\right], \tau_{\sigma_i}\right\} + \mathbb{E}\left[n_i(T) \mid \beta\right]\left(\frac{11}{T\Delta_{\min}^2} + \sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right). \tag{323}$$

Where the final bound is from Lemma F.2. Now we must bound the expectation term $\mathbb{E}\left[n_i(T) \mid \Gamma\right]$. For this we recognize that during the PE-stage, the critical event which determines the number of samples further accrued for a high-cost arm is whether the final PE episode $Z > a^*$ or not. In the case when $Z \leq a^*$, there are no further samples of arm $i$ accrued beyond the BAI-stage.

$$\mathbb{E}\left[n_i(T) \mid \Gamma\right] = \mathbb{E}\left[n_i(T) \mid \left\{Z > a^*\right\}, \Gamma\right]\Pr\left(Z > a^* \mid \Gamma\right) + \mathbb{E}\left[n_i(T) \mid \left\{Z \leq a^*\right\}, \Gamma\right]\Pr\left(Z \leq a^* \mid \Gamma\right) \tag{324}$$

$$\leq \tau_{\sigma_i} + \mathbb{E}\left[n_i(T) \mid \left\{Z > a^*\right\}, \Gamma\right]\Pr\left(B_{a^*} \mid \Gamma\right) \quad \left(\text{since } \left\{Z > a^*\right\} \subseteq B_{a^*}\right) \tag{325}$$

$$\leq \tau_{\sigma_i} + \mathbb{E}\left[n_i(T) \mid \left\{Z > a^*\right\}, \Gamma\right]\cdot\left(\frac{32}{T\Delta_{a^*}^2} + \frac{43}{T\Delta_{Q,a^*}^2}\right). \tag{326}$$

Where the final line follows by using the bound developed in Equation 305. Returning to bounding $\mathbb{E}\left[n_i\left(T\right)\right]$ we obtain the bound stated in Lemma F.5,

$$\mathbb{E}\left[n_i(T)\right] \leq \tau_{\sigma_i} + \mathbb{E}\left[n_i(T) \mid \{Z > a^*\}, \Gamma\right] \cdot \left(\frac{32}{T\Delta_{a^*}^2} + \frac{43}{T\Delta_{Q,a^*}^2}\right) + \mathbb{E}\left[n_i(T) \mid \beta\right] \left(\frac{11}{T\Delta_{\min}^2} + \sum_{j \neq i^*} \frac{32}{T\Delta_j^2}\right) \tag{327}$$

$$< 1 + \frac{32 \log\left(T\Delta_i^2\right)}{\Delta_i^2} + \mathbb{E}\left[n_i(T) \mid \{Z > a^*\}, \Gamma\right] \cdot \left(\frac{32}{T\Delta_{a^*}^2} + \frac{43}{T\Delta_{Q,a^*}^2}\right) + \mathbb{E}\left[n_i(T) \mid \beta\right] \left(\frac{11}{T\Delta_{\min}^2} + \sum_{j \neq i^*} \frac{32}{T\Delta_j^2}\right). \tag{328}$$

□

Finally we have all the pieces needed to prove Theorem 3.2. We combine the results obtained in Lemmas F.3, F.4, and F.5 by adding together the contributions to regret of all three categories of arms while collating like terms.

*Proof of Theorem 3.2.* Using Equation 16 from the Regret decomposition Lemma C.5 we can express and bound the Expected Cumulative Cost Regret as,

$$\mathbb{E}\left[\text{Cost\_Reg}\left(T,\nu\right)\right] \tag{329}$$

$$=\sum_{i=1}^{K}\Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\right] \tag{330}$$

$$=\sum_{i<a^*}\Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\right]+\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\right]+\Delta_{C,i^*}\mathbb{E}\left[n_{i^*}\left(T\right)\right] \tag{331}$$

$$=\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\right]+\Delta_{C,i^*}\mathbb{E}\left[n_{i^*}\left(T\right)\right]\quad\text{(because }\Delta_{C,i}=0,\ \forall\, i\leq a^*\text{)} \tag{332}$$

$$\begin{aligned}<&\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}+\mathbb{E}\left[n_i\left(T\right)\mid\left\{Z>a^*\right\},\Gamma\right]\cdot\left(\frac{32}{T\Delta_{a^*}^2}+\frac{43}{T\Delta_{Q,a^*}^2}\right)\right.\\&\left.\quad+\mathbb{E}\left[n_i\left(T\right)\mid\beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\right)\\&+\Delta_{C,i^*}\left(1+\max\left\{\frac{32\log\left(T\Delta_{\min}^2\right)}{\Delta_{\min}^2},\left\{\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}\right\}_{i\leq a^*}\right\}+\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}\right.\\&\left.\quad+2\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)+\mathbb{E}\left[n_{i^*}\left(T\right)\mid\beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\right)\end{aligned} \tag{333}$$

$$\begin{aligned}=&\sum_{i>a^*}\Delta_{C,i}\left(1+\mathbb{E}\left[n_i\left(T\right)\mid\beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\right)+\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\\&+\Delta_{C,i^*}\left(\max_{\Delta\in\{\Delta_{\min}\}\cup\{\Delta_{Q,j}\}_{j\leq a^*}}\left\{\frac{32\log\left(T\Delta^2\right)}{\Delta^2}\right\}+\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}+2\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\right)\\&+\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\mathbb{E}\left[n_i\left(T\right)\mid\left\{Z>a^*\right\},\Gamma\right]\cdot\left(\frac{32}{T\Delta_{a^*}^2}+\frac{43}{T\Delta_{Q,a^*}^2}\right)\end{aligned} \tag{334}$$

$$\begin{aligned}=&\sum_{i>a^*}\Delta_{C,i}+\Delta_{C,\max}\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\sum_{i>a^*}\mathbb{E}\left[n_i\left(T\right)\mid\beta\right]+\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\\&+\Delta_{C,i^*}\left(\max_{\Delta\in\{\Delta_{\min}\}\cup\{\Delta_{Q,j}\}_{j\leq a^*}}\left\{\frac{32\log\left(T\Delta^2\right)}{\Delta^2}\right\}+\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}+2\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\right)\\&+\max_{i>a^*,i\neq i^*}\Delta_{C,i}\left(\frac{32}{T\Delta_{a^*}^2}+\frac{43}{T\Delta_{Q,a^*}^2}\right)\sum_{i>a^*,i\neq i^*}\mathbb{E}\left[n_i\left(T\right)\mid\left\{Z>a^*\right\},\Gamma\right]\end{aligned} \tag{335}$$

$$\begin{aligned}\leq&\sum_{i>a^*}\Delta_{C,i}+\Delta_{C,\max}\left(\frac{11}{\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{\Delta_j^2}\right)+\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\\&+\Delta_{C,i^*}\left(\max_{\Delta\in\{\Delta_{\min}\}\cup\{\Delta_{Q,j}\}_{j\leq a^*}}\left\{\frac{32\log\left(T\Delta^2\right)}{\Delta^2}\right\}+\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}+2\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\right)\\&+\max_{i>a^*,i\neq i^*}\Delta_{C,i}\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\text{ (linearity of expectation, total samples }T\text{).}\end{aligned} \tag{336}$$

Where $\Delta_{C,\max}=\max_{i\in[K]}\Delta_{C,i}$ is the largest cost gap among all arms. Equation 336 is the upper bound on Expected Cumulative Cost Regret stated in Theorem 3.2.

Proceeding identically, for Quality Regret we shall have,

$$\mathbb{E}\left[\text{Quality\_Reg}\left(T,\nu\right)\right] \tag{337}$$

$$=\sum_{i=1}^{K}\Delta_{Q,i}^{+}\mathbb{E}\left[n_i\left(T\right)\right] \tag{338}$$

$$=\sum_{i<a^*}\Delta_{Q,i}\mathbb{E}\left[n_i\left(T\right)\right]+\sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^{+}\mathbb{E}\left[n_i\left(T\right)\right] \tag{339}$$

$$<\sum_{i<a^*}\Delta_{Q,i}\left(1+\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}^2}+\frac{43}{\Delta_{Q,i}^2}+\mathbb{E}\left[n_i\left(T\right)\mid\beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\right) \tag{340}$$

$$+\sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^{+}\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}+\mathbb{E}\left[n_i\left(T\right)\mid\left\{Z>a^*\right\},\Gamma\right]\cdot\left(\frac{32}{T\Delta_{a^*}^2}+\frac{43}{T\Delta_{Q,a^*}^2}\right)\right.$$

$$\left.+\,\mathbb{E}\left[n_i\left(T\right)\mid\beta\right]\cdot\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\right)$$

$$\leq\sum_{i<a^*}\left(\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}+\frac{43}{\Delta_{Q,i}}\right)+\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)\Delta_{Q,\max}^{+}\sum_{i=1}^{K}\mathbb{E}\left[n_i(T)\mid\beta\right]$$

$$+\left(\frac{32}{T\Delta_{a^*}^2}+\frac{43}{T\Delta_{Q,a^*}^2}\right)\max_{i>a^*}\Delta_{Q,i}^{+}\sum_{i>a^*,i\neq i^*}\mathbb{E}\left[n_i(T)\mid\left\{Z>a^*\right\},\Gamma\right]+\sum_{i=1}^{K}\Delta_{Q,i}^{+} \tag{341}$$

$$+\sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^{+}\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2} \tag{342}$$

$$\leq\sum_{i<a^*}\left(\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}+\frac{43}{\Delta_{Q,i}}\right)+\sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^{+}\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}$$

$$+\max_{i>a^*}\Delta_{Q,i}^{+}\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)+\Delta_{Q,\max}^{+}\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)+\sum_{i=1}^{K}\Delta_{Q,i}^{+}. \tag{343}$$

Where Equation 343 is the upper bound on Expected Cumulative Quality Regret states in Theorem 3.2 and follows from the Linearity of the Expectation operator and the total sample budget being $T$. Similar to the description that followed the proof □

Similar to the description that followed the proof of Theorem 3.1, we have for Cost Regret,

$$\underbrace{\Delta_{C,i^*}\left(1+\max_{\Delta\in\{\Delta_{\min}\}\cup\{\Delta_{Q,j}\}_{j\leq a^*}}\left\{\frac{32\log\left(T\Delta^2\right)}{\Delta^2}\right\}\right)}_{\substack{\text{Contribution from } i^* \text{ under nominal}\\ \text{termination in PE-stage episode } a^*}}+\underbrace{\sum_{i>a^*,i\neq i^*}\Delta_{C,i}\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\right)}_{\substack{\text{Contribution from high-cost arms}\\ \text{with a proper end to the BAI-stage}}}$$

$$+\underbrace{\Delta_{C,\max}\left(\frac{11}{\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{\Delta_j^2}\right)}_{\substack{\text{Contribution from improper}\\ \text{end to BAI stage}}}+\underbrace{\Delta_{C,i^*}\left(\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}^2}+\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)\right)}_{\substack{\text{Contribution from } i^* \text{ under}\\ \text{mis-termination in PE-stage episode } \leq a^*}}$$

$$+\underbrace{\max_{i>a^*}\Delta_{C,i}\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)}_{\substack{\text{Contribution from PE-stage episodes } >a^*\\ \text{in case of mis-termination during ep } a^*}}.$$

And for Quality Regret,

$$\underbrace{\sum_{i=1}^{a^*-1}\left(\Delta_{Q,i}+\frac{32\log\left(T\Delta_{Q,i}^2\right)}{\Delta_{Q,i}}\right)}_{\substack{\text{Contribution from } i<a^* \text{ under nominal}\\ \text{termination in PE-stage episode } a^*}}+\underbrace{\sum_{i>a^*,i\neq i^*}\Delta_{Q,i}^+\left(1+\frac{32\log\left(T\Delta_i^2\right)}{\Delta_i^2}\right)}_{\substack{\text{Contribution from PE-stage episodes } >a^*\\ \text{in case of mis-termination during ep } a^*}}$$

$$+\underbrace{\Delta_{Q,\max}^+\left(\frac{11}{T\Delta_{\min}^2}+\sum_{j\neq i^*}\frac{32}{T\Delta_j^2}\right)}_{\substack{\text{Contribution from PE-stage episodes } >a^*\\ \text{in case of mis-termination during ep } a^*}}+\underbrace{\sum_{i=1}^{a^*-1}\frac{43}{\Delta_{Q,i}}}_{\substack{\text{Contribution from}\\ i<a^* \text{ under}\\ \text{mis-termination in}\\ \text{PE-stage episode } \leq a^*}}+\underbrace{\max_{i>a^*}\Delta_{Q,i}^+\left(\frac{32}{\Delta_{a^*}^2}+\frac{43}{\Delta_{Q,a^*}^2}\right)}_{\substack{\text{Contribution from PE-stage episodes } >a^*\\ \text{in case of mis-termination during ep } a^*}}.$$