NONSTATIONARY MULTI-ARMED BANDITS WITH HETEROGENEOUS ARM CLASSES

Anonymous authors
Paper under double-blind review

ABSTRACT

In real-world applications such as online recommendation, nonstationarity is rarely uniform: trending items fluctuate rapidly, while the broader catalog remains stable. Existing nonstationary bandit algorithms ignore this structure, incurring regret that scales with the global product of total arms and total switches. To address this, we formulate nonstationary multi-armed bandits with heterogeneous arm classes, where K arms are partitioned into M known classes of sizes K_m with individual change budgets L_m. We design an adaptive algorithm requiring no prior knowledge of these budgets that achieves expected dynamic regret of order $\sqrt{MT\sum_m (L_m+1)K_m}$, tightly pairing each class's volatility with only its own size. Finally, we establish complementary lower bounds showing that this rate is optimal up to logarithmic factors for fixed M, and that the $\sqrt{M}$ class-search overhead is fundamentally unavoidable in general.

1 INTRODUCTION

Multi-armed bandit (MAB) algorithms operating in nonstationary environments provide a foundational framework for sequential decision-making under uncertainty, with widespread applications in online recommendation, ad allocation, and dynamic pricing (Garivier & Moulines, 2008; Besbes et al., 2014). In such domains, the underlying reward distributions of the arms can undergo abrupt switches or continuous drifts over a horizon of T rounds. To protect against arbitrary distribution shifts, the classical literature characterizes nonstationarity through a *global change budget*, typically bounded by L total distribution switches across the entire pool of K arms. Under this formulation, existing minimax-optimal algorithms achieve an expected dynamic regret of $\Theta(\sqrt{(L+1)KT})$ against the sequence of optimal arms at each round (Auer et al., 2019; Chen et al., 2019).

Although mathematically elegant, measuring nonstationarity via a single global parameter L overlooks a critical reality: volatility is rarely uniform across actions. In online platforms, a small subset of trending or viral items undergoes rapid shifts in user preference, while the vast majority of the catalog remains essentially stationary. Conventional nonstationary bandit algorithms fail in such asymmetric settings because their change-detection or windowing mechanisms operate globally. A sudden distribution shift in a single viral item triggers a global reset across all arms, needlessly wiping out accumulated observations on stable catalog items.

Consequently, existing algorithms incur dynamic regret that scales with the global product of total arms K and total switches L. When frequent changes are confined to a tiny fraction of the action space, this forces the learner to pay for the entire catalog size at the worst-case switching rate. Ideally, an algorithm should isolate nonstationarity to its source, pairing each subset's volatility strictly with its own size. This asymmetry exposes a fundamental theoretical question: *can an online learner achieve regret that decouples localized change budgets from the global arm pool, and what is the unavoidable price of searching across heterogeneous classes?*

To address this challenge, we introduce the framework of *nonstationary multi-armed bandits with heterogeneous arm classes*. Over a horizon of T rounds, a learner interacts with a set of K arms partitioned into M known disjoint classes $\mathcal{A}_1, \ldots, \mathcal{A}_M$ with respective sizes $K_m = |\mathcal{A}_m|$. In practical systems, such partitions are readily available from domain metadata, which naturally differentiates volatile items (e.g., breaking news, seasonal promotions) from stable baselines (e.g., evergreen media, staple catalog goods). Rather than imposing a monolithic bound on environmental

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shifts, each class $\mathcal{A}_m$ is governed by an individual change budget L_m, denoting the maximum number of rounds in which reward distributions within class m undergo switches. This formulation captures localized volatility through the cumulative instance difficulty: $D = \sum_{m=1}^{M}(L_m + 1)K_m$. Whenever frequent switches are confined to relatively small subsets of arms, D can be orders of magnitude smaller than the standard global metric $(L + 1)K$, establishing a benchmark that reflects the true refined complexity of the environment.

To exploit this structure, we develop *Exponential Backfill for Heterogeneous Arm Classes* (`EBF-H`). The algorithm is built on two core principles: localized change detection and class-aware backfilling. Rather than resetting globally upon detecting a shift, `EBF-H` isolates resets strictly to the affected class, preserving sample history for stationary arms. Simultaneously, it uses a careful backfilling schedule to ensure that arms that temporarily appear sub-optimal are re-examined before nonstationarity makes them optimal again. We prove that `EBF-H` achieves expected dynamic regret of $\widetilde{O}(\sqrt{MTD})$, where $\widetilde{O}(\cdot)$ hides logarithmic factors, decoupling the classes up to an explicit factor of $\sqrt{M}$.

We complement this upper bound with two lower bounds confirming its near-optimality. First, a direct-sum bound of $\Omega(\sqrt{TD})$ proves that our rate is optimal up to logarithmic factors whenever M is fixed. Second, an $\Omega(\sqrt{MKT})$ lower bound demonstrates that the $\sqrt{M}$ factor is not an algorithmic artifact, but the fundamental cost of identifying which class currently contains the best arm.

Summary of contributions. In summary, this paper provides a near-complete characterization of nonstationary bandits under localized volatility. We introduce the class-structured nonstationary bandit framework, parameterizing asymmetric volatility via class sizes K_m and change budgets L_m, and defining the cumulative instance difficulty $D = \sum_{m=1}^{M}(L_m + 1)K_m$. We develop `EBF-H`, an algorithm achieving expected dynamic regret of $\widetilde{O}(\sqrt{MTD})$ that prevents shifts in volatile classes from contaminating stationary ones. In the supplementary material, we extend this guarantee to an adaptive algorithm (`AEBF-H`) that requires no prior knowledge of the budgets $\{L_m\}_{m=1}^{M}$. Finally, we prove a direct-sum lower bound of $\Omega(\sqrt{TD})$ establishing near-optimality for fixed M, alongside an $\Omega(\sqrt{MKT})$ bound confirming the $\sqrt{M}$ factor is the unavoidable price of searching across classes.

1.1 Related Work

Global nonstationarity: switching and variation budgets. Classical nonstationary bandit formulations characterize environmental shifts through a monolithic global budget, such as total switches L (Garivier & Moulines, 2008; Auer et al., 2019; Chen et al., 2019) or total variation V_T (Besbes et al., 2014). Because these metrics pool nonstationarity across all actions, a single shifting arm forces algorithms to reset or window observations globally, incurring dynamic regret of order $\sqrt{(L+1)KT}$. Our setting directly complements this line of research: it smoothly recovers the classical switching rate when $M = 1$, while offering a structural refinement when nonstationarity is localized, enabling learners to preserve statistical history in stable subsets of arms.

Reward structure vs dynamics structure. A rich literature accelerates exploration by exploiting dependencies across arms, including linear and contextual bandits (Li et al., 2010; Abbasi-Yadkori et al., 2011), general structured bandits (Combes & Proutiere, 2014), and latent or clustered bandits (Gentile et al., 2014; Maillard & Mannor, 2014). These frameworks operate fundamentally in the *reward space*, assuming correlated payoffs or shared latent parameters. Our work explores an orthogonal structural dimension in the *temporal space*: arms within a class share volatility budgets rather than reward values. Because arm means within a class remain independent and unconstrained, this formulation offers a complementary way to incorporate domain structure into online learning.

Specialized temporal dynamics: restless, rotting and corrupted bandits. Other important lines of work address structured temporal evolution under specific physical or behavioral dynamics. Restless and rotting bandits (Slivkins & Upfal, 2008; Seznec et al., 2020) model systems governed by continuous stochastic processes (e.g., Brownian motion) or wear-and-tear degradation where rewards diminish with pulls. Similarly, corrupted bandit models (Lykouris et al., 2018; Gupta et al., 2019) provide robust guarantees for predominantly stationary systems subject to bounded adversarial perturbations. Our piecewise-stationary class model complements these dynamics by capturing

abrupt, recurring regime shifts that occur at disparate frequencies across subclasses, without requiring Markovian, monotonic, or small-total-perturbation assumptions.

For additional related work see Section E.

2 PRELIMINARIES

Heterogeneous nonstationary bandits. We adopt the stochastic switching-bandit setup of Auer et al. (2019) and introduce a class-structured extension in which the arms are partitioned into known classes. The online game lasts T rounds and has a finite set of $K \geq 2$ arms, $\mathcal{A} = \{1, \ldots, K\}$. The arms are partitioned into M known, nonempty, disjoint classes $\mathcal{A}_1, \ldots, \mathcal{A}_M$, where $\bigcup_{m=1}^{M} \mathcal{A}_m = \mathcal{A}$. Class m contains $K_m := |\mathcal{A}_m|$ arms, so $K = \sum_{m=1}^{M} K_m$. For each arm a, $\mathrm{class}(a)$ denotes the unique class index m such that $a \in \mathcal{A}_m$. Arm a at round t has reward $R_{a,t} \in [0, 1]$. Writing $\mathbb{E}$ for expectation, its mean is $\mu_a(t) := \mathbb{E}[R_{a,t}]$. For each class m, define its realized set of change rounds by

$$\mathcal{L}_m := \{t \in \{2, \ldots, T\} : \exists a \in \mathcal{A}_m \text{ such that } \mu_a(t) \neq \mu_a(t-1)\}.$$

We assume $|\mathcal{L}_m| \leq L_m$ for a class-specific budget $L_m \in \{0, \ldots, T-1\}$. A round is counted once for class m whenever the mean of at least one arm in that class changes, even if several means change simultaneously. These change rounds partition $\{1, \ldots, T\}$ into $|\mathcal{L}_m| + 1 \leq L_m + 1$ maximal intervals on which every mean in class m is constant; we call them the *true class-m stationary segments*. Within each maximal interval on which an individual arm's mean is constant, its rewards are independent and identically distributed (i.i.d.) from a fixed law; these intervals are the arm's stationary segments. Thus $M = 1$ recovers the classical switching-bandit model. The budgets may be conservative upper bounds on the realized numbers of changes.

Online interaction protocol. Before interaction begins, the environment fixes the entire mean sequence $\{\mu_a(t)\}_{a \in \mathcal{A},\ t=1,\ldots,T}$. The reward variables are mutually independent and satisfy $\mathbb{E}[R_{a,t}] = \mu_a(t)$. At round t, the learner chooses an arm a_t using only its internal randomness and the actions and rewards observed before round t, and then observes only $R_{a_t,t}$. The learner knows T and the class partition. Algorithm `EBF-H` also receives $L_1, \ldots, L_M$, whereas `AEBF-H` does not receive these budgets. Fixing the mean sequence before interaction ensures that it cannot react to the learner's random choices. We call any instance satisfying the preceding conditions *admissible*.

Dynamic regret. Fix a total order on the arms and choose the first maximizing arm in this order whenever a tie occurs. Define the optimal arm at round t by $a_t^* \in \arg\max_{a \in \mathcal{A}} \mu_a(t)$. The learner's dynamic regret is

$$\mathrm{Regret}_T := \sum_{t=1}^{T} \big(\mu_{a_t^*}(t) - \mu_{a_t}(t)\big).$$

Unlike static regret, this benchmark allows the optimal arm to change over time. Our results bound $\mathbb{E}[\mathrm{Regret}_T]$, where the expectation is over the reward variables and the learner's randomization.

Additional notation. For an integer $n \geq 0$, write $[n] := \{1, \ldots, n\}$, with $[0] := \emptyset$; for integers $s \leq t$, write $[s, t] := \{s, s+1, \ldots, t\}$, with $[t, t-1] := \emptyset$. $|S|$ denotes the cardinality of finite set S, $\mathbf{1}\{E\}$ denotes the indicator of event E, and $\mathbb{P}$ denotes probability. Unless stated otherwise, sums indexed by m range over $[M]$. Finally, we define the total class difficulty by $D := \sum_{m=1}^{M} (L_m + 1) K_m$.

3 THE EBF-H ALGORITHM AND MAIN RESULTS

In this section, we present *Exponential Backfill for Heterogeneous Arm Classes* (`EBF-H`). The primary challenge in heterogeneous nonstationary environments is preventing high-frequency distribution shifts in small, volatile classes from degrading learning progress in stable ones. `EBF-H` resolves this by decoupling change detection across classes while actively monitoring dormant arms through a multi-scale exponential backfill schedule.

Interval statistics and confidence bounds. Because global averages can obscure localized changes by mixing old and new distributions, EBF−H evaluates statistics on sub-intervals within each class's current algorithmic phase. For each class $m \in [M]$, the algorithm tracks a phase start round $\phi_m \in [T]$, initialized to $\phi_m = 1$. For an arm $a \in \mathcal{A}_m$ and a consecutive sub-interval $I \subseteq [\phi_m, t-1]$, we define the pull count $N_a(I) = \sum_{\tau \in I} \mathbf{1}\{a_\tau = a\}$, empirical mean $\hat{\mu}_a(I) = \frac{\sum_{\tau \in I} R_{a,\tau} \mathbf{1}\{a_\tau = a\}}{\max\{N_a(I),1\}}$, confidence radius $\mathrm{CB}_a(I) = \sqrt{\frac{8\log(2TK)}{\max\{N_a(I),1\}}}$, and empirical bounds $\hat{\mu}_a^{(\pm q)}(I) = \hat{\mu}_a(I) \pm q\,\mathrm{CB}_a(I)$. With no observations ($I = \emptyset$), we set $\hat{\mu}_a(\emptyset) = 0$ and $\mathrm{CB}_a(\emptyset) = \sqrt{8\log(2TK)}$.

Mechanism 1: class-isolated change detection. At the start of each round t, EBF−H tests each class $m \in [M]$ for distribution shifts independently. Specifically, it checks whether there exists an arm $a \in \mathcal{A}_m$ and consecutive intervals $I_1, I_2 \subseteq [\phi_m, t-1]$ whose confidence intervals are disjoint:

$$\hat{\mu}_a^{(-1)}(I_1) > \hat{\mu}_a^{(+1)}(I_2). \tag{1}$$

If condition equation 1 is met, a distribution shift is declared in class m, triggering an isolated reset: $\phi_m \leftarrow t$. Crucially, observations collected for arms in all other classes $m' \neq m$ are strictly preserved, preventing cross-class contamination.

Mechanism 2: arm classification and dominance gaps. After evaluating resets, the algorithm classifies each arm $a \in \mathcal{A}$ as either *good* ($\mathcal{G}_t$) or *bad* ($\mathcal{B}_t$). An arm a is declared bad if another arm $a' \neq a$ provides statistical evidence of superiority across valid phase intervals $I \subseteq [\phi_{\mathrm{class}(a)}, t-1]$ and $I' \subseteq [\phi_{\mathrm{class}(a')}, t-1]$:

$$\hat{\mu}_{a'}^{(-3)}(I') > \hat{\mu}_a^{(+3)}(I). \tag{2}$$

Let $X_{t,a}$ be the set of all triples (a', I, I') satisfying equation 2. If $X_{t,a} \neq \emptyset$, $a \in \mathcal{B}_t$; otherwise, $a \in \mathcal{G}_t$. The conservative margin of 3 times the confidence radii guarantees that the blocking arm in a maximizing certificate,

$$(\bar{a}_t(a), I_t(a), \overline{I}_t(a)) := \operatorname*{arg\,max}_{(a',I,I') \in X_{t,a}} \left\{ \hat{\mu}_{a'}^{(-1)}(I') - \hat{\mu}_a^{(+1)}(I) \right\},$$

is provably good (Proposition 4.3). We record this maximum gap as the *dominance gap* $\Delta_t(a) := \max_{(a',I,I') \in X_{t,a}} \{\hat{\mu}_{a'}^{(-1)}(I') - \hat{\mu}_a^{(+1)}(I)\}$ and define the log scale $i_t(a) := \lceil -\log_2 \Delta_t(a) \rceil$.

Mechanism 3: Class-Aware Exponential Backfill. When an arm's reward distribution changes, it may secretly become optimal while outdated observations from before the shift continue to certify it as bad. If the learner completely ignored bad arms, it would never collect fresh samples to detect this improvement. Conversely, pulling bad arms indiscriminately would incur intolerable regret. To balance this trade-off, EBF−H uses *exponential backfill*, maintaining a set of monitoring counters $s_{a,j}$ that act as allocated quotas of exploratory pulls.

Each counter $s_{a,j}$ corresponds to a target gap resolution 2^{-j} indexed by log scales $j \in \{0, \ldots, J_{\max}\}$, where $J_{\max} := \lceil \log_2 T \rceil$. A positive counter value $s_{a,j} > 0$ represents a pending request to inspect arm a with enough samples to resolve a gap of size 2^{-j}. At each round, each counter is refreshed to a full quota of $2^{2j+5}\Lambda$ pulls (where $\Lambda := \lceil \log(2TK) \rceil$) with independent probability:

$$2^{-j} \sqrt{\frac{L_{\mathrm{class}(a)} + 1}{TK_{\mathrm{class}(a)}}}.$$

This refresh rule captures the core statistical trade-off of monitoring:

- *Resolving finer gaps requires more pulls:* By Hoeffding's inequality, distinguishing a gap of size 2^{-j} requires $O(2^{2j})$ observations, which is why the allocated quota quadruples with each increase in scale j.
- *Finer checks are launched less frequently:* To keep the expected monitoring regret bounded, the probability of initiating a check at scale j halves with each step (2^{-j}).
- *Rates are calibrated to class volatility:* The refresh probability scales with $\sqrt{(L_m + 1)/K_m}$, ensuring that arms in volatile classes are monitored frequently while arms in stable classes are queried sparingly.

Algorithm 1: EBF−H: Exponential Backfill for Heterogeneous Arm Classes

Input: Horizon T, classes $\mathcal{A}_1, \dots, \mathcal{A}_M$, budgets $L_1, \dots, L_M$.

Set $J_{\max} \leftarrow \lceil \log_2 T \rceil$, $\Lambda \leftarrow \lceil \log(2TK) \rceil$, $\phi_m \leftarrow 1$ for all $m \in [M]$, and $s_{a,j} \leftarrow 0$.
for $t = 1, \dots, T$ **do**
  // 1. Class-Isolated Change Detection
  foreach $m \in [M]$ **do**
    if $\exists a \in \mathcal{A}_m, I_1, I_2 \subseteq [\phi_m, t-1]$ *s.t.* $\hat{\mu}_a^{(-1)}(I_1) > \hat{\mu}_a^{(+1)}(I_2)$ **then**
      $\phi_m \leftarrow t$.
    end
  end
  // 2. Arm Classification (Details in Appendix A, Algorithm 2)
  $(\mathcal{G}_t, \mathcal{B}_t, \{\Delta_t(a), i_t(a)\}_{a \in \mathcal{B}_t}) \leftarrow \texttt{ClassifyArms}(t, \phi_1, \dots, \phi_M)$.
  // 3. Exponential Backfill Refresh
  foreach $a \in \mathcal{A}, j \in \{0, \dots, J_{\max}\}$ **do**
    With prob. $2^{-j}\sqrt{(L_{\text{class}(a)}+1)/(TK_{\text{class}(a)})}$, set $s_{a,j} \leftarrow 2^{2j+5}\Lambda$.
  end
  // 4. Two-Stage Action Selection
  $\mathcal{E}_t \leftarrow \mathcal{G}_t \cup \{a \in \mathcal{B}_t : \exists j \le i_t(a) \text{ s.t. } s_{a,j} > 0\}$.
  Draw m_t uniformly from $\{m : \mathcal{A}_m \cap \mathcal{E}_t \neq \emptyset\}$; draw j_t uniformly from $\{0, \dots, J_{\max}\}$.
  Play $a_t \leftarrow \arg\min_{a \in \mathcal{A}_{m_t} \cap \mathcal{E}_t} \tau_a^{j_t}$, observe $R_{a_t,t}$, and update histories.
  if $a_t \in \mathcal{B}_t$ **then**
    foreach $j \in \{0, \dots, \min\{i_t(a_t), J_{\max}\}\}$ **do**
      $s_{a_t,j} \leftarrow \max\{0, s_{a_t,j} - 1\}$.
    end
  end
end

Action Selection Policy. Armed with the classified arms and monitoring quotas, the learner decides which arm to sample at round t. This process involves three stages:

1. **Forming the eligible pool:** All good arms ($\mathcal{G}_t$) have not been ruled out by evidence and are always eligible for selection. In contrast, a bad arm $a \in \mathcal{B}_t$ is eligible only if it has an active monitoring request ($s_{a,j} > 0$) at a scale no finer than its estimated gap ($j \le i_t(a)$):
$$\mathcal{E}_t := \mathcal{G}_t \cup \big\{a \in \mathcal{B}_t : \exists j \le i_t(a) \text{ such that } s_{a,j} > 0\big\}. \tag{3}$$
The constraint $j \le i_t(a)$ is crucial: it prevents a large quota intended for a very subtle gap (large j) from being wasted on an arm with a massive performance deficit (small $i_t(a)$).
2. **Class selection:** The learner first samples a class m_t uniformly at random among all classes that contain at least one eligible arm. Drawing classes uniformly prevents large catalog classes from drowning out small, highly volatile classes, guaranteeing that volatile arms receive sufficient monitoring opportunities regardless of catalog size.
3. **Arm selection:** Within the selected class m_t, the learner chooses which eligible arm to pull. Rather than pulling greedily, the algorithm samples a random scale $j_t \in \{0, \dots, J_{\max}\}$ and identifies how recently each candidate received its 2^{j_t}-th most recent pull in the current phase, denoted $\tau_a^{j_t}$ (setting $\tau_a^{j_t} = -\infty$ if it has fewer pulls). It then plays the arm whose observations are least recent: $a_t \leftarrow \arg\min_{a \in \mathcal{A}_{m_t} \cap \mathcal{E}_t} \tau_a^{j_t}$. This priority mechanism operates like a multi-scale round-robin, ensuring that no eligible arm is neglected and that confidence intervals shrink evenly across the class.

Updating Counters. After playing a_t and observing its reward $R_{a_t,t}$, the algorithm updates its empirical records. If the played arm was bad, it decrements its applicable active counters $s_{a_t,j}$ by one for all $0 \le j \le \min\{i_t(a_t), J_{\max}\}$, gradually exhausting its monitoring quota until the check concludes. Pulls of good arms leave the counters unchanged.

The complete execution of EBF−H is summarized in Algorithm 1. By maintaining prefix sums of pull counts and rewards, each interval statistic is computable in $O(1)$ time, yielding an overall runtime of $O(K^2T^5)$ and $O(KT)$ memory. The following theorem establishes our main dynamic regret guarantee.

Theorem 3.1 (Dynamic regret bound of EBF−H)**.** *For every admissible nonstationary environment, the expected dynamic regret of EBF−H satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le 4096 \log^2(2TK)\sqrt{MTD} + 1 = \widetilde{O}(\sqrt{MTD}).$$

Extension to unknown budgets. While EBF−H takes the budgets $\{L_m\}_{m=1}^M$ as input to calibrate its backfill probabilities, in many realistic settings these volatility budgets are unknown. In Appendix D, we show that this requirement can be completely eliminated. We develop *Adaptive Exponential Backfill for Heterogeneous Arm Classes* (AEBF−H), an algorithm that takes only the horizon T and the class partition as inputs. Rather than relying on a static budget L_m, AEBF−H dynamically scales the monitoring rate of class m based on the empirical number of resets detected in that class up to the current round. The proof separates classification and monitoring observations and chooses monitoring candidates independently of obligations. This preserves launch-seed independence. Bounding launch costs by detected resets and summing waiting times across detected phases gives the identical leading dynamic regret bound:

$$\mathbb{E}[\mathrm{Regret}_T] = \widetilde{O}(\sqrt{MTD}).$$

To keep our main presentation focused on the core statistical mechanics, we concentrate on EBF−H in the main text and provide the full pseudocode and analysis of AEBF−H in Appendix D.

4 Dynamic Regret Analysis

In this section, we prove Theorem 3.1 by developing the proof architecture for EBF−H. We establish that the algorithm's localized resets strictly isolate nonstationarity, and we decompose the dynamic regret into three distinct statistical penalties. Full proofs of all lemmas are deferred to Appendix B.

4.1 Foundations: Concentration, Local Resets, and Representatives

We begin by establishing uniform confidence bounds and bounding the number of class resets. For a stationary interval I, let $\mu_a(I)$ denote the common mean of arm a on I. The following lemma gives a uniform concentration bound over all such intervals.

Lemma 4.1 (Uniform interval concentration)**.** *Let $\mathcal{C}$ be the presampled-block event of Appendix B.2, independent of the algorithm's coins. Then $\mathbb{P}(\mathcal{C}) \ge 1 - T^{-1}$, and on $\mathcal{C}$, $|\hat{\mu}_a(I) - \mu_a(I)| \le \mathrm{CB}_a(I)$ for every arm a and nonempty interval I within one stationary segment.*

Lemma 4.1 follows from Hoeffding's inequality applied to pre-sampled rewards combined with a union bound over all fixed sequence blocks. On $\mathcal{C}$, the algorithm's change-detection test is strictly conservative:

Lemma 4.2 (Resets are paid by true changes)**.** *On the concentration event $\mathcal{C}$, class m is reset at most L_m times, dividing the horizon for class m into at most $L_m + 1$ algorithmic phases.*

If no change occurs in class m during its active phase, the true mean remains identical across all sub-intervals, so their confidence intervals must overlap on $\mathcal{C}$. Hence, every reset requires a genuine distribution shift in class m. Next, recall that each bad arm $a \in \mathcal{B}_t$ is associated with a blocking arm $\bar{a}_t(a)$ maximizing the certificate separation. The following structural result ensures that our reference arm is always trustworthy:

Proposition 4.3 (Blocking arms are good)**.** *For every round $t \in [T]$ and every arm $a \in \mathcal{B}_t$, the blocking arm satisfies $\bar{a}_t(a) \in \mathcal{G}_t$.*

If $\bar{a}_t(a)$ were bad, another arm would certify it as inferior; the $3\times$ radius margin in equation 2 would then imply that this third arm certifies a with a strictly larger gap, contradicting the maximality of $\bar{a}_t(a)$.

4.2 The Three-Term Regret Decomposition

Writing $\mu^*(t) := \mu_{a_t^*}(t)$, we assign a *representative* to both the played arm a_t and the instantaneous optimal arm a_t^*:

$$\hat{a}_t := \begin{cases} a_t & \text{if } a_t \in \mathcal{G}_t, \\ \bar{a}_t(a_t) & \text{if } a_t \in \mathcal{B}_t, \end{cases} \qquad \hat{a}_t^* := \begin{cases} a_t^* & \text{if } a_t^* \in \mathcal{G}_t, \\ \bar{a}_t(a_t^*) & \text{if } a_t^* \in \mathcal{B}_t. \end{cases}$$

By Proposition 4.3, both $\hat{a}_t$ and $\hat{a}_t^*$ are guaranteed to be good arms. Adding and subtracting their expected rewards yields the exact instantaneous decomposition:

$$\mu^*(t) - \mu_{a_t}(t) = \underbrace{\mu^*(t) - \mu_{\hat{a}_t^*}(t)}_{\textbf{Term (iii): delayed detection}} + \underbrace{\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)}_{\textbf{Term (ii): comparison good arms}} + \underbrace{\mu_{\hat{a}_t}(t) - \mu_{a_t}(t).}_{\textbf{Term (i): monitoring bad arms}} \tag{4}$$

The terms can be signed; each isolates a distinct statistical challenge:

- **Term (i) (monitoring bad arms):** Measures the cost of exploring an arm currently labeled bad. It is zero when $a_t \in \mathcal{G}_t$, and otherwise bounded by its dominance gap and confidence corrections.
- **Term (ii) (comparing good arms):** Measures the regret between two good representatives. Because neither arm has been ruled out by evidence, their gap is controlled by their remaining confidence radii.
- **Term (iii) (delayed detection):** Measures the loss when the optimal arm a_t^* is temporarily misclassified as bad due to outdated observations from before a recent change.

4.3 Controlling the Regret Components

We bound the sums conditional on $\mathcal{C}$. Here $\mathrm{CB}_a(t)$ uses observations before t since the latest true change or reset of a's class (its current MSI).

Term (i): monitoring bad arms. Term (i) is zero for a good played arm. Otherwise, the certificate and no-reset comparisons give $\mu_{\hat{a}_t}(t) - \mu_{a_t}(t) \le 2\mathrm{CB}_{\hat{a}_t}(t) + 3\Delta_t(a_t) + 2\mathrm{CB}_{a_t}(t)$.

Lemma 4.4 (Bounding Term (i))**.** *On the concentration event $\mathcal{C}$, the cumulative cost of monitoring bad arms satisfies:* $\mathbb{E}\Big[\sum_{t=1}^{T}\big(\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)\big) \mid \mathcal{C}\Big] \le \widetilde{O}(\sqrt{MTD})$.

Proof intuition. Every bad-arm pull charges $3\Delta_t(a_t) \le 6 \cdot 2^{-j}$ to the latest refresh of one active counter $s_{a_t,j}$ with $j \le i_t(a_t)$. A scale-j refresh receives at most $2^{2j+5}\Lambda$ such charges, costing $O(2^j\Lambda)$. Multiplying by its refresh probability $2^{-j}\sqrt{(L_m+1)/(TK_m)}$ cancels 2^j. Summing across scales, arms, and rounds bounds the dominance-gap contribution. Eligible-arm and played-arm confidence sums control the two remaining terms (Appendix B.5.1).

Term (ii): comparing good arms. Term (ii) compares the two good representatives $\hat{a}_t^*$ and $\hat{a}_t$.

Lemma 4.5 (Bounding Term (ii))**.** *On the concentration event $\mathcal{C}$, the cumulative regret between good representatives satisfies:* $\mathbb{E}\Big[\sum_{t=1}^{T}\big(\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)\big) \mid \mathcal{C}\Big] \le \widetilde{O}(\sqrt{MTD})$.

Proof intuition. Because both representatives remain good, the three-radius certificate test and concentration on their current MSI histories give $\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t) \le 4\mathrm{CB}_{\hat{a}_t^*}(t) + 4\mathrm{CB}_{\hat{a}_t}(t)$. Summing these confidence radii over the horizon is controlled by the rate at which eligible arms are sampled under our two-stage policy (Appendix B.5.2).

Term (iii): delayed change detection. Term (iii) is non-zero only when the optimal arm a_t^* is misclassified as bad due to stale observations.

Lemma 4.6 (Bounding Term (iii))**.** *On the concentration event $\mathcal{C}$, the cumulative regret from delayed detection satisfies:* $\mathbb{E}\Big[\sum_{t=1}^{T}\big(\mu^*(t) - \mu_{\hat{a}_t^*}(t)\big) \mid \mathcal{C}\Big] \le \widetilde{O}(\sqrt{MTD})$.

Proof intuition. For a bad optimal arm, use the certificate residual $r_t := \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))$, not its regret gap. The no-reset comparison gives Term (iii)$\le r_t - \Delta_t(a_t^*) + 2\mathrm{CB}_{\hat{a}_t^*}(t)$. After bounding the confidence and class-threshold terms, consider hard rounds with $r_t > \Delta_t(a_t^*)$ and $2^{-j} < r_t \le 2^{1-j}$. Then $j \le i_t(a_t^*)$. On a fixed true class-m stationary segment, the wait for its fixed best arm's first scale-j refresh has expectation at most $2^j\sqrt{TK_m/(L_m+1)}$. Multiplication by 2^{1-j} bounds the waiting contribution. After refresh, exhausting the counter on such a hard round would force a reset; otherwise the arm is eligible and $r_t \le 2\mathrm{CB}_{a_t^*}(t)$. A reset eliminates positive residuals for the rest of that segment. Summing over segments and scales, and adding at most one for residuals below $2^{-J_{\max}}$, proves the claim (Appendix B.5.3).

4.4 Synthesis: Summing over Maximal Static Intervals

To complete the proof of Theorem 3.1, we bound the cumulative sum of confidence radii across the horizon. For each class $m \in [M]$, we partition the horizon into *Maximal Static Intervals* (MSIs). A new MSI begins whenever: the reward distribution of some arm in $\mathcal{A}_m$ changes (at most L_m times), or class m undergoes a change-detection reset (at most L_m times on $\mathcal{C}$ by Lemma 4.2). Thus, class m is partitioned into at most $2L_m + 1$ MSIs, during which both the true means and the phase start time ϕ_m remain completely fixed. Across all M classes, the total number of arm-interval pairs $(a, \mathcal{I})$ is bounded by:

$$\sum_{m=1}^{M} (2L_m + 1)K_m \leq 2 \sum_{m=1}^{M} (L_m + 1)K_m = 2D.$$

Within each MSI $\mathcal{I}$ where arm a is pulled $n_{a,\mathcal{I}}$ times. Standard confidence-sum bounds (Lattimore & Szepesvári, 2020) guarantee that the sum of radii over pulls is bounded by $O(\sqrt{\log(TK) n_{a,\mathcal{I}}})$. Applying the Cauchy–Schwarz inequality across all arm-interval pairs yields:

$$\sum_{m=1}^{M} \sum_{\mathcal{I}} \sum_{a \in \mathcal{A}_m} \sqrt{n_{a,\mathcal{I}}} \leq \sqrt{\left(\sum_{m=1}^{M} (2L_m + 1)K_m \right) \left(\sum_{m=1}^{M} \sum_{\mathcal{I}} \sum_{a \in \mathcal{A}_m} n_{a,\mathcal{I}} \right)} \leq \sqrt{2D \cdot T}.$$

Origin of the $\sqrt{M}$ overhead. For the widest-confidence eligible arm, the recency rule supplies a same-class comparison arm of at least half its radius, played with probability at least $p = 1/[M(J_{\max} + 1)]$. Each arm–MSI pair advances from a given pull count only once. With at most $2D$ pairs, the expected number of rounds at any comparison-arm count is at most $2D/p$. Splitting the confidence sum at a count threshold gives $O(\sqrt{DT \log(2TK)/p})$ (Lemma B.9), yielding the final bound: $\mathbb{E}[\mathrm{Regret}_T \mid \mathcal{C}] = \widetilde{O}(\sqrt{MTD})$.

Since $\mathbb{P}(\mathcal{C}^c) \leq T^{-1}$ and regret is bounded by T, the failure event contributes at most $T \cdot T^{-1} = 1$ in expectation, concluding the proof of Theorem 3.1.

5 Lower Bound

In this section, we establish the fundamental information-theoretic limits of nonstationary bandits with heterogeneous arm classes. Our upper bound reflects two distinct statistical challenges:

1. *Within-class re-learning:* The sample cost of identifying newly optimal arms within each class following localized distribution shifts.
2. *Cross-class search:* The exploration overhead of identifying which of the M classes currently contains the globally optimal arm.

Below, we present two complementary lower bounds that isolate these two phenomena. Both results hold for randomized learners, even when the algorithm has full prior knowledge of the class budgets $\{L_m\}_{m=1}^{M}$.

Within-class difficulty: the direct-sum bound. Our first result applies to arbitrary class configurations and demonstrates that the cumulative instance difficulty D cannot be circumvented.

Theorem 5.1 (Direct-Sum Lower Bound)**.** *Fix any collection of class sizes $\{K_m\}_{m=1}^{M}$ and change budgets $\{L_m\}_{m=1}^{M}$. If the horizon satisfies $T \geq 2D$, then for any policy, there exists an admissible nonstationary environment such that $\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{256}\sqrt{TD}$.*

Intuition. Theorem 5.1 establishes that localized nonstationarity is fundamentally additive: an algorithm cannot amortize the cost of changes in one class using samples gathered from another. If $D \leq 4K$, a stationary K-arm instance suffices. Otherwise, class m receives $\lfloor L_m/2 \rfloor$ independent learning blocks, each bracketed by baseline rounds and using two changes; their total weight satisfies $\sum_m \lfloor L_m/2 \rfloor K_m \geq D/4$. By tuning the sub-interval lengths proportionally to class sizes and calibrating the mean gaps to bound the Kullback–Leibler (KL) divergence of the history, Pinsker's inequality prevents the learner from reliably identifying the best arm in each epoch, yielding the additive rate $\sqrt{TD}$.

Cross-class exploration: the price of class search. While Theorem 5.1 captures within-class learning, it does not account for the uncertainty of locating an active class when optimal arms migrate across the partition. Our second result isolates this cross-class search cost in balanced environments with two changes per class.

Theorem 5.2 (Class-Search Lower Bound)**.** *Let K be divisible by M, and suppose all classes are balanced with $K_m = K/M$ with change budgets $L_m = 2$ for all $m \in [M]$. If $T \geq 6MK$, then for any policy, there exists an admissible environment such that $\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{512}\sqrt{MKT}$.*

Intuition. In this construction, M active macro-epochs lie between initial and terminal baseline blocks. In each epoch, a single class is chosen according to a uniformly random latent permutation of $[M]$, and a single arm within that class is temporarily elevated above the baseline. Even if an oracle reveals the active class and optimal arm in hindsight at the end of each macro-epoch, the identities of all future active classes remain uniformly distributed among the unvisited classes. Consequently, the learner is forced to repeatedly search over $\Theta(M)$ candidate classes, incurring an unavoidable $\sqrt{M}$ exploration penalty that scales the regret up to $\Omega(\sqrt{MKT})$.

Tightness and Minimax Implications. Together, Theorems 5.1 and 5.2 provide a comprehensive characterization of the minimax regret landscape:

- **Near-optimality for fixed M:** For any fixed number of classes ($M = O(1)$), our upper bound $\widetilde{O}(\sqrt{MTD})$ matches the direct-sum lower bound $\Omega(\sqrt{TD})$ up to logarithmic factors across all horizon regimes $T \geq 2D$.
- **Necessity of the $\sqrt{M}$ class-search penalty:** In the balanced setting of Theorem 5.2, the difficulty parameter evaluates to $D = \sum_{m=1}^{M}(2+1)(K/M) = 3K$. On this family of instances, our upper bound evaluates to $\widetilde{O}(\sqrt{MKT})$, precisely matching the lower bound of Theorem 5.2 up to logarithmic factors. This resolves the status of the $\sqrt{M}$ factor: it is not an algorithmic defect of `EBF-H`, but an unavoidable class-search cost on this balanced family.

Proof Architecture. Both lower bounds rely on a standard information-theoretic change-of-measure argument for bandits (Lattimore & Szepesvári, 2020). We construct a prior distribution over candidate mean vectors and evaluate the trajectory of the learner under a stationary null baseline versus alternative hypotheses where a specific candidate arm is shifted by Δ. The KL divergence between the resulting path measures is bounded by the expected number of pulls allocated to the perturbed arm. In the direct-sum bound, independence across classes allows the conditional regrets to sum additively across blocks via the tower property of expectation. In the class-search bound, the uniform prior over future class permutations ensures that the conditional indistinguishability holds at every macro-epoch, regardless of past feedback. Complete proofs are provided in Appendix C.

6 Discussion and Future Work

This work introduced nonstationary multi-armed bandits with heterogeneous arm classes to model environments with localized volatility. We proposed `EBF-H`, an algorithm achieving expected dynamic regret of order $\sqrt{MTD}$ by confining distribution shifts strictly to their source classes. Complementary lower bounds establish that this guarantee is optimal up to logarithmic factors for fixed M, while confirming that $\sqrt{M}$ factor is an unavoidable cost of searching across classes.

Several compelling directions remain open. First, while our upper and lower bounds match for fixed M, pinning down the exact minimax rate when M grows large in highly imbalanced regimes remains an open theoretical question. Second, relaxing the assumption of a known partition to learn latent volatility classes online presents a significant challenge. Third, designing lighter change-detection approximations, such as grid-based or sliding-window change-detection tests, would substantially reduce per-round computational overhead for practical large-scale deployments. Finally, extending localized volatility to linear and contextual bandits offers a natural path toward broader real-world applications.

AI USE STATEMENT

Generative AI assistance was used to check and develop mathematical arguments, verify references, and draft and revise the exposition. The authors are responsible for the final content.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

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

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158. PMLR, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pp. 521–529. PMLR, 2014.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2250–2259. PMLR, 2020.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International conference on machine learning*, pp. 757–765. PMLR, 2014.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Bowen Jiang, Bo Jiang, Jian Li, Tao Lin, Xinbing Wang, and Chenghu Zhou. Online restless bandits with unobserved states. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15041–15066. PMLR, 23–29 Jul 2023.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pp. 1739–1776. PMLR, 2018.

Thodoris Lykouris, Vahab S. Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122. ACM, 2018. doi: 10.1145/3188745.3188918.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pp. 136–144. PMLR, 2014.

Alexander Rakhlin and Karthik Sridharan. BISTRO: an efficient relaxation-based method for contextual bandits. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, JMLR Workshop and Conference Proceedings, pp. 1977–1985. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/rakhlin16.html.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12017–12026, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/263fc48aae39f219b4c71d9d4bb4aed2-Abstract.html.

Julien Seznec, Pierre Menard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3784–3794. PMLR, 2020.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pp. 343–354, 2008.

Siwei Wang, Longbo Huang, and John Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33:11878–11889, 2020.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pp. 4300–4354. PMLR, 2021.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 26865–26894. PMLR, 2022.

A Additional Algorithms

Algorithm 2: `ClassifyArms`$(t, \phi_1, \ldots, \phi_M)$

Input: Current time t, phase start times $\phi_1, \ldots, \phi_M$, and interval statistics from Section 3 computed from the rewards before t.

Output: Good arms $\mathcal{G}_t$, bad arms $\mathcal{B}_t$, and each bad arm's blocking arm, certificate intervals, dominance gap, and scale.

Initialize $\mathcal{B}_t \leftarrow \emptyset$ `// All intervals below are consecutive round intervals; all ties use the fixed rule`

foreach *arm* $a \in \mathcal{A}$ **do**

  $X_{t,a} \leftarrow \{(a', I, I') : a' \in \mathcal{A} \setminus \{a\}, I \subseteq [\phi_{\text{class}(a)}, t-1], I' \subseteq [\phi_{\text{class}(a')}, t-1], \hat{\mu}_{a'}^{(-3)}(I') > \hat{\mu}_a^{(+3)}(I)\}$

  if $X_{t,a} = \emptyset$ **then**

    continue `// a is good`

  end

  $\mathcal{B}_t \leftarrow \mathcal{B}_t \cup \{a\}$

  $(\bar{a}_t(a), I_t(a), \overline{I}_t(a)) \leftarrow \arg\max_{(a', I, I') \in X_{t,a}} \left\{\hat{\mu}_{a'}^{(-1)}(I') - \hat{\mu}_a^{(+1)}(I)\right\}$

  $\Delta_t(a) \leftarrow \max_{(a', I, I') \in X_{t,a}} \{\hat{\mu}_{a'}^{(-1)}(I') - \hat{\mu}_a^{(+1)}(I)\}$

  $i_t(a) \leftarrow \lceil -\log_2(\Delta_t(a)) \rceil$

end

$\mathcal{G}_t \leftarrow \mathcal{A} \setminus \mathcal{B}_t$

return the good and bad arms and the stored certificates, gaps, and scales

B Proofs for Sections 3 and 4

Our goal is to prove Theorem 3.1, the dynamic-regret upper bound for Algorithm `EBF-H` proving that

$$\mathbb{E}[\text{Regret}_T] \le 4096 \log^2(2TK)\sqrt{MTD} + 1, \qquad \text{where } D = \sum_{m=1}^{M} (L_m + 1)K_m.$$

Our following analysis distinguishes intervals on which a class's means remain constant from intervals containing a true rewards change. On the former, concentration bounds define the difference between empirical and true means; on the latter, reward observations collected before a change may no longer represent the current means. We therefore partition each class's timeline at its true change rounds and further split the resulting stationary segments at algorithmic resets. This gives the *maximal static intervals* (MSIs), within which both the class means and its phase start remain fixed. Uniform concentration on stationary intervals then ensures that a reset cannot occur unless a true change has occurred since the preceding phase began. Each reset can consequently be charged to a distinct true change, bounding the number of MSIs in terms of the class change budget.

We next show that every bad arm has a good blocking arm, which provides a representative for comparisons even when the played or optimal arm is classified as bad. Using the MSI partition, we bound the cumulative confidence radii of played arms and then extend this bound to eligible arms through the randomized class–scale selection rule. These bounds control the estimation-error contributions in the regret decomposition.

What remains then is the cost of monitoring bad arms and the delay caused by certificates that still contain pre-change observations. We charge monitoring pulls to counter refreshes. To bound the delayed-detection contribution, we further divide each true stationary segment into rounds before the relevant counter is refreshed, rounds after that refresh but before the next reset, and rounds after the reset. Bounding these contributions and accounting for concentration failure completes the proof of Theorem 3.1.

B.1 Notation and interval structure

Below we detail the interval structure and definitions used in the proofs.

Maximal static intervals and current-MSI statistics. Our analysis compares empirical observations with the current means using observations from a single algorithmic phase of each arms class. However, a class phase may contain undetected mean changes. We therefore split each class's timeline at both true changes and resets. The resulting *maximal static intervals* (MSIs) are the longest intervals on which both the class means and its phase start remain fixed.

Definition B.1 (Maximal static intervals)**.** *For each class $m \in [M]$, let*

$$\mathcal{R}_m := \{t \in [T] : \textit{class } m \textit{ is reset at the start of round } t\}.$$

A maximal static interval *of class m is a nonempty interval $I = [s, e] \subseteq [T]$, maximal under inclusion, such that*

$$\big(\mathcal{L}_m \cup \mathcal{R}_m\big) \cap \{s+1, \dots, e\} = \emptyset.$$

For each class, the MSIs partition $[T]$. The first MSI starts at round 1, and every subsequent MSI starts at a true change or reset round. Each ends immediately before the next MSI, or at T if it is the last.

For an arm a and round t, let τ be the starting round of the MSI of class $\mathrm{class}(a)$ containing t. We evaluate the current-MSI statistics using only observations before t:

$$N_a(t) := N_a([\tau, t-1]), \qquad \hat{\mu}_a(t) := \hat{\mu}_a([\tau, t-1]), \qquad \mathrm{CB}_a(t) := \mathrm{CB}_a([\tau, t-1]),$$

and, for every $q \geq 0$,

$$\hat{\mu}_a^{(\pm q)}(t) := \hat{\mu}_a^{(\pm q)}([\tau, t-1]).$$

At the start of an MSI, $\tau = t$, so the empty-history convention applies. These statistics are used in the analysis; the algorithm's classification certificates still use phase intervals, which may contain pre-change observations.

Blocking arms. In the regret decomposition, we compare the played arm with the optimal arm through two intermediate arms classified as good. This separates the comparison between good arms from the contributions of playing a bad arm and failing to recognize an optimal arm. A good arm is blocking itself, hence represents itself; a bad arm is represented by its blocking arm.

For each bad arm $a \in \mathcal{B}_t$, the classification rule in Algorithm 1 guarantees that its certificate set $X_{t,a}$ is nonempty. The algorithm selects a maximizing certificate using its fixed tie-breaking rule, and $\bar{a}_t(a)$ denotes the other arm appearing in that certificate, called the *blocking arm* of a. Formally,

Definition B.2 (Blocking arm)**.** *Fix a round $t \in [T]$ and a bad arm $a \in \mathcal{B}_t$. Let $X_{t,a}$ be its set of valid classification certificates:*

$$X_{t,a} := \left\{ (b, I, I') : \begin{array}{l} b \in \mathcal{A} \setminus \{a\}, \\ I \subseteq [\phi_{\mathrm{class}(a)}, t-1], \\ I' \subseteq [\phi_{\mathrm{class}(b)}, t-1], \\ \hat{\mu}_b^{(-3)}(I') > \hat{\mu}_a^{(+3)}(I) \end{array} \right\},$$

where I and I' are intervals of consecutive rounds, and the phase starts are evaluated after the round-t reset step.

Since a is classified as bad, $X_{t,a}$ is nonempty. Select the triple

$$\big(\bar{a}_t(a), I_t(a), \bar{I}_t(a)\big) \in \operatorname*{arg\,max}_{(b,I,I') \in X_{t,a}} \left\{ \hat{\mu}_b^{(-1)}(I') - \hat{\mu}_a^{(+1)}(I) \right\},$$

using the algorithm's fixed deterministic tie-breaking rule. The arm $\bar{a}_t(a)$ is called the blocking arm *of a at round t. The associated intervals $I_t(a)$ and $\bar{I}_t(a)$ contain the observations used for a and its blocking arm, respectively.*

The maximizing triple exists because the certificate set is finite and nonempty, and the tie-breaking rule makes its selection unique. By Proposition 4.3, the selected blocking arm satisfies $\bar{a}_t(a) \in \mathcal{G}_t$. This is a property of its classification, not a claim that its current mean exceeds that of a.

Thus this arm is well defined, and Proposition 4.3 establishes that $\bar{a}_t(a) \in \mathcal{G}_t$. Using blocking arms, we can define the representative of a bad arm.

Definition B.3 (Representatives)**.** *For each round $t \in [T]$, define the representative of the played arm a_t by*

$$\hat{a}_t := \begin{cases} a_t, & a_t \in \mathcal{G}_t, \\ \bar{a}_t(a_t), & a_t \in \mathcal{B}_t, \end{cases}$$

and the representative of the optimal arm a_t^* by

$$\hat{a}_t^* := \begin{cases} a_t^*, & a_t^* \in \mathcal{G}_t, \\ \bar{a}_t(a_t^*), & a_t^* \in \mathcal{B}_t. \end{cases}$$

Both representatives therefore belong to $\mathcal{G}_t \subseteq \mathcal{E}_t$. Adding and subtracting their mean rewards gives equation 4. The construction uses historical certificates and does not assume that a blocking arm's current mean exceeds that of the arm it represents.

Certificate residuals. To analyze rounds on which the optimal arm is classified as bad, we measure the discrepancy between its current mean and the historical upper bound used in its classification certificate. We denote the set of these rounds by $\mathcal{U}$, formally defined as

$$\mathcal{U} := \{t \in [T] : a_t^* \in \mathcal{B}_t\}.$$

For each $t \in \mathcal{U}$, define the *certificate residual*

$$r_t := \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)).$$

Here $I_t(a_t^*)$ is the optimal arm's selected certificate interval, and $\hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))$ is its empirical mean on that interval plus one confidence radius. Thus a positive residual means that the optimal arm's current mean exceeds its historical certificate upper bound.

We separate small residual contributions, which can be summed directly, from larger ones that require the monitoring argument. For each class $m \in [M]$, let

$$T_m := |\{t \in [T] : a_t^* \in \mathcal{A}_m\}|$$

be the total number of rounds whose optimal arm belongs to class m. These counts include all such rounds, not only those in $\mathcal{U}$, and satisfy $\sum_{m=1}^{M} T_m = T$. Denote $\Lambda = \lceil \log(2TK) \rceil$, where log is the natural logarithm. Define the class threshold

$$\zeta_m := \begin{cases} \Lambda \sqrt{\dfrac{MK_m(L_m+1)}{T_m}}, & T_m > 0, \\ 0, & T_m = 0. \end{cases}$$

After separating the confidence-radius contribution, the bound for Term (iii) involves $r_t - \Delta_t(a_t^*)$. This difference is nonpositive when the residual does not exceed the dominance gap. If the residual exceeds the dominance gap but not its class threshold, the difference is at most that threshold. We therefore reserve the monitoring argument for rounds whose residual exceeds both:

$$\mathcal{U}_{\text{large}} := \left\{ t \in \mathcal{U} : r_t > \max\left\{ \Delta_t(a_t^*), \zeta_{\text{class}(a_t^*)} \right\} \right\}.$$

To match these residuals to the monitoring scales, we group the rounds in $\mathcal{U}_{\text{large}}$ by the optimal arm's class and the magnitude of the residual. For each $m \in [M]$ and integer $j \geq 1$, define

$$\mathcal{U}_{m,j} := \left\{ t \in \mathcal{U}_{\text{large}} : a_t^* \in \mathcal{A}_m,\ 2^{-j} < r_t \leq 2^{1-j} \right\}. \tag{5}$$

These sets and thresholds are used only in the analysis, not by the algorithm.

B.2 Concentration Bounds

The intervals used in the analysis may be selected based on observed rewards. We therefore need a concentration bound that holds for all stationary intervals at once, rather than only for an interval fixed in advance.

For the analysis, imagine generating a reward sequence for each arm and each of its stationary segments before interaction, independently of the algorithm's random coins. Whenever an arm is pulled, it reveals the next unused reward in the corresponding sequence. We prove concentration simultaneously for every consecutive block of every sequence. Since the observations collected within any stationary round interval form such a block, this also covers intervals chosen using the observed history. Intervals with no observations are handled by the empty-history convention.

The resulting event $\mathcal{C}$ depends only on the presampled reward sequences, not on which rewards the algorithm chooses to observe. Conditioning on this event therefore leaves the distribution of the algorithm's independent randomness unchanged.

Lemma B.4 (Restatement of Lemma 4.1)**.** *Represent the rewards by presampling one independent sequence for each arm and each of its stationary segments. For a segment of length d, with mean μ and sequence $Y_1, \dots, Y_d$, let $\mathcal{C}$ be the event that, simultaneously for every such arm–segment sequence and all $1 \le r \le s \le d$,*

$$\left| \frac{1}{s-r+1} \sum_{u=r}^{s} Y_u - \mu \right| \le \sqrt{\frac{8 \log(2TK)}{s-r+1}}.$$

Then $\mathbb{P}(\mathcal{C}) \ge 1 - T^{-1}$. Moreover, on $\mathcal{C}$, every arm a and every nonempty interval I inside one of its stationary segments satisfy

$$\left|\hat{\mu}_a(I) - \mu\right| \le \mathrm{CB}_a(I).$$

Here μ is the segment's common mean.

Proof. The proof idea is first to apply Hoeffding's inequality to fixed blocks of presampled rewards, obtain simultaneous concentration over all such blocks, and then show that this event covers the intervals selected by the algorithm.

Step 1: Represent the rewards by independent sequences. For each arm $a \in \mathcal{A}$, let s_a be its number of stationary segments. For $i \in \{1, \dots, s_a\}$, denote the length and common mean of its i-th segment by $d_{a,i}$ and $\mu_{a,i}$, respectively. Since these segments partition the horizon, $\sum_{i=1}^{s_a} d_{a,i} = T$. For each pair (a, i), presample $Y_{a,i,1}, \dots, Y_{a,i,d_{a,i}}$ independently from the reward distribution of arm a in that segment. All entries are mutually independent and independent of the algorithm's random coins.

The k-th pull of arm a within its i-th segment reveals $Y_{a,i,k}$. Conditional on the observed history and the selected arm, the next revealed entry has the prescribed reward distribution. Thus the action–reward history has the same law as in the original interaction. We note this is the reward-stack representation of Lattimore & Szepesvári (2020, Section 4.6), applied separately to each arm's stationary segments.

Step 2: Apply Hoeffding's theorem to a fixed block. Fix an arm a, a segment i, and deterministic indices $1 \le r \le s \le d_{a,i}$. Write

$$n := s - r + 1, \qquad \overline{Y}_{a,i}(r,s) := \frac{1}{n} \sum_{k=r}^{s} Y_{a,i,k}.$$

The n summands are independent, take values in $[0,1]$, and have common mean $\mu_{a,i}$. Therefore, Theorem 2 of Hoeffding (1963), with all lower bounds equal to zero and all upper bounds equal to one, gives

$$\mathbb{P}\big(\overline{Y}_{a,i}(r,s) - \mu_{a,i} > \varepsilon\big) \le \exp(-2n\varepsilon^2) \qquad (\varepsilon > 0).$$

Applying the same theorem to the variables $1 - Y_{a,i,k}$ bounds the lower tail. Adding the two tail probabilities yields

$$\mathbb{P}\big(\big|\overline{Y}_{a,i}(r,s) - \mu_{a,i}\big| > \varepsilon\big) \le 2\exp(-2n\varepsilon^2).$$

Taking $\varepsilon = \sqrt{8\log(2TK)/n}$, we obtain

$$\mathbb{P}\left(\big|\overline{Y}_{a,i}(r,s) - \mu_{a,i}\big| > \sqrt{\frac{8\log(2TK)}{s-r+1}} \right) \le 2(2TK)^{-16}.$$

Step 3: Obtain concentration for all blocks simultaneously. The event $\mathcal{C}$ in the lemma is the intersection of the successful concentration events from Step 2 over all arms, segments, and block endpoints. A sequence of length $d_{a,i}$ contains exactly $d_{a,i}(d_{a,i}+1)/2$ nonempty consecutive blocks. Hence their total number is at most

$$\sum_{a\in\mathcal{A}}\sum_{i=1}^{s_a}\frac{d_{a,i}(d_{a,i}+1)}{2}\le\sum_{a\in\mathcal{A}}\left(\sum_{i=1}^{s_a}d_{a,i}\right)^2=KT^2.$$

The union bound, together with Step 2, therefore gives

$$\mathbb{P}(\bar{\mathcal{C}})\le 2KT^2(2TK)^{-16}\le T^{-1}.$$

In particular, $\mathbb{P}(\mathcal{C})\ge 1-T^{-1}$. Because the event is defined using only the presampled sequences, it is independent of the algorithm's random coins.

Step 4: Transfer the bound to adaptively selected intervals. Fix a realization in $\mathcal{C}$, an arm a, and a nonempty interval $I=[u,v]$ contained in the i-th stationary segment of that arm. Let b be the starting round of this segment, and define

$$h:=N_a([b,u-1]),\qquad n:=N_a(I).$$

Here h counts the entries revealed from this segment's sequence before I, while n counts those revealed during I.

Suppose first that $n\ge 1$. The rewards observed from arm a during I are exactly

$$Y_{a,i,h+1},\dots,Y_{a,i,h+n}.$$

Consequently,

$$\hat{\mu}_a(I)=\overline{Y}_{a,i}(h+1,h+n).$$

Although h and n may depend on the observed history, $\mathcal{C}$ controls every block simultaneously. Applying its defining inequality to this block gives

$$|\hat{\mu}_a(I)-\mu_{a,i}|\le\sqrt{\frac{8\log(2TK)}{n}}=:\mathrm{CB}_a(I).$$

This is a pathwise application of the simultaneous event; it does not condition on the random values of h or n.

If $n=0$, the empty-history convention gives $\hat{\mu}_a(I)=0$ and $\mathrm{CB}_a(I)=\sqrt{8\log(2TK)}\ge 1$. Since $\mu_{a,i}\in[0,1]$, the same inequality holds deterministically.

The arm and interval were arbitrary, so the claimed empirical concentration bounds hold simultaneously on $\mathcal{C}$. □

The conclusion is that under $\mathcal{C}$, every current-MSI statistic therefore concentrates around the arm's current mean.

B.3 Reset and classification properties

We first use concentration to bound the number of resets, as formally stated next.

Lemma B.5 (Restatement of Lemma 4.2)**.** *Assume event $\mathcal{C}$ holds, then every class m is reset at most L_m times.*

The proof idea is that under $\mathcal{C}$, if no mean in a class has changed during its current phase, confidence intervals computed from different parts of that phase for the same arm all contain its common mean. They therefore overlap, so the reset condition cannot hold. Full proof is given below.

Proof. Fix a class m and a round t at which the algorithm resets that class. Let ϕ_m denote its phase start immediately before the reset. By the change-detection rule in Algorithm 1, line 4, there exist an arm $a\in\mathcal{A}_m$ and consecutive intervals $I_1,I_2\subseteq[\phi_m,t-1]$ such that

$$\hat{\mu}_a^{(-1)}(I_1)>\hat{\mu}_a^{(+1)}(I_2).$$

Both witness intervals have non-empty pull counts. Under the empty-history convention, a zero-pull interval has a negative lower bound and an upper bound larger than one, whereas every empirical lower bound is at most one and every empirical upper bound is nonnegative. Thus a zero-pull interval cannot appear on either side of the strict inequality. In particular, both intervals are nonempty.

Suppose, toward a contradiction, that class m has no change round in $\{\phi_m + 1, \dots, t-1\}$. Then arm a has the constant mean $\mu_a(\phi_m)$ throughout $[\phi_m, t-1]$. On $\mathcal{C}$,

$$|\hat{\mu}_a(I_r) - \mu_a(\phi_m)| \le \mathrm{CB}_a(I_r), \qquad r \in \{1, 2\}.$$

Consequently,

$$\hat{\mu}_a^{(-1)}(I_1) \le \mu_a(\phi_m) \le \hat{\mu}_a^{(+1)}(I_2),$$

contradicting the reset condition. Hence at least one change round of class m lies strictly after the phase start and before t.

Charge the reset to any such change round. After the reset at t, the new phase starts at t, so any later reset must be charged to a change round strictly after t. Thus distinct resets receive distinct charges. Since class m has at most L_m change rounds, it can be reset at most L_m times. □

On $\mathcal{C}$, every reset of class m follows a true change within the phase being reset. Since each reset starts a new phase, the same change cannot justify a later reset. Class m therefore undergoes at most L_m resets. Its first MSI starts at round 1, and every subsequent MSI starts at a true change or reset round. Together, the at most L_m true changes and L_m resets thus partition the horizon into at most $2L_m + 1$ MSIs. Formally,

Lemma B.6 (Number of MSIs)**.** *Under the concentration event $\mathcal{C}$, for each class m, the number of MSIs is at most $2L_m + 1$.*

Proof. Fix a class m. By Lemma 4.2, it has at most L_m phase resets. Together with its at most L_m change rounds, these events partition the rounds into at most $2L_m + 1$ MSIs. □

Consequently, on $\mathcal{C}$, the total number of arm–MSI pairs is at most

$$\sum_{m=1}^{M} (2L_m + 1)K_m \le 2 \sum_{m=1}^{M} (L_m + 1)K_m = 2D.$$

This count includes pairs whose arms are never pulled. They will contribute zero to the played-arm confidence sum.

B.3.1 Blocking arms are good

We next show that every bad arm's selected blocking arm is classified as good.

Proposition B.7 (Restatement of Proposition 4.3)**.** *For every round t and arm $a \in \mathcal{B}_t$, its blocking arm $\bar{a}_t(a)$ is good, meaning $\bar{a}_t(a) \in \mathcal{G}_t$.*

Proof. Fix a round t and an arm $a \in \mathcal{B}_t$, and write $a' := \bar{a}_t(a)$. Suppose, toward a contradiction, that $a' \in \mathcal{B}_t$. Let the selected maximal certificate for a' be the tuple $\big(\bar{a}_t(a'), I_t(a'), \overline{I}_t(a')\big) \in X_{t,a'}$. Its second component is the observation interval for a', and its third component is the observation interval for $\bar{a}_t(a')$, the arm blocking a'. It thus holds true that $\hat{\mu}^{(-3)}_{\bar{a}_t(a')}(\overline{I}_t(a')) > \hat{\mu}^{(+3)}_{a'}(I_t(a'))$.

Let the selected certificate for a be the tuple $\big(a', I_t(a), \overline{I}_t(a)\big) \in X_{t,a}$. Here the second component is the observation interval for a, whereas the third component is the observation interval for a'.

The certificate by which a' blocks a likewise satisfies $\hat{\mu}^{(-3)}_{a'}(\overline{I}_t(a)) > \hat{\mu}^{(+3)}_{a}(I_t(a))$.

Step 1: Verify a second blocking certificate for a. Let us consider the candidate tuple $\big(\bar{a}_t(a'), I_t(a), \overline{I}_t(a')\big)$. This retains the interval for a from its selected certificate and takes the blocking arm and its observation interval from the selected certificate for a'. To verify this candidate, we connect the two original certificates through $\overline{I}_t(a)$ and $I_t(a')$: both are observation intervals for a', although they use in different certificates.

Because a' is bad after the reset step, its class was not reset at round t: a newly reset arm has no observations, so its $+3$ bound exceeds one, whereas every arm's -3 bound is at most one. Thus the change-detection test is false on the current phase of a'. Applying that fact to the two intervals of a' gives

$$\hat{\mu}_{a'}^{(-1)}(\overline{I}_t(a)) \le \hat{\mu}_{a'}^{(+1)}(I_t(a')).$$

Equivalently,

$$\hat{\mu}_{a'}(I_t(a')) \ge \hat{\mu}_{a'}(\overline{I}_t(a)) - \mathrm{CB}_{a'}(\overline{I}_t(a)) - \mathrm{CB}_{a'}(I_t(a')).$$

Combining this inequality with the two blocking certificates yields

$$\begin{aligned}
&\hat{\mu}_{\bar{a}_t(a')}^{(-3)}(\overline{I}_t(a')) \\
&> \hat{\mu}_{a'}^{(+3)}(I_t(a')) \\
&\ge \hat{\mu}_{a'}(\overline{I}_t(a)) - \mathrm{CB}_{a'}(\overline{I}_t(a)) + 2\mathrm{CB}_{a'}(I_t(a')) \\
&\ge \hat{\mu}_{a'}^{(-1)}(\overline{I}_t(a)) \\
&= \hat{\mu}_{a'}^{(-3)}(\overline{I}_t(a)) + 2\mathrm{CB}_{a'}(\overline{I}_t(a)) \\
&> \hat{\mu}_{a}^{(+3)}(I_t(a)).
\end{aligned}$$

Thus $\big(\bar{a}_t(a'), I_t(a), \overline{I}_t(a')\big)$ satisfies the blocking inequality for a. Moreover, $\bar{a}_t(a') \neq a$: otherwise the preceding strict separation would give two intervals of a in its current phase with disjoint one-confidence-bound intervals, and the change-detection step would have reset its class before classification. The displayed triplet is therefore a valid candidate in $X_{t,a}$.

Step 2: Compare the empirical separations. We compare the new candidate tuple from Step 1 with the originally selected tuple $\big(a', I_t(a), \overline{I}_t(a)\big)$. Both use the same interval $I_t(a)$ for the blocked arm a, so their one-radius separations subtract the same upper bound $\hat{\mu}_a^{(+1)}(I_t(a))$. It therefore suffices to show that the new candidate's blocking-arm lower bound is strictly larger.

The inequalities of Step 1 also give $\hat{\mu}_{\bar{a}_t(a')}^{(-3)}(\overline{I}_t(a')) > \hat{\mu}_{a'}^{(-1)}(\overline{I}_t(a))$. Consequently,

$$\begin{aligned}
&\hat{\mu}_{\bar{a}_t(a')}^{(-1)}(\overline{I}_t(a')) - \hat{\mu}_a^{(+1)}(I_t(a)) \\
&= \hat{\mu}_{\bar{a}_t(a')}^{(-3)}(\overline{I}_t(a')) + 2\mathrm{CB}_{\bar{a}_t(a')}(\overline{I}_t(a')) - \hat{\mu}_a^{(+1)}(I_t(a)) \\
&> \hat{\mu}_{a'}^{(-1)}(\overline{I}_t(a)) - \hat{\mu}_a^{(+1)}(I_t(a)).
\end{aligned}$$

Hence this candidate has a strictly larger empirical separation than the certificate defining $\bar{a}_t(a) = a'$, contradicting its maximizing property. Therefore $\bar{a}_t(a) \in \mathcal{G}_t$. □

The preceding proposition ensures that $\mathcal{G}_t$ is nonempty at every round: if any arm is bad, its blocking arm is good; otherwise, every arm is good. Since all good arms are eligible, $\mathcal{E}_t$ is also nonempty. Moreover, both representatives in the regret decomposition are good and hence eligible. To control their confidence contributions, we therefore bound the cumulative confidence radii of eligible arms.

B.4 Cumulative confidence-sum bounds

We first consider the simpler sum over arms actually played. Within an MSI, each pull increases the corresponding arm's observation count, allowing us to sum its confidence radii in terms of its total number of pulls.

Lemma B.8 (Played-arm confidence sum)**.** *Assume that the concentration event $\mathcal{C}$ defined in Lemma 4.1 holds. Then, for the realized played-arm sequence $a_1, \ldots, a_T$,*

$$\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) \le 3\sqrt{8\log(2TK)\, T \sum_{m=1}^{M} (2L_m + 1)K_m}.$$

In particular, this is $\widetilde{O}(\sqrt{TD})$.

Proof. Assume $\mathcal{C}$ holds and fix a realization of the played-arm sequence. We group the pulls by their arm and MSI, bound the contribution of each group, and then sum these bounds.

Step 1: Group the pulls by arm and MSI. Let $\mathcal{P}$ be the set of pairs (a, I), where a is an arm and I is an MSI of its class. For each pair, define $n_{a,I} := N_a(I) = \sum_{t \in I} \mathbf{1}\{a_t = a\}$. Thus $n_{a,I}$ is the total number of pulls of arm a during I.

Every round contributes to exactly one pair: the played arm and the unique MSI of its class containing that round. Consequently, $\sum_{(a,I)\in\mathcal{P}} n_{a,I} = T$. Also, by Lemma B.6, class m has at most $2L_m + 1$ MSIs on $\mathcal{C}$. Each such MSI is paired with the K_m arms in that class, so $|\mathcal{P}| \le \sum_{m=1}^{M} (2L_m + 1)K_m$.

Step 2: Bound the contribution of one arm within one MSI. Fix $(a, I) \in \mathcal{P}$, and write $n := n_{a,I}$. If $n = 0$, this pair contributes nothing. Otherwise, let $t_1 < \cdots < t_n$ be the rounds in I at which arm a is played. Here t_k is the global round of the k-th pull of arm a within I.

Immediately before round t_k, arm a has exactly $k-1$ observations in this MSI. Since the current-MSI statistics exclude the current round's observation, for every $k \in \{1, \dots, n\}$,

$$N_a(t_k) = k - 1, \qquad \mathrm{CB}_a(t_k) = \sqrt{\frac{8\log(2TK)}{\max\{k-1, 1\}}}.$$

To sum these contributions, write $q := k - 1$ for the number of preceding pulls of arm a in I. As the pull index k ranges from 1 to n, this count q ranges from 0 to $n - 1$. Therefore,

$$\begin{aligned}
\sum_{\substack{t \in I \\ a_t = a}} \mathrm{CB}_a(t) &= \sum_{k=1}^{n} \mathrm{CB}_a(t_k) \\
&= \sqrt{8\log(2TK)} \sum_{q=0}^{n-1} \frac{1}{\sqrt{\max\{q, 1\}}} \\
&= \sqrt{8\log(2TK)} \left(1 + \sum_{q=1}^{n-1} \frac{1}{\sqrt{q}}\right).
\end{aligned}$$

The separate term 1 corresponds to $q = 0$, namely the first pull, before which no observation of arm a has been collected in this MSI. The remaining sum accounts for the second through the n-th pulls.

We bound the remaining sum by an integral. For each integer $q \ge 2$ and every $x \in [q-1, q]$, we have $x \le q$, so $x^{-1/2} \ge q^{-1/2}$. Integrating over this interval of length one gives

$$\frac{1}{\sqrt{q}} \le \int_{q-1}^{q} \frac{dx}{\sqrt{x}}.$$

For $n \ge 2$, separate the term $q = 1$ and apply this inequality to the other terms:

$$\begin{aligned}
\sum_{q=1}^{n-1} \frac{1}{\sqrt{q}} &= 1 + \sum_{q=2}^{n-1} \frac{1}{\sqrt{q}} \\
&\le 1 + \int_{1}^{n-1} \frac{dx}{\sqrt{x}} \\
&= 2\sqrt{n-1} - 1 \\
&\le 2\sqrt{n}.
\end{aligned}$$

When $n = 2$, the sum from $q = 2$ to $n - 1$ and the integral are both zero. When $n = 1$, the entire sum from $q = 1$ to $n - 1$ is empty, so the bound by $2\sqrt{n}$ also holds. Consequently, for every $n \ge 1$,

$$1 + \sum_{q=1}^{n-1} \frac{1}{\sqrt{q}} \le 1 + 2\sqrt{n} \le 3\sqrt{n},$$

where the final inequality uses $1 \le \sqrt{n}$. Substituting this into the confidence sum gives

$$\sum_{\substack{t \in I \\ a_t = a}} \mathrm{CB}_a(t) \le 3\sqrt{8\log(2TK)}\,\sqrt{n_{a,I}}.$$

This inequality also holds when $n_{a,I} = 0$, since both sides are zero.

Step 3: Sum over all arm–MSI pairs. Using the bound from Step 2,

$$\begin{aligned}\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) &= \sum_{(a,I)\in\mathcal{P}} \sum_{\substack{t \in I \\ a_t = a}} \mathrm{CB}_a(t) \\ &\le 3\sqrt{8\log(2TK)} \sum_{(a,I)\in\mathcal{P}} \sqrt{n_{a,I}}.\end{aligned}$$

Apply Cauchy–Schwarz to the sequences $(1)_{(a,I)\in\mathcal{P}}$ and $(\sqrt{n_{a,I}})_{(a,I)\in\mathcal{P}}$:

$$\begin{aligned}\sum_{(a,I)\in\mathcal{P}} \sqrt{n_{a,I}} &\le \sqrt{\left(\sum_{(a,I)\in\mathcal{P}} 1\right)\left(\sum_{(a,I)\in\mathcal{P}} n_{a,I}\right)} \\ &= \sqrt{|\mathcal{P}|\,T} \\ &\le \sqrt{\sum_{m=1}^{M} (2L_m+1)K_m\,T}.\end{aligned}$$

The equality uses the fact that the pull counts sum to T, and the last inequality uses the bound on $|\mathcal{P}|$ from Step 1. Hence

$$\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) \le 3\sqrt{8\log(2TK)\,T\sum_{m=1}^{M}(2L_m+1)K_m}.$$

Finally, since $2L_m + 1 \le 2(L_m + 1)$ for every class m,

$$\sum_{m=1}^{M} (2L_m+1)K_m \le 2\sum_{m=1}^{M}(L_m+1)K_m = 2D.$$

Substituting this inequality and using $3\sqrt{8\cdot 2} = 12$ gives

$$\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) \le 12\sqrt{\log(2TK)\,TD},$$

which is $\widetilde{O}(\sqrt{TD})$. □

We note that since every term corresponds to an arm that was actually played, this bound has no additional $\sqrt{M}$ factor.

In the previous proof, the bound on the number of arm–MSI pairs then controls the total contribution across the horizon. For eligible arms that are not played, this direct counting argument does not apply. We instead use the randomized class–scale selection and recency rule to relate their radii to those of arms the algorithm samples. This yields a bound on the sum of the largest eligible-arm radius at each round, which applies in particular to both representatives, as formally stated next.

Lemma B.9 (Eligible-arm confidence sum)**.** *Let $\mathcal{C}$ be the concentration event from Lemma 4.1. For every admissible instance, Algorithm* `EBF-H` *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t) \,\middle|\, \mathcal{C}\right] \le 64\,\log(2TK)\sqrt{MTD}.$$

Consequently, for any possibly random sequence $a'_1, \dots, a'_T$ satisfying $a'_t \in \mathcal{E}_t$ at every round,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{CB}_{a'_t}(t) \,\middle|\, \mathcal{C}\right] \le 64 \log(2TK)\sqrt{MTD}.$$

In particular, both expectations are $\widetilde{O}\left(\sqrt{MTD}\right)$.

The proof idea is that an eligible arm need not be played when its confidence radius is counted in the sum. We therefore cannot simply count its successive pulls as in the preceding lemma. Instead, at each round we identify a comparison arm whose radius is at least half the largest eligible-arm radius and which the algorithm plays with a controlled probability. Whenever this comparison arm is played, its MSI observation count increases. We use these increases to bound the number of rounds with small observation counts. On the remaining rounds, the counts are large enough that the confidence radii themselves are small. This is formally given next.

Proof. Let $\ell := \log(2TK)$ and recall that $N_a(t)$ counts observations of arm a before round t in its current MSI, and $\mathrm{CB}_a(t) = \sqrt{\frac{8\ell}{\max\{N_a(t),1\}}}$. We retain the randomness of the interaction throughout the proof; the bound is on a conditional expectation, rather than on every realized sequence of plays.

Step 1: Choose a comparison arm with a comparable radius. At each round t, after the counter refreshes and before the class and scale draws, choose $\widetilde{a}_t \in \arg\max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t)$, using the fixed tie-breaking rule. Such an arm exists because $\mathcal{E}_t$ is nonempty.

Let $k_t := N_{\widetilde{a}_t}(t)$ be the number of observations this arm has already received in its current MSI. Choose the scale $h_t := \lceil \log_2(k_t + 1) \rceil$. This scale is used only in the analysis; it is not necessarily the scale drawn by the algorithm. Since $0 \le k_t \le T-1$, we have $h_t \in \{0, \dots, J_{\max}\}$. The definition also gives $k_t < 2^{h_t} \le 2(k_t + 1)$.

Define the comparison arm by $\widetilde{a}'_t \in \arg\min_{a\in\mathcal{E}_t\cap\mathcal{A}_{\mathrm{class}(\widetilde{a}_t)}} \tau_a^{h_t}$, using the same rule as in line 14 of Algorithm 1. Thus $\widetilde{a}'_t$ is the arm the algorithm would play if it selected the class of $\widetilde{a}_t$ and scale h_t.

We now compare the two arms' observation counts. Let u_t be the starting round of their common current MSI. They share this MSI because they belong to the same class. Recall that $\tau_a^{h_t}$ is the round of arm a's 2^{h_t}-th most recent pull in its current phase, or $-\infty$ if that many pulls do not exist.

Arm $\widetilde{a}_t$ has only $k_t < 2^{h_t}$ observations in its current MSI. Its 2^{h_t}-th most recent phase observation must therefore precede u_t, if it exists. By the minimizing choice of $\widetilde{a}'_t$,

$$\tau_{\widetilde{a}'_t}^{h_t} \le \tau_{\widetilde{a}_t}^{h_t} < u_t.$$

Consequently, $\widetilde{a}'_t$ also has fewer than 2^{h_t} observations in this MSI:

$$N_{\widetilde{a}'_t}(t) < 2^{h_t} \le 2(k_t + 1).$$

If $k_t = 0$, then $h_t = 0$, so this inequality implies $N_{\widetilde{a}'_t}(t) = 0$. If $k_t \ge 1$, then $2(k_t + 1) \le 4k_t$. In either case,

$$\max\{N_{\widetilde{a}'_t}(t), 1\} \le 4\max\{k_t, 1\}.$$

Substituting into the confidence-radius formula gives

$$\frac{\mathrm{CB}_{\widetilde{a}_t}(t)}{\mathrm{CB}_{\widetilde{a}'_t}(t)} = \sqrt{\frac{\max\{N_{\widetilde{a}'_t}(t), 1\}}{\max\{k_t, 1\}}} \le 2.$$

Therefore, at every round,

$$\max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t) \le 2\mathrm{CB}_{\widetilde{a}'_t}(t).$$

It remains to bound the sum of the comparison arms' radii.

Step 2: Bound the probability of playing the comparison arm. For the conditioning argument, let $\mathcal{H}_t$ be the sigma-field generated by all presampled reward sequences, all algorithmic random draws

before round t, and the counter-refresh coins at round t. These variables determine the eligible set, the two arms from Step 1, and the scale h_t. The round-t class and scale draws have not been included.

Conditional on $\mathcal{H}_t$, the algorithm selects uniformly among at most M classes containing an eligible arm. The class of $\widetilde{a}_t$ is one of these classes. Independently, the algorithm selects uniformly among the $J_{\max}+1$ maintained scales. Thus

$$\mathbb{P}(m_t = \mathrm{class}(\widetilde{a}_t),\ j_t = h_t \,|\, \mathcal{H}_t) \geq \frac{1}{M(J_{\max}+1)} =: p.$$

When these two choices occur, the algorithm plays $\widetilde{a}'_t$. It may also play this arm under other scale choices, so

$$\mathbb{P}(a_t = \widetilde{a}'_t \,|\, \mathcal{H}_t) \geq p.$$

The event $\mathcal{C}$ depends only on the presampled reward sequences, which are included in $\mathcal{H}_t$. The fresh class and scale draws therefore retain these probabilities when we subsequently condition on $\mathcal{C}$.

Step 3: Count rounds with a given number of preceding observations. Fix a nonnegative integer n. Here n is a possible observation count, not a round number or the total number of pulls over the horizon. Define the random counts

$$V_n := \sum_{t=1}^{T} \mathbf{1}\{N_{\widetilde{a}'_t}(t) = n\},$$
$$S_n := \sum_{t=1}^{T} \mathbf{1}\{N_{\widetilde{a}'_t}(t) = n,\ a_t = \widetilde{a}'_t\}.$$

The first count, V_n, includes every round when the comparison arm has exactly n preceding observations in its current MSI. The second, S_n, includes only those rounds when the algorithm actually plays that comparison arm. Both counts are zero for $n \geq T$.

First, we bound S_n. Every round counted by S_n is the $(n+1)$-st pull of a particular arm within a particular MSI. The same arm–MSI pair cannot have two different $(n+1)$-st pulls. Hence each pair contributes at most once to S_n, even though the comparison arm may change between rounds. By Lemma B.6, on $\mathcal{C}$,

$$S_n \leq \sum_{m=1}^{M} (2L_m + 1)K_m \leq 2D.$$

Next, we relate V_n to S_n. The comparison arm and its preceding observation count are determined by $\mathcal{H}_t$. Therefore, Step 2 gives

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}\{N_{\widetilde{a}'_t}(t) = n,\ a_t = \widetilde{a}'_t\} \,\big|\, \mathcal{H}_t\big] \\
&\quad = \mathbf{1}\{N_{\widetilde{a}'_t}(t) = n\}\mathbb{P}(a_t = \widetilde{a}'_t \,|\, \mathcal{H}_t) \\
&\quad \geq p\,\mathbf{1}\{N_{\widetilde{a}'_t}(t) = n\}.
\end{aligned}$$

Taking conditional expectations given $\mathcal{C}$ and summing over t, the tower property (Lattimore & Szepesvári, 2020, Theorem 2.12(5)) yields

$$p\,\mathbb{E}[V_n \mid \mathcal{C}] \leq \mathbb{E}[S_n \mid \mathcal{C}] \leq 2D.$$

Thus, for every nonnegative integer n,

$$\mathbb{E}[V_n \mid \mathcal{C}] \leq \frac{2D}{p}.$$

This bounds the number of rounds with observation count n, including rounds when the comparison arm is not played.

Step 4: Sum the confidence radii over observation counts. Recall that $\ell = \log(2TK)$, and that V_n counts the rounds at which the comparison arm has exactly n preceding observations in its current MSI. At every such round, its radius equals $\sqrt{\frac{8\ell}{\max\{n,1\}}}$. The combined contribution of these rounds

is therefore $V_n\sqrt{\frac{8\ell}{\max\{n,1\}}}$. The comparison arms may differ between these rounds, but their radii are equal because their observation counts are equal.

Every round has exactly one observation count $n \in \{0, \dots, T-1\}$. Thus each round is counted in exactly one of $V_0, \dots, V_{T-1}$, and $\sum_{n=0}^{T-1} V_n = T$. Adding the contributions for these different counts gives, on every realization,

$$\sum_{t=1}^{T} \mathrm{CB}_{\widetilde{a}'_t}(t) = \sqrt{8\ell} \sum_{n=0}^{T-1} \frac{V_n}{\sqrt{\max\{n,1\}}}.$$

Taking conditional expectations and using linearity, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{CB}_{\widetilde{a}'_t}(t) \,\middle|\, \mathcal{C}\right] = \sqrt{8\ell} \sum_{n=0}^{T-1} \frac{\mathbb{E}[V_n \mid \mathcal{C}]}{\sqrt{\max\{n,1\}}}.$$

Small observation counts give large radii, but Step 3 limits how often each such count occurs: $\mathbb{E}[V_n \mid \mathcal{C}] \le \frac{2D}{p}$. Large observation counts may occur many times, but each corresponding radius is small. We use these two bounds on opposite sides of the threshold $\beta := \frac{Tp}{2D}$. The threshold is a positive real number; the observation counts compared with it are nonnegative integers.

First handle the case $\beta < 1$. In this case, $T < \frac{2D}{p}$. Every arm's confidence radius is at most $\sqrt{8\ell}$. Hence we can bound the largest eligible-arm radius directly, without using the comparison arm:

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{E}_t} \mathrm{CB}_a(t) \,\middle|\, \mathcal{C}\right] \le T\sqrt{8\ell} \le \sqrt{8\ell T \frac{2D}{p}} = 4\sqrt{\frac{\ell T D}{p}}.$$

The second inequality uses $T^2 \le 2TD/p$. This case already gives a smaller bound than the one obtained below. We therefore assume $\beta \ge 1$ for the next two parts.

Small counts: fewer than β preceding observations. The contribution of rounds with $N_{\widetilde{a}'_t}(t) < \beta$ can be written as

$$\begin{aligned}\mathbb{E}\left[\sum_{t \in [T] N_{\widetilde{a}'_t}(t) < \beta} \mathrm{CB}_{\widetilde{a}'_t}(t) \,\middle|\, \mathcal{C}\right] &= \sqrt{8\ell} \sum_{\substack{0 \le n \le T-1 \\ n < \beta}} \frac{\mathbb{E}[V_n \mid \mathcal{C}]}{\sqrt{\max\{n,1\}}} \\ &\le \sqrt{8\ell}\,\frac{2D}{p} \sum_{\substack{0 \le n \le T-1 \\ n < \beta}} \frac{1}{\sqrt{\max\{n,1\}}}\end{aligned}$$

where the inequality applies the bound from Step 3 to each count n. All sums over n below also range over integers.

To bound the remaining sum, the terms $n = 0$ and $n = 1$, when present, contribute at most two. For each integer $n \ge 2$, the function $x^{-1/2}$ is at least $n^{-1/2}$ throughout $[n-1, n]$. Therefore,

$$\frac{1}{\sqrt{n}} \le \int_{n-1}^{n} \frac{dx}{\sqrt{x}}.$$

For the integers $n < \beta$, these integration intervals have disjoint interiors and lie inside $[1, \beta]$. Since the integrand is nonnegative,

$$\sum_{\substack{0 \le n \le T-1 \\ n < \beta}} \frac{1}{\sqrt{\max\{n,1\}}} \le 2 + \int_{1}^{\beta} \frac{dx}{\sqrt{x}} = 2 + 2(\sqrt{\beta} - 1) = 2\sqrt{\beta}.$$

This remains valid when $\beta = 1$: only the term $n = 0$ is included, and the upper bound is two.

Substituting this estimate and the definition of β gives

$$\mathbb{E}\left[\sum_{\substack{t\in[T]\\ N_{\widetilde{a}'_t}(t)<\beta}} \mathrm{CB}_{\widetilde{a}'_t}(t)\,\middle|\,\mathcal{C}\right] \le \sqrt{8\ell}\,\frac{2D}{p}\cdot 2\sqrt{\beta} = \sqrt{8\ell}\,\frac{4D}{p}\sqrt{\frac{Tp}{2D}} = 8\sqrt{\frac{\ell TD}{p}}.$$

Large counts: at least β preceding observations. On a round with $N_{\widetilde{a}'_t}(t) \ge \beta$, the confidence-radius formula gives

$$\mathrm{CB}_{\widetilde{a}'_t}(t) = \sqrt{\frac{8\ell}{\max\{N_{\widetilde{a}'_t}(t),1\}}} \le \sqrt{\frac{8\ell}{\beta}}.$$

There are at most T such rounds. Thus, on every realization,

$$\sum_{\substack{t\in[T]\\ N_{\widetilde{a}'_t}(t)\ge\beta}} \mathrm{CB}_{\widetilde{a}'_t}(t) \le T\sqrt{\frac{8\ell}{\beta}} = T\sqrt{\frac{16\ell D}{Tp}} = 4\sqrt{\frac{\ell TD}{p}}.$$

The same bound holds after taking conditional expectations given $\mathcal{C}$.

Combine the two contributions. Each round has either fewer than β observations or at least β observations, so the two sums cover all rounds without overlap. Adding their conditional bounds yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{CB}_{\widetilde{a}'_t}(t)\,\middle|\,\mathcal{C}\right] \le (8+4)\sqrt{\frac{\ell TD}{p}}.$$

By Step 1,

$$\max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t) \le 2\mathrm{CB}_{\widetilde{a}'_t}(t)$$

at every round. Consequently,

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t)\,\middle|\,\mathcal{C}\right] \le 24\sqrt{\frac{\ell TD}{p}}.$$

This bound also covers the case $\beta < 1$, for which we obtained the smaller coefficient four.

Recall that

$$p = \frac{1}{M(J_{\max}+1)}.$$

Moreover, since $J_{\max} = \lceil \log_2 T \rceil$, $T \ge 1$, and $K \ge 2$,

$$J_{\max} + 1 \le \log_2 T + 2 \le 3\log(2TK) = 3\ell.$$

Substituting these relations gives

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t)\,\middle|\,\mathcal{C}\right] &\le 24\sqrt{\ell M(J_{\max}+1)TD}\\ &\le 24\sqrt{3}\,\ell\sqrt{MTD}\\ &\le 32\sqrt{3}\,\log(2TK)\sqrt{MTD}.\end{aligned}$$

The last inequality uses $24 \le 32$, giving the constant stated in the lemma. In particular, the expectation is $\widetilde{O}\left(\sqrt{MTD}\right)$.

Finally, let $a'_1, \ldots, a'_T$ be any possibly random sequence satisfying $a'_t \in \mathcal{E}_t$ at every round. On every realization,

$$\sum_{t=1}^{T} \mathrm{CB}_{a'_t}(t) \le \sum_{t=1}^{T} \max_{a\in\mathcal{E}_t} \mathrm{CB}_a(t).$$

Taking conditional expectations therefore gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{CB}_{a'_t}(t) \,\middle|\, \mathcal{C}\right] \le 64\,\log(2TK)\sqrt{MTD}.$$

No independence assumption on the choices a'_t is needed, because the comparison with the maximum holds on every realization. □

B.5 Regret decomposition and its three contributions

The confidence bounds now apply to both representative arms. It remains to separate the comparisons they can control directly from the cost of sampling bad arms and the residual left by an outdated optimal-arm certificate.

Lemma B.10 (Three-term regret decomposition)**.** *For the representatives in Definition B.3, define*

$$R_1 := \sum_{t=1}^{T} \big(\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)\big),$$
$$R_2 := \sum_{t=1}^{T} \big(\mu_{\hat{a}^*_t}(t) - \mu_{\hat{a}_t}(t)\big),$$
$$R_3 := \sum_{t=1}^{T} \big(\mu_{a^*_t}(t) - \mu_{\hat{a}^*_t}(t)\big).$$

Then $\mathrm{Regret}_T = R_1 + R_2 + R_3$ *on every realization. The three sums correspond to Terms (i), (ii), and (iii) of equation 4; their summands need not all be nonnegative.*

Proof. For every round t, add and subtract the means of the two representatives:

$$\mu_{a^*_t}(t) - \mu_{a_t}(t) = \big(\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)\big) + \big(\mu_{\hat{a}^*_t}(t) - \mu_{\hat{a}_t}(t)\big) + \big(\mu_{a^*_t}(t) - \mu_{\hat{a}^*_t}(t)\big).$$

Summing over $t \in [T]$ proves the identity. □

We bound each contribution from above. The good representatives are already covered by the eligible-arm confidence bound. For a bad arm, the historical dominance gap must be accompanied by confidence corrections before it can be compared with current means.

B.5.1 Term (i): monitoring bad arms

Lemma B.11 (Restatement of Lemma 4.4)**.** *For the concentration event $\mathcal{C}$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} [\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)] \,\middle|\, \mathcal{C}\right] \le \widetilde{O}\Big(\sqrt{MTD}\Big).$$

Proof. If the played arm is good, then $\hat{a}_t = a_t$ and the summand is zero. We therefore fix a round t on which $a_t \in \mathcal{B}_t$. By Definition B.3, $\hat{a}_t = \bar{a}_t(a_t)$, and Proposition 4.3 gives $\hat{a}_t \in \mathcal{G}_t$.

The blocking certificate implies

$$\begin{aligned}\Delta_t(a_t) &= \hat{\mu}^{(-1)}_{\hat{a}_t}(\overline{I}_t(a_t)) - \hat{\mu}^{(+1)}_{a_t}(I_t(a_t)) \\ &\ge 2\mathrm{CB}_{\hat{a}_t}(\overline{I}_t(a_t)) + 2\mathrm{CB}_{a_t}(I_t(a_t)).\end{aligned} \tag{6}$$

Using the first line of this display, we have the exact decomposition

$$\begin{aligned}
&\mu_{\hat{a}_t}(t) - \mu_{a_t}(t) \\
&= \underbrace{\mu_{\hat{a}_t}(t) - \hat{\mu}^{(-1)}_{\hat{a}_t}(\overline{I}_t(a_t)) - \Delta_t(a_t)}_{\text{(1a)}} \\
&\quad + \underbrace{3\Delta_t(a_t)}_{\text{(1b)}} \\
&\quad + \underbrace{\hat{\mu}^{(+1)}_{a_t}(I_t(a_t)) - \Delta_t(a_t) - \mu_{a_t}(t)}_{\text{(1c)}}. \qquad (7)
\end{aligned}$$

The certificate and the unit reward range imply

$$0 \le \hat{\mu}^{(+1)}_{a_t}(I_t(a_t)) < \hat{\mu}^{(-1)}_{\hat{a}_t}(\overline{I}_t(a_t)) \le 1.$$

Thus each of (1a) and (1c) is at most one, since $\Delta_t(a_t) > 0$ and both true means lie in $[0, 1]$. If the corresponding arm's MSI history is empty, its confidence radius is at least one, proving the claimed bound by twice that radius. Below we therefore consider nonempty MSI histories.

Bounding Term (1a). The concentration event on the MSI of $\hat{a}_t$ gives $\mu_{\hat{a}_t}(t) \le \hat{\mu}^{(+1)}_{\hat{a}_t}(t)$. Hence

$$\begin{aligned}
\text{(1a)} &\le \hat{\mu}^{(+1)}_{\hat{a}_t}(t) - \hat{\mu}^{(-1)}_{\hat{a}_t}(\overline{I}_t(a_t)) - \Delta_t(a_t) \\
&\le \hat{\mu}^{(+1)}_{\hat{a}_t}(t) - \hat{\mu}^{(-1)}_{\hat{a}_t}(\overline{I}_t(a_t)) - 2\mathrm{CB}_{\hat{a}_t}(\overline{I}_t(a_t)) \\
&= \hat{\mu}^{(-1)}_{\hat{a}_t}(t) - \hat{\mu}^{(+1)}_{\hat{a}_t}(\overline{I}_t(a_t)) + 2\mathrm{CB}_{\hat{a}_t}(t) \\
&\le 2\mathrm{CB}_{\hat{a}_t}(t).
\end{aligned}$$

The final inequality is the contrapositive of the change-detection rule, applied to the two displayed intervals in the current phase of $\hat{a}_t$. Since $\hat{a}_t \in \mathcal{G}_t \subseteq \mathcal{E}_t$, Lemma B.9 bounds the conditional expectation of the sum of (1a) by $\widetilde{O}\left(\sqrt{MTD}\right)$.

Bounding Term (1c). The same argument for a_t, using equation 6, gives

$$\begin{aligned}
\text{(1c)} &\le \hat{\mu}^{(+1)}_{a_t}(I_t(a_t)) - \Delta_t(a_t) - \hat{\mu}^{(-1)}_{a_t}(t) \\
&\le \hat{\mu}^{(-1)}_{a_t}(I_t(a_t)) - \hat{\mu}^{(+1)}_{a_t}(t) + 2\mathrm{CB}_{a_t}(t) \\
&\le 2\mathrm{CB}_{a_t}(t).
\end{aligned}$$

The played-arm confidence-sum bound, Lemma B.8, therefore bounds the conditional expectation of the sum of (1c) by $\widetilde{O}\left(\sqrt{TD}\right)$.

Bounding Term (1b). Because the bad arm a_t is eligible, let $j \in \{0, \dots, J_{\max}\}$ be the smallest scale such that $j \le i_t(a_t)$ and $s_{a_t,j} > 0$ after the round-t refreshes and before the play. By the definition of the dominance scale, $\Delta_t(a_t) \le 2^{-i_t(a_t)+1} \le 2^{-j+1}$, so $3\Delta_t(a_t) \le 6 \cdot 2^{-j}$. Assign this amount to the decrease of $s_{a_t,j}$ caused by the play.

Every such decrease has a unique most recent update of the same counter. An update sets the counter to $2^{2j+5}\Lambda$, so it can be followed by at most this many decreases before the next update. If a later update overwrites a positive value, unused counter units are discarded, which can only reduce the number of assigned decreases. Therefore, for every realization of the algorithm's random choices, the total dominance-gap charge assigned to one update at scale j is at most $2^{2j+5}\Lambda \cdot 6 \cdot 2^{-j}$.

The algorithm maintains only scales $j \in \{0, \dots, J_{\max}\}$. Because the counter-update coins are independent of $\mathcal{C}$, their probabilities are unchanged after conditioning on that event. Thus the

conditional expected charge generated by the updates at any one round is at most

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{a\in\mathcal{A}_m}\sum_{j=0}^{J_{\max}} 2^{-j}\sqrt{\frac{L_m+1}{TK_m}}\cdot 2^{2j+5}\Lambda\cdot 6\cdot 2^{-j}\\
&= 6\cdot 2^5\Lambda\cdot(J_{\max}+1)\sum_{m\in[M]} K_m\sqrt{\frac{L_m+1}{TK_m}}\\
&= 6\cdot 2^5\Lambda\cdot(J_{\max}+1)\sum_{m\in[M]}\sqrt{\frac{(L_m+1)K_m}{T}}\\
&\le 6\cdot 2^5\Lambda\cdot(J_{\max}+1)\cdot\sqrt{\frac{MD}{T}},
\end{aligned}$$

where the last step is Cauchy–Schwarz. Summing this bound over all rounds gives

$$\mathbb{E}\left[\sum_{\substack{t\in[T]:\\ a_t\in\mathcal{B}_t}} 3\Delta_t(a_t)\,\middle|\,\mathcal{C}\right] \le 6\cdot 2^5\Lambda(J_{\max}+1)\sqrt{MTD} = \widetilde{O}\Big(\sqrt{MTD}\Big).$$

Combining Terms (1a), (1b), (1c) via equation 7 completes the proof. □

Only the dominance-gap part was charged to counter refreshes. The comparison with the blocker's current mean used the eligible-arm confidence sum, and the comparison with the played arm's current mean used the played-arm sum. The next term has no counter charge, because both arms in the comparison are good.

B.5.2 Term (II): comparing good representatives

Lemma B.12 (Restatement of Lemma 4.5)**.** *For the concentration event $\mathcal{C}$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}[\mu_{\hat{a}_t^*}(t)-\mu_{\hat{a}_t}(t)]\,\middle|\,\mathcal{C}\right] \le \widetilde{O}\Big(\sqrt{MTD}\Big).$$

Proof. For every round t, both $\hat{a}_t$ and $\hat{a}_t^*$ are good by Proposition 4.3. If they are equal, the summand is zero. Otherwise, the concentration event applied to the MSI of each arm gives

$$\begin{aligned}
\mu_{\hat{a}_t^*}(t)-\mu_{\hat{a}_t}(t) &\le \hat{\mu}^{(+1)}_{\hat{a}_t^*}(t)-\hat{\mu}^{(-1)}_{\hat{a}_t}(t)\\
&= \hat{\mu}_{\hat{a}_t^*}(t)-\hat{\mu}_{\hat{a}_t}(t)+\mathrm{CB}_{\hat{a}_t^*}(t)+\mathrm{CB}_{\hat{a}_t}(t)\\
&\le 4\mathrm{CB}_{\hat{a}_t^*}(t)+4\mathrm{CB}_{\hat{a}_t}(t).
\end{aligned}$$

The final inequality follows because the goodness of $\hat{a}_t$ precludes the two MSI-history intervals from certifying it as bad:

$$\hat{\mu}^{(-3)}_{\hat{a}_t^*}(t) \le \hat{\mu}^{(+3)}_{\hat{a}_t}(t).$$

If either MSI history is empty, the conclusion follows instead from the unit reward range and the fact that the corresponding confidence bound is at least one.

Both representatives belong to $\mathcal{G}_t \subseteq \mathcal{E}_t$. Taking conditional expectations and applying Lemma B.9 therefore gives

$$\mathbb{E}\left[\sum_{t=1}^{T}\big[\mu_{\hat{a}_t^*}(t)-\mu_{\hat{a}_t}(t)\big]\,\middle|\,\mathcal{C}\right] \le 4\mathbb{E}\left[\sum_{t=1}^{T}\big[\mathrm{CB}_{\hat{a}_t^*}(t)+\mathrm{CB}_{\hat{a}_t}(t)\big]\,\middle|\,\mathcal{C}\right] = \widetilde{O}\Big(\sqrt{MTD}\Big).$$

□

The remaining term is nonzero only when the optimal arm is classified as bad. Its blocker is good, but the optimal arm need not be eligible. We separate the blocker's estimation error and the class-threshold contribution, then follow the monitoring counter to bound the remaining certificate residuals.

B.5.3 TERM (III): DELAYED DETECTION

Lemma B.13 (Restatement of Lemma 4.6)**.** *For the concentration event $\mathcal{C}$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}[\mu_{a_t^*}(t)-\mu_{\hat{a}_t^*}(t)]\,\middle|\,\mathcal{C}\right]\le\widetilde{O}\Big(\sqrt{MTD}\Big).$$

Proof. The difficulty in bounding this term is that an optimal arm can still be classified as bad: its certificate may rely on observations collected before its mean changed. Its representative is good, but the optimal arm itself need not be eligible. We therefore cannot apply the eligible-arm confidence bound directly to both arms.

We first separate the error in the blocking arm's historical estimate from the discrepancy in the optimal arm's certificate. The former is controlled by the eligible-arm confidence sum. We then split the latter into small contributions, bounded by class thresholds, and large residuals that require monitoring. For the large residuals, we fix a true stationary segment and a monitoring scale, and consider three periods: before the first counter refresh, after that refresh but before the next reset, and after the reset. We bound the waiting time in the first period, prove eligibility on the relevant rounds in the second, and show that no positive residual remains in the third. Finally, we sum these contributions over classes, segments, and scales.

Throughout, pointwise inequalities are proved on $\mathcal{C}$; expectations retain the randomness of the interaction conditional on this event.

Step 1: Express the loss through the selected certificate. Recall that $\mathcal{U}:=\{t\in[T]:a_t^*\in\mathcal{B}_t\}$ is the set of rounds on which the optimal arm is classified as bad. Outside this set, $\hat{a}_t^*=a_t^*$, so Term (iii) is zero.

For $t\in\mathcal{U}$, the representative is the blocking arm $\hat{a}_t^*=\bar{a}_t(a_t^*)$. By Proposition 4.3, it belongs to $\mathcal{G}_t\subseteq\mathcal{E}_t$. Let the selected certificate be $\big(\hat{a}_t^*,I_t(a_t^*),\overline{I}_t(a_t^*)\big)$. The second interval contains the observations used for the optimal arm; the third contains those used for its blocking arm. Define the certificate residual by

$$r_t:=\mu_{a_t^*}(t)-\hat{\mu}^{(+1)}_{a_t^*}(I_t(a_t^*))\qquad(t\in\mathcal{U}).$$

This compares the optimal arm's current mean with its historical certificate upper bound, not with another arm's current mean.

By the definition of the dominance gap,

$$\Delta_t(a_t^*)=\hat{\mu}^{(-1)}_{\hat{a}_t^*}(\overline{I}_t(a_t^*))-\hat{\mu}^{(+1)}_{a_t^*}(I_t(a_t^*))\ge 0. \tag{8}$$

Subtracting this identity from the definition of r_t gives

$$\mu_{a_t^*}(t)-\hat{\mu}^{(-1)}_{\hat{a}_t^*}(\overline{I}_t(a_t^*))=r_t-\Delta_t(a_t^*). \tag{9}$$

Adding the difference between the blocker's certificate lower bound and its current mean to both sides recovers the Term (iii) summand. This separates the blocking-arm comparison from the residual difference $r_t-\Delta_t(a_t^*)$.

Step 2: Separate small residuals from those requiring monitoring. For each class $m\in[M]$, recall the count

$$T_m:=\big|\{t\in[T]:a_t^*\in\mathcal{A}_m\}\big|.$$

These counts include all rounds, not only those in $\mathcal{U}$, and satisfy $\sum_{m=1}^{M}T_m=T$. Recall also $\Lambda=\lceil\log(2TK)\rceil$, where log is the natural logarithm, and the thresholds

$$\zeta_m:=\begin{cases}\Lambda\sqrt{\dfrac{MK_m(L_m+1)}{T_m}}, & T_m>0,\\ 0, & T_m=0.\end{cases}$$

If $r_t \le \Delta_t(a_t^*)$, the residual difference is nonpositive. If r_t exceeds that gap but not its class threshold, the difference is at most the threshold. We therefore need the monitoring argument only for the remaining residuals, indexed by

$$\mathcal{U}_{\text{large}} := \left\{ t \in \mathcal{U} : r_t > \max\left\{ \Delta_t(a_t^*), \zeta_{\text{class}(a_t^*)} \right\} \right\}.$$

On this set, the residual difference is at most r_t. There are at most T_m threshold contributions from class m. Summing equation 9 after adding the blocking-arm comparison therefore yields

$$\begin{aligned}
&\sum_{t=1}^{T} \left[\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \right] \\
&\quad \le \underbrace{\sum_{t \in \mathcal{U}} \left[\hat{\mu}_{\hat{a}_t^*}^{(-1)}(\overline{I}_t(a_t^*)) - \mu_{\hat{a}_t^*}(t) \right]}_{\text{(3a): blocking-arm comparison}} \\
&\qquad + \underbrace{\sum_{m=1}^{M} T_m \zeta_m}_{\text{(3b): class thresholds}} + \underbrace{\sum_{t \in \mathcal{U}_{\text{large}}} r_t}_{\text{(3c): large residuals}} .
\end{aligned} \tag{10}$$

We next bound these three contributions.

Step 3: Bound the blocking-arm comparison and thresholds. For brevity, denote the conditional expected eligible-arm confidence sum by

$$S_{\text{elig}} := \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{E}_t} \text{CB}_a(t) \,\middle|\, \mathcal{C} \right].$$

As throughout the analysis, a statistic evaluated at t uses only observations before t in the arm's current MSI.

Term (3a). Fix $t \in \mathcal{U}$. The blocker's class was not reset at round t: after a reset its phase history would be empty, making its three-radius lower bound negative and hence unable to exceed the optimal arm's nonnegative three-radius upper bound. The reset test in equation 1 is therefore false for this class. Applied to its certificate interval and current MSI history, it gives

$$\hat{\mu}_{\hat{a}_t^*}^{(-1)}(\overline{I}_t(a_t^*)) \le \hat{\mu}_{\hat{a}_t^*}^{(+1)}(t) \le \mu_{\hat{a}_t^*}(t) + 2\text{CB}_{\hat{a}_t^*}(t).$$

The second inequality follows from concentration on the current MSI. If that history has no observations, its upper bound is its confidence radius, which is at least one. The first inequality then follows from the certificate lower bound being at most one, and the second follows from the mean being nonnegative. Thus the same comparison holds in the zero-observation case.

Subtracting the blocker's mean, summing, and using its eligibility, we obtain

$$\mathbb{E}\left[\sum_{t \in \mathcal{U}} \left[\hat{\mu}_{\hat{a}_t^*}^{(-1)}(\overline{I}_t(a_t^*)) - \mu_{\hat{a}_t^*}(t) \right] \,\middle|\, \mathcal{C} \right] \le 2S_{\text{elig}}.$$

Term (3b). The threshold definition and Cauchy–Schwarz give

$$\begin{aligned}
\sum_{m=1}^{M} T_m \zeta_m &= \Lambda \sqrt{M} \sum_{m=1}^{M} \sqrt{T_m K_m (L_m + 1)} \\
&\le \Lambda \sqrt{M} \sqrt{\sum_{m=1}^{M} T_m} \sqrt{\sum_{m=1}^{M} K_m (L_m + 1)} \\
&= \Lambda \sqrt{MTD}.
\end{aligned}$$

The zero-count classes contribute zero in the first equality. The last equality uses the total count T and the definition of D. It remains to bound Term (3c).

Step 4: Partition the large-residual rounds. Recall the scale sets from equation 5:

$$\mathcal{U}_{m,j} = \{t \in \mathcal{U}_{\text{large}} : a_t^* \in \mathcal{A}_m, \quad 2^{-j} < r_t \le 2^{1-j}\}, \qquad j \ge 1.$$

On $\mathcal{U}_{\text{large}}$, we have $0 < r_t \le 1$: the true mean is at most one and the certificate upper bound is nonnegative. Each such round therefore belongs to exactly one class and one residual scale. The algorithm maintains scales up to $J_{\max} = \lceil \log_2 T \rceil$. Residuals at most $2^{-J_{\max}}$ have total contribution at most $T\,2^{-J_{\max}} \le 1$. We treat the remaining scales $j \in \{1, \dots, J_{\max}\}$ below; this set is empty when $T = 1$.

For class m, write its true stationary segments as

$$J_{m,i} = [b_{m,i}, e_{m,i}], \qquad i \in \{0, \dots, |\mathcal{L}_m|\}, \qquad d_{m,i} := e_{m,i} - b_{m,i} + 1.$$

Thus $b_{m,i}$, $e_{m,i}$, and $d_{m,i}$ are the segment's first round, last round, and length. These segments have constant class means but, unlike MSIs, may contain algorithmic resets. Let $a^\star_{m,i}$ be the best arm in class m on this segment, using the same fixed arm order as a_t^*. The segment and its best arm are fixed by the instance. In particular,

$$t \in J_{m,i}, \quad a_t^* \in \mathcal{A}_m \quad \Longrightarrow \quad a_t^* = a^\star_{m,i}.$$

The global optimal arm may leave this class as other classes change; whenever it belongs to this class, it is the same class-best arm.

Fix a class m, segment index i, and maintained scale j. For each $t \in J_{m,i}$, define $X_{m,i,j,t}$ to be the indicator that the counter $s_{a^\star_{m,i},j}$ is refreshed at round t. The refresh probability is

$$p_{m,j} := 2^{-j}\sqrt{\frac{L_m + 1}{TK_m}}.$$

These indicators are independent across rounds and independent of $\mathcal{C}$, which depends only on the presampled reward sequences. Each refresh sets the counter to $2^{2j+5}\Lambda$, whether the arm is good or bad. Since $L_m + 1 \le T$ and $K_m \ge 1$, we have $0 < p_{m,j} \le 1$.

Let the first refresh round in this segment be

$$\sigma_{m,i,j} := \min\big(\{t \in J_{m,i} : X_{m,i,j,t} = 1\} \cup \{e_{m,i} + 1\}\big).$$

The value $e_{m,i} + 1$ means that no refresh occurs in the segment. Let $\eta_{m,i,j}$ be the first class-m reset round in the segment strictly after $\sigma_{m,i,j}$, or $e_{m,i} + 1$ if there is no such reset. Define

$$\begin{aligned}
\mathcal{R}^{(1)}_{m,i,j} &:= \{t \in J_{m,i} : t < \sigma_{m,i,j}\},\\
\mathcal{R}^{(2)}_{m,i,j} &:= \{t \in J_{m,i} : \sigma_{m,i,j} \le t < \eta_{m,i,j}\},\\
\mathcal{R}^{(3)}_{m,i,j} &:= \{t \in J_{m,i} : \eta_{m,i,j} \le t\}.
\end{aligned}$$

These regions are, respectively, before refresh, after refresh but before the next reset, and after that reset. They partition the segment. If no refresh occurs, only Region 1 is nonempty; if no subsequent reset occurs, Region 3 is empty. A reset at the start of the refresh round occurs before the refresh and does not end Region 2.

For $k \in \{1, 2, 3\}$, define the residual sum in Region k by

$$Z^{(k)}_{m,i,j} := \sum_{t \in \mathcal{U}_{m,j} \cap \mathcal{R}^{(k)}_{m,i,j}} r_t.$$

All these sums are nonnegative, and empty sums are zero. Every large-residual round at a maintained scale has a unique class, segment, scale, and region, so

$$\sum_{m=1}^{M} \sum_{j=1}^{J_{\max}} \sum_{t \in \mathcal{U}_{m,j}} r_t = \sum_{m=1}^{M} \sum_{i=0}^{|\mathcal{L}_m|} \sum_{j=1}^{J_{\max}} \sum_{k=1}^{3} Z^{(k)}_{m,i,j}.$$

We now bound these region sums in chronological order.

Step 5: Bound the three monitoring periods.

Region 1: before the first refresh. Define the waiting count

$$W_{m,i,j} := |\mathcal{R}^{(1)}_{m,i,j}| = \sigma_{m,i,j} - b_{m,i}. \tag{11}$$

This counts all rounds before the first refresh, excluding the refresh round. It is zero if the first round refreshes the counter, and equals $d_{m,i}$ if the counter is never refreshed in the segment. For an integer $q \in \{1, \dots, d_{m,i}\}$, at least q rounds precede the refresh exactly when the first q coins fail:

$$\{W_{m,i,j} \geq q\} = \bigcap_{u=b_{m,i}}^{b_{m,i}+q-1} \{X_{m,i,j,u} = 0\}.$$

Independence of the refresh coins, also after conditioning on $\mathcal{C}$, gives

$$\mathbb{P}(W_{m,i,j} \geq q \mid \mathcal{C}) = \begin{cases} (1-p_{m,j})^q, & 1 \leq q \leq d_{m,i}, \\ 0, & q > d_{m,i}, \end{cases} \tag{12}$$

for every integer $q \geq 1$. Using the tail-sum identity (Lattimore & Szepesvári, 2020, Proposition 2.8),

$$\begin{aligned} \mathbb{E}[W_{m,i,j} \mid \mathcal{C}] &= \sum_{q=1}^{d_{m,i}} (1-p_{m,j})^q \\ &\leq \sum_{q=1}^{\infty} (1-p_{m,j})^q = \frac{1-p_{m,j}}{p_{m,j}} \leq \frac{1}{p_{m,j}}. \end{aligned}$$

This estimate does not require a refresh to occur before the segment ends. Each summand of $Z^{(1)}_{m,i,j}$ is at most 2^{1-j}, and there are at most $W_{m,i,j}$ such summands. Hence, pathwise, $Z^{(1)}_{m,i,j} \leq 2^{1-j} W_{m,i,j}$, and therefore

$$\mathbb{E}[Z^{(1)}_{m,i,j} \mid \mathcal{C}] \leq \frac{2^{1-j}}{p_{m,j}} = 2\sqrt{\frac{TK_m}{L_m+1}}.$$

This uses no independence between the large-residual rounds and the refresh coins; the bound follows by counting all waiting rounds. Summing over at most $L_m + 1$ segments in class m and the $J_{\max}$ maintained residual scales gives

$$\begin{aligned} &\mathbb{E}\left[\sum_{m=1}^{M} \sum_{i=0}^{|\mathcal{L}_m|} \sum_{j=1}^{J_{\max}} Z^{(1)}_{m,i,j} \,\middle|\, \mathcal{C} \right] \\ &\leq 2J_{\max} \sum_{m=1}^{M} \sqrt{TK_m(L_m+1)} \\ &\leq 2J_{\max}\sqrt{MTD}. \end{aligned}$$

The last inequality is Cauchy–Schwarz over the classes, using $\sum_m (L_m+1)K_m = D$.

Region 2: after refresh and before the next reset. Fix a round $t \in \mathcal{U}_{m,j} \cap \mathcal{R}^{(2)}_{m,i,j}$, and write $s := \sigma_{m,i,j}$. The optimal arm at t is $a^\star_{m,i}$, whose scale-j counter was refreshed at s. We show first that this counter is still positive, then that the arm is eligible, and finally that its residual is bounded by its current-MSI confidence radius.

The counter cannot be exhausted on this round. Counter values here are evaluated after the round-t refreshes and before the play. If $t = s$, the counter is positive by its refresh. Otherwise, suppose it were zero. Since the most recent refresh at or after s set it to $2^{2j+5}\Lambda$, it must have been decremented at least this many times before t. Each decrement requires a pull of this arm. Thus

$$\mathrm{CB}_{a^*_t}([s, t-1]) \leq \sqrt{\frac{8\log(2TK)}{2^{2j+5}\Lambda}} \leq 2^{-j-1},$$

where the second inequality uses $\Lambda \geq \log(2TK)$. There is no class-m change on $[s, t-1]$. Concentration and $r_t > 2^{-j}$ therefore imply

$$\begin{aligned} \hat{\mu}^{(-1)}_{a^*_t}([s, t-1]) &\geq \mu_{a^*_t}(t) - 2\mathrm{CB}_{a^*_t}([s, t-1]) \\ &\geq \mu_{a^*_t}(t) - 2^{-j} \\ &> \hat{\mu}^{(+1)}_{a^*_t}(I_t(a^*_t)). \end{aligned}$$

Both intervals lie in the current class phase: no reset occurs strictly after s and before or at this round in Region 2. The displayed inequality would therefore trigger the reset test at t, contradicting membership in Region 2. Consequently, $s_{a_t^*,j} > 0$.

The positive counter makes the optimal arm eligible. Membership in $\mathcal{U}_{m,j}$ gives

$$0 < \Delta_t(a_t^*) < r_t \le 2^{1-j}.$$

Taking negative base-two logarithms gives $-\log_2 \Delta_t(a_t^*) > j - 1$, so

$$i_t(a_t^*) = \lceil -\log_2 \Delta_t(a_t^*) \rceil \ge j.$$

The arm is bad, its scale-j counter is positive, and j is a maintained scale not exceeding its dominance scale. These are exactly the conditions for its inclusion in $\mathcal{E}_t$.

The residual is bounded by a confidence radius. Because the optimal arm is bad at t, its class was not reset at the start of that round: a newly reset arm has no observations to support a bad-arm certificate. In particular, a reset at s is impossible when $t = s$. If $N_{a_t^*}(t) > 0$, concentration on the current MSI and the absence of a reset witness give, respectively,

$$\mu_{a_t^*}(t) \le \hat{\mu}^{(+1)}_{a_t^*}(t), \qquad \hat{\mu}^{(-1)}_{a_t^*}(t) \le \hat{\mu}^{(+1)}_{a_t^*}(I_t(a_t^*)).$$

Subtracting the certificate upper bound from the first inequality and using the second gives

$$r_t \le \hat{\mu}^{(+1)}_{a_t^*}(t) - \hat{\mu}^{(-1)}_{a_t^*}(t) = 2\mathrm{CB}_{a_t^*}(t).$$

If the current-MSI count is zero, the same conclusion follows from $r_t \le 1$ and $\mathrm{CB}_{a_t^*}(t) \ge 1$. Since this arm is eligible, each Region 2 residual is at most twice the largest eligible-arm radius at that round. Every relevant round has a unique class, segment, and residual scale, so

$$\sum_{m=1}^{M} \sum_{i=0}^{|\mathcal{L}_m|} \sum_{j=1}^{J_{\max}} Z^{(2)}_{m,i,j} \le 2 \sum_{t=1}^{T} \max_{a \in \mathcal{E}_t} \mathrm{CB}_a(t).$$

There is no additional factor for the number of scales: no round is counted at two different residual scales. Taking conditional expectations bounds this contribution by $2S_{\mathrm{elig}}$.

Region 3: after the reset. Suppose that $t \in \mathcal{U}_{m,j} \cap \mathcal{R}^{(3)}_{m,i,j}$. Class m has already reset within its current true stationary segment. The optimal arm's current certificate interval must therefore lie entirely inside that segment:

$$I_t(a_t^*) \subseteq [\phi_m, t-1] \subseteq J_{m,i}.$$

Concentration now gives

$$\mu_{a_t^*}(t) \le \hat{\mu}^{(+1)}_{a_t^*}(I_t(a_t^*)), \qquad r_t \le 0.$$

This contradicts $r_t > 0$ on $\mathcal{U}_{m,j}$. Hence the intersection is empty and $Z^{(3)}_{m,i,j} = 0$.

Step 6: Combine the bounds. The Region 1 contribution is at most $2J_{\max}\sqrt{MTD}$ in conditional expectation, Region 2 contributes at most $2S_{\mathrm{elig}}$, and Region 3 contributes zero. Adding the at-most-one contribution of residuals below the smallest maintained scale proves

$$\mathbb{E}\left[\sum_{t \in \mathcal{U}_{\mathrm{large}}} r_t \,\middle|\, \mathcal{C} \right] \le 2S_{\mathrm{elig}} + 2J_{\max}\sqrt{MTD} + 1. \tag{13}$$

This completes the bound on Term (3c). Substituting it and the two bounds from Step 3 into equation 10, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \left[\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \right] \,\middle|\, \mathcal{C} \right] \le 4S_{\mathrm{elig}} + (\Lambda + 2J_{\max})\sqrt{MTD} + 1.$$

The coefficient four consists of two eligible-arm confidence sums from the blocking-arm comparison and two from Region 2. Finally, Lemma B.9 bounds S_{elig} by $O(\log(2TK)\sqrt{MTD})$. Since Λ and $J_{\max}$ are also $O(\log(2TK))$, the preceding display proves the claimed $\widetilde{O}(\sqrt{MTD})$ bound. □

B.6 PROOF OF THE MAIN REGRET THEOREM

The three contributions are controlled by two confidence sums, the counter-refresh charge, the class thresholds, and the first-refresh waiting bound. We retain their explicit constants below and then remove the conditioning on concentration.

Theorem B.14 (Restatement of Theorem 3.1)**.** *For every admissible instance, Algorithm* `EBF-H` *satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le 4096 \log^2(2TK)\sqrt{MTD} + 1.$$

Equivalently, its expected dynamic regret is $\widetilde{O}(\sqrt{MTD})$.

Proof. Fix an admissible instance. Recall that T is the horizon, $K \ge 2$ is the number of arms, and $M \ge 1$ is the number of classes. The arm set of class m is $\mathcal{A}_m$, with K_m arms and change budget L_m. Write

$$D = \sum_{m=1}^{M} (L_m + 1)K_m, \qquad \ell := \log(2TK), \qquad \Gamma := \sqrt{MTD}.$$

Here and below, $\log$ denotes the natural logarithm. Recall also the algorithm's parameters

$$\Lambda = \lceil \ell \rceil, \qquad J_{\max} = \lceil \log_2 T \rceil.$$

If $T = 1$, the claim follows immediately from $0 \le \mathrm{Regret}_T \le 1$. We therefore assume $T \ge 2$.

Recall $\mathcal{C}$ is the concentration event of Lemma 4.1. We first bound regret conditional on $\mathcal{C}$, and then account for its complement.

Step 1: Collect the two confidence-sum bounds. For an arm a and round t, $\mathrm{CB}_a(t)$ denotes its confidence radius computed using observations before t in the current *maximal static interval* (MSI) of its class. An MSI starts at round one and whenever the class undergoes a true mean change or an algorithmic reset. Thus both the class means and its phase start remain fixed within an MSI; see Section B.1.

Let $\mathcal{E}_t$ be the algorithm's eligible set at round t, and let a_t be the arm played at that round. Define the two *conditional expected* confidence sums

$$S_{\mathrm{elig}} := \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{E}_t} \mathrm{CB}_a(t) \,\middle|\, \mathcal{C}\right], \qquad S_{\mathrm{play}} := \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) \,\middle|\, \mathcal{C}\right].$$

On $\mathcal{C}$, Lemma B.6 bounds the total number of arm–MSI pairs by

$$\sum_{m=1}^{M} (2L_m + 1)K_m \le 2D.$$

The explicit bound proved in Lemma B.9 therefore gives

$$S_{\mathrm{elig}} \le 32\sqrt{\ell(J_{\max} + 1)}\,\Gamma.$$

For the played arms, Lemma B.8 gives, on every realization in $\mathcal{C}$,

$$\sum_{t=1}^{T} \mathrm{CB}_{a_t}(t) \le 3\sqrt{8\ell T \sum_{m=1}^{M} (2L_m + 1)K_m} \le 12\sqrt{\ell TD}.$$

Taking conditional expectations and using $M \ge 1$ yields

$$S_{\mathrm{play}} \le 12\sqrt{\ell}\,\Gamma.$$

Step 2: Separate the three regret contributions. Write a_t^* for the optimal arm at round t, and $\mu_a(t)$ for the mean reward of arm a at that round. Recall that $\hat{a}_t$ equals a_t when the played arm is good and equals its blocking arm otherwise. The representative $\hat{a}_t^*$ is defined in the same way for a_t^*. By Proposition 4.3, both representatives are good and hence belong to $\mathcal{E}_t$.

Recall the cumulative versions of Terms (i), (ii), and (iii):

$$R_1 := \sum_{t=1}^{T} \big(\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)\big),$$
$$R_2 := \sum_{t=1}^{T} \big(\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)\big),$$
$$R_3 := \sum_{t=1}^{T} \big(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\big).$$

Lemma B.10 gives the pathwise identity

$$\mathrm{Regret}_T = R_1 + R_2 + R_3.$$

The individual contributions need not all be nonnegative; we bound each one from above.

Step 3: Bound Term (i), the cost of playing bad arms. Let $\mathcal{B}_t$ denote the bad-arm set, and let $\Delta_t(a)$ be the dominance gap of a bad arm a, defined by its maximizing classification certificate in Section 3. Term (i) is zero when $a_t \notin \mathcal{B}_t$. Otherwise, the pointwise bound proved in Section B.5.1 is

$$\mu_{\hat{a}_t}(t) - \mu_{a_t}(t) \le 2\mathrm{CB}_{\hat{a}_t}(t) + 3\Delta_t(a_t) + 2\mathrm{CB}_{a_t}(t).$$

The first confidence radius is at most the largest eligible-arm radius, because $\hat{a}_t \in \mathcal{E}_t$.

The counter-assignment argument in Section B.5.1 bounds the remaining dominance-gap contribution by

$$\mathbb{E}\left[\sum_{\substack{1\le t\le T\\ a_t\in\mathcal{B}_t}} 3\Delta_t(a_t) \,\middle|\, \mathcal{C}\right] \le 6\cdot 2^5\Lambda(J_{\max}+1)\sum_{m=1}^{M}\sqrt{T(L_m+1)K_m}$$
$$\le 192\Lambda(J_{\max}+1)\Gamma.$$

The last inequality is Cauchy–Schwarz over the classes. Summing the pointwise bound and taking conditional expectations therefore yields

$$\mathbb{E}[R_1 \mid \mathcal{C}] \le 2S_{\mathrm{elig}} + 2S_{\mathrm{play}} + 192\Lambda(J_{\max}+1)\Gamma.$$

Step 4: Bound Term (ii), the comparison of good representatives. The pointwise comparison proved in Section B.5.2 gives

$$\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t) \le 4\mathrm{CB}_{\hat{a}_t^*}(t) + 4\mathrm{CB}_{\hat{a}_t}(t).$$

Both representatives are eligible. Each radius is therefore at most $\max_{a\in\mathcal{E}_t}\mathrm{CB}_a(t)$, and hence

$$\mathbb{E}[R_2 \mid \mathcal{C}] \le 8S_{\mathrm{elig}}.$$

Step 5: Bound Term (iii), the cost of delayed detection. We recall the residual notation from Section B.1. Let

$$\mathcal{U} := \{t \in \{1,\dots,T\} : a_t^* \in \mathcal{B}_t\}.$$

For $t \in \mathcal{U}$, let $I_t(a_t^*)$ be the optimal arm's certificate interval. For any arm a and interval I, $\hat{\mu}_a^{(+1)}(I)$ denotes the empirical mean of arm a on I plus its confidence radius. Define

$$r_t := \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)).$$

Thus r_t is the difference between the optimal arm's current mean and its historical certificate upper bound, not its instantaneous regret.

For each class m, recall the optimal-class count and threshold

$$T_m := \big|\{t \in \{1,\dots,T\} : a_t^* \in \mathcal{A}_m\}\big|, \qquad \zeta_m := \begin{cases} \Lambda\sqrt{\dfrac{MK_m(L_m+1)}{T_m}}, & T_m > 0,\\ 0, & T_m = 0. \end{cases}$$

Each round has a unique optimal arm under the fixed tie-breaking rule, so $\sum_{m=1}^{M} T_m = T$. Write $\mathrm{class}(a)$ for the class containing arm a, and define

$$\mathcal{U}_{\mathrm{large}} := \left\{ t \in \mathcal{U} : r_t > \max\left\{ \Delta_t(a_t^*), \zeta_{\mathrm{class}(a_t^*)} \right\} \right\}.$$

For $t \in \mathcal{U}$, the blocking-certificate comparison in Section B.5.3 gives

$$\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \le r_t - \Delta_t(a_t^*) + 2\mathrm{CB}_{\hat{a}_t^*}(t).$$

Outside $\mathcal{U}$, Term (iii) is zero. If $r_t \le \Delta_t(a_t^*)$, the residual difference in this display is nonpositive. If the residual exceeds the dominance gap but not the class threshold, it is bounded by that threshold. All other rounds belong to $\mathcal{U}_{\mathrm{large}}$, where we upper-bound the residual difference by r_t. Consequently,

$$\mathbb{E}[R_3 \mid \mathcal{C}] \le 2S_{\mathrm{elig}} + \sum_{m=1}^{M} T_m \zeta_m + \mathbb{E}\left[\sum_{t \in \mathcal{U}_{\mathrm{large}}} r_t \,\middle|\, \mathcal{C} \right].$$

The threshold contribution. Using the threshold definition and Cauchy–Schwarz,

$$\begin{aligned}
\sum_{m=1}^{M} T_m \zeta_m &= \Lambda\sqrt{M} \sum_{m=1}^{M} \sqrt{T_m K_m (L_m + 1)} \\
&\le \Lambda\sqrt{M} \sqrt{\sum_{m=1}^{M} T_m} \sqrt{\sum_{m=1}^{M} K_m (L_m + 1)} \\
&= \Lambda\Gamma.
\end{aligned}$$

The large-residual contribution. We use equation 13, proved within Section B.5.3. Group rounds in $\mathcal{U}_{\mathrm{large}}$ by their residual scale j, defined by $2^{-j} < r_t \le 2^{1-j}$. For each maintained scale $j \in \{1, \dots, J_{\max}\}$ and true class-stationary segment, the proof separates rounds before the first scale-j counter refresh, rounds after that refresh but before the next class reset, and rounds after that reset.

Before the first refresh, the conditional expected residual contribution of each class-m segment and scale is at most

$$2\sqrt{\frac{TK_m}{L_m + 1}}.$$

Since class m has at most $L_m + 1$ true stationary segments, summing gives

$$2J_{\max} \sum_{m=1}^{M} \sqrt{TK_m(L_m + 1)} \le 2J_{\max}\Gamma.$$

On the large-residual rounds at this scale after the refresh and before a class reset, the optimal arm is eligible and its residual is at most twice its current MSI confidence radius. Each large-residual round has a unique residual scale, so the total conditional expected contribution of these rounds is at most $2S_{\mathrm{elig}}$, with *no additional scale factor*. After a class reset within the segment, concentration on the new phase makes the residual nonpositive whenever the optimal arm belongs to this class. Thus no large-residual contribution from this class remains in that segment. Finally, residuals at most $2^{-J_{\max}}$ sum to at most $T\,2^{-J_{\max}} \le 1$. Combining these contributions gives

$$\mathbb{E}\left[\sum_{t \in \mathcal{U}_{\mathrm{large}}} r_t \,\middle|\, \mathcal{C} \right] \le 2S_{\mathrm{elig}} + 2J_{\max}\Gamma + 1.$$

Substituting the threshold and residual bounds into the bound for R_3, we obtain

$$\mathbb{E}[R_3 \mid \mathcal{C}] \le 4S_{\mathrm{elig}} + (\Lambda + 2J_{\max})\Gamma + 1.$$

The coefficient four consists of two confidence-sum contributions from the blocking-arm comparison and two from the post-refresh residuals.

Step 6: Combine the three bounds and simplify the constants. Adding the bounds for R_1, R_2, R_3 gives

$$\mathbb{E}[\mathrm{Regret}_T \mid \mathcal{C}] \le (2+8+4)S_{\mathrm{elig}} + 2S_{\mathrm{play}} + \big[192\Lambda(J_{\max}+1) + \Lambda + 2J_{\max}\big]\Gamma + 1.$$

Using Step 1 and the identities $14 \cdot 32 = 448$ and $2 \cdot 12 = 24$, this becomes

$$\mathbb{E}[\mathrm{Regret}_T \mid \mathcal{C}] \le \Big[448\sqrt{\ell(J_{\max}+1)} + 24\sqrt{\ell} + 192\Lambda(J_{\max}+1) + \Lambda + 2J_{\max}\Big]\Gamma + 1.$$

Since $K \ge 2$, $T \ge 1$, and $D \ge K$, we have

$$\ell \ge \log 4 > 1, \qquad \Lambda \le 2\ell, \qquad J_{\max}+1 \le 3\ell, \qquad \Gamma \ge 1.$$

Hence, $192\Lambda(J_{\max}+1) \le 1152\ell^2$. For the remaining terms, we use $\sqrt{\ell} \le \ell \le \ell^2$ and $2J_{\max} \le 6\ell$. The additive one can also be absorbed because $1 \le \ell^2\Gamma$. Therefore,

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}_T \mid \mathcal{C}] &\le \big(448\sqrt{3} + 24 + 1152 + 2 + 6\big)\ell^2\Gamma + 1 \\ &\le \big(448\sqrt{3} + 1185\big)\ell^2\Gamma \\ &\le 4096\ell^2\Gamma.\end{aligned}$$

Step 7: Account for failure of concentration. By Lemma 4.1, $\mathbb{P}(\mathcal{C}^c) \le T^{-1}$, where $\mathcal{C}^c$ denotes the complementary event. Every instantaneous regret lies in $[0,1]$, so $0 \le \mathrm{Regret}_T \le T$ on every realization. Splitting the expectation according to whether $\mathcal{C}$ occurs gives

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}_T] &\le \mathbb{P}(\mathcal{C})\mathbb{E}[\mathrm{Regret}_T \mid \mathcal{C}] + T\mathbb{P}(\mathcal{C}^c) \\ &\le 4096\ell^2\Gamma + 1 \\ &= 4096\log^2(2TK)\sqrt{MTD} + 1,\end{aligned}$$

as stated. □

C Proofs for Section 5

We prove the two lower bounds from Section 5. Both are built from the same basic statistical obstacle: over a block of b rounds, identifying one slightly better arm among k candidates costs $\Omega(\sqrt{kb})$ regret. The direct-sum lower bound repeats this difficulty independently across classes, while the class-search lower bound additionally hides which class contains the improved arm.

All constructions below use Bernoulli rewards and are sampled before interaction, and hence satisfy the stochastic model of Section 2. The results apply to arbitrary randomized learners, even when the class budgets are known.

C.1 Auxiliary Lower-Bound Tools

We first prove the block-level lower bound used in both constructions. For $p, q \in (0,1)$, write

$$\mathrm{kl}(p\|q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

For $0 < \epsilon \le 1/4$, we use the elementary bounds

$$\mathrm{kl}(1/2\|1/2+\epsilon) \le 4\epsilon^2, \qquad \mathrm{kl}(1/2-\epsilon\|1/2+\epsilon) \le 16\epsilon^2.$$

C.1.1 Axillary Lower Bounds

We first prove a conditional Bernoulli version of the standard minimax lower bound for stochastic multi-armed bandits (Lattimore & Szepesvári, 2020, Chapter 15); see also Auer et al. (2002). We include it because we need a version that can be applied conditionally after an arbitrary preceding history and, for singleton classes, relative to an outside baseline arm.

Lemma C.1 (Conditional block lower bound)**.** *Fix any candidate set S of $k \geq 1$ arms and a block of $b \geq k$ rounds. If $k = 1$, assume that the bandit contains at least one arm outside S. There is a prior over Bernoulli mean assignments during the block for which every possibly randomized learner has Bayes regret at least*

$$\frac{1}{32\sqrt{2}}\sqrt{kb}.$$

The statement remains valid conditional on any preceding history independent of the random parameter assigned to the block.

Moreover, if S is one class and the block is bracketed by baseline rounds on which every arm has mean $1/2$, every assignment in the prior creates exactly two change rounds in that class.

Proof. We first consider $k \geq 2$. Choose one arm J uniformly from S. During the block, arm J has mean $1/2 + \epsilon$, while every other arm has mean $1/2$.

For each $j \in [k]$, let P_j be the block-transcript law when arm j is improved, let $Z_j \in [0, b]$ be its number of pulls, and let $\mathbb{E}_j$ denote expectation under P_j. Let P_0 be the all-baseline law and let $\mathbb{E}_0$ denote its expectation.

Once the transcript preceding the block is fixed, the learner follows the same decision rule under P_0 and under every P_j. The standard bandit divergence decomposition (Lattimore & Szepesvári, 2020, Lemma 15.1) therefore gives

$$\mathrm{KL}(P_0 \| P_j) = \mathbb{E}_0 Z_j \ \mathrm{kl}(1/2 \| 1/2 + \epsilon) \leq 4\epsilon^2 \mathbb{E}_0 Z_j.$$

Since $0 \leq Z_j \leq b$, Pinsker's inequality and the standard expectation–total-variation bound (Lattimore & Szepesvári, 2020, Eq. (14.12)) imply

$$\mathbb{E}_j Z_j \leq \mathbb{E}_0 Z_j + b\,\mathrm{TV}(P_j, P_0) \leq \mathbb{E}_0 Z_j + b\epsilon\sqrt{2\mathbb{E}_0 Z_j}.$$

Because only one arm is pulled in each round,

$$\sum_{j=1}^{k} \mathbb{E}_0 Z_j \leq b.$$

Averaging over j and applying Cauchy–Schwarz gives

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{E}_j Z_j \leq \frac{b}{k} + b\epsilon\sqrt{\frac{2b}{k}}.$$

Choose

$$\epsilon = \frac{1}{8\sqrt{2}}\sqrt{\frac{k}{b}}.$$

Since $b \geq k$, we have $\epsilon \leq 1/4$, and since $k \geq 2$, the preceding display is at most $5b/8$. Hence the Bayes regret is

$$\frac{1}{k}\sum_{j=1}^{k} \epsilon\big(b - \mathbb{E}_j Z_j\big) \geq \frac{\epsilon b}{4} = \frac{1}{32\sqrt{2}}\sqrt{kb}.$$

It remains to handle $k = 1$. Let the unique arm in S have mean $1/2 + \epsilon$ or $1/2 - \epsilon$, each with probability $1/2$, while every arm outside S remains at mean $1/2$. Set

$$\epsilon = \frac{1}{8\sqrt{2b}}.$$

Let P_+ and P_- be the corresponding transcript laws, and let Z denote the number of pulls of the candidate arm. Under P_+, the expected regret is $\epsilon(b - \mathbb{E}_+ Z)$, whereas under P_- it is $\epsilon\mathbb{E}_- Z$. Moreover,

$$\mathrm{KL}(P_- \| P_+) \leq 16\epsilon^2 \mathbb{E}_- Z \leq 16\epsilon^2 b.$$

Pinsker's inequality therefore gives

$$\mathrm{TV}(P_+, P_-) \leq \frac{1}{4}.$$

Using again that $0 \le Z \le b$, the Bayes regret is at least

$$\begin{aligned}\frac{\epsilon}{2}\big(b - \mathbb{E}_+ Z + \mathbb{E}_- Z\big) &\ge \frac{\epsilon b}{2}\big(1 - \mathrm{TV}(P_+, P_-)\big) \\ &\ge \frac{3\epsilon b}{8} \\ &\ge \frac{1}{32\sqrt{2}}\sqrt{b}.\end{aligned}$$

The argument is unchanged after conditioning on any preceding history independent of the random parameter assigned to the block: the learner may enter the block in an arbitrary state, but the block parameter retains the prior above.

Finally, when S is a single class and the block is bracketed by baseline rounds, that class changes once when the block begins and once when it returns to the baseline. Thus the block uses exactly two change rounds in that class. □

Next, we add uncertainty over the identity of the active class, as stated in the next lemma. The argument is a variant of the standard minimax lower-bound construction for switching bandits, which partitions the horizon into hard stationary epochs and randomizes the identity of the improved arm (Auer et al., 2002; Besbes et al., 2014; Auer et al., 2019). We use a random permutation of the classes so that the same construction also respects the class-specific change budgets.

Lemma C.2 (Random-permutation block construction)**.** *Consider $M \ge 1$ classes with $k \ge 1$ designated arms in each class, where $Mk \ge 2$. For every horizon $h \ge 6M^2k$ and every learner, there exists an instance with at most two change rounds in each class such that the expected regret over these h rounds is at least*

$$\frac{1}{512}M\sqrt{kh}.$$

The same Bayes-regret bound holds conditional on any preceding history that is independent of the random parameters of the construction.

Proof. Partition the h rounds into an initial baseline block, M active blocks, and a terminal baseline block, each of length

$$B := \left\lfloor \frac{h}{M+2} \right\rfloor.$$

All remaining rounds are also baseline. On baseline rounds, every arm has mean $1/2$.

Before interaction, draw a uniformly random permutation of the M classes and independently choose one designated arm uniformly from each class. In active block u, the u-th class in the permutation is active. Let

$$s := M - u + 1$$

be the number of classes that have not yet appeared, and set

$$\epsilon_u := \frac{1}{8\sqrt{2}}\sqrt{\frac{sk}{B}}.$$

During this block, the selected arm of the active class has mean $1/2 + \epsilon_u$, while every other arm has mean $1/2$.

For the analysis, reveal to the learner the active class and improved arm after every completed active block. This can only make the problem easier. Conditional on the entire past and these revelations, the next active class is uniform among the s unused classes, and its improved arm is uniform among their sk designated arms.

Whenever $sk \ge 2$, this conditional block law is exactly the multiple-candidate construction of Lemma C.1 with sk candidate arms. When $sk = 1$, we use only the nonnegativity of regret and do not invoke the singleton construction of that lemma.

Since $h \ge 6M^2k$,

$$\frac{h}{M+2} \ge 2Mk, \qquad \text{and hence} \qquad B \ge Mk \ge sk.$$

Therefore, whenever $sk \geq 2$, the conditional expected regret in block u is at least

$$\frac{1}{32\sqrt{2}}\sqrt{skB}.$$

Suppose first that $M \geq 2$. Keeping only the stages with $s > M/2$,

$$\sum_{\substack{s\in[M]\\ sk\geq 2}} \sqrt{s} \geq \frac{M}{2}\sqrt{\frac{M}{2}} \geq \frac{M^{3/2}}{4}.$$

Also,

$$B = \left\lfloor \frac{h}{M+2} \right\rfloor \geq \frac{h}{2(M+2)} \geq \frac{h}{6M}.$$

Hence the total expected regret is at least

$$\frac{1}{32\sqrt{2}}\sqrt{kB} \sum_{\substack{s\in[M]\\ sk\geq 2}} \sqrt{s} \geq \frac{1}{128\sqrt{12}} M\sqrt{kh} \geq \frac{1}{512} M\sqrt{kh}.$$

If $M = 1$, then $k \geq 2$, and the single active block gives

$$\frac{1}{32\sqrt{2}}\sqrt{kB} \geq \frac{1}{32\sqrt{12}}\sqrt{kh} \geq \frac{1}{512}\sqrt{kh}.$$

Finally, each class is active in at most one block. Its selected arm moves from the baseline at the beginning of that block and returns to the baseline afterward, so every class undergoes at most two change rounds. Thus every instance in the prior is admissible.

Averaging over the prior yields a fixed admissible instance with at least the displayed expected regret. The same argument applies conditional on any preceding history independent of the permutation and improved-arm identities. □

C.2 Proofs of Our Main Lower Bounds

We now apply the two auxiliary constructions to prove the lower bounds stated in Section 5.

C.2.1 Direct-sum construction for arbitrary class parameters

Theorem C.3 (Restatement of Theorem 5.1)**.** *Fix any collection of class sizes $\{K_m\}_{m=1}^M$ and change budgets $\{L_m\}_{m=1}^M$. If the horizon satisfies $T \geq 2D$, then for any policy, there exists an admissible nonstationary environment such that*

$$\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{256}\sqrt{TD}.$$

Proof. Recall that $D = \sum_{m=1}^M (L_m + 1)K_m$. The construction uses independent learning blocks inside the different classes. Each such block is surrounded by baseline rounds, so it spends exactly two changes and presents the learner with a fresh instance of Lemma C.1.

We distinguish two cases.

Case 1: $D \leq 4K$. In this regime a stationary instance is already sufficiently hard. Apply Lemma C.1 with all K arms as candidates over the full game length T. Since

$$T \geq 2D \geq 2K,$$

the lemma applies and gives a stationary instance with

$$\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{32\sqrt{2}}\sqrt{KT}.$$

Every stationary instance is admissible for every collection of change budgets. Moreover,

$$\frac{1}{32\sqrt{2}}\sqrt{KT} \geq \frac{1}{64\sqrt{2}}\sqrt{TD} \geq \frac{1}{256}\sqrt{TD}.$$

Case 2: $D > 4K$**.** For each class m, create

$$q_m := \left\lfloor \frac{L_m}{2} \right\rfloor$$

learning blocks, each of length

$$b_m := \left\lfloor \frac{TK_m}{D} \right\rfloor.$$

Arrange all these blocks in any fixed order, with one baseline round before the first block, between consecutive blocks, and after the last block. All arms have mean $1/2$ on baseline rounds and on any unused rounds.

Since $T \geq 2D$,

$$\frac{TK_m}{D} \geq 2K_m \geq 2.$$

Hence the floor loses at most a factor of two:

$$b_m \geq \frac{TK_m}{2D} \geq K_m.$$

Thus the conditional block lower bound applies to every class-m block.

We next verify that the construction fits within the horizon. The learning blocks occupy at most

$$\begin{aligned} \sum_m q_m b_m &\leq \frac{T}{D} \sum_m q_m K_m \\ &\leq \frac{T}{2}, \end{aligned}$$

where the last inequality uses $2q_m \leq L_m$.

The number of required baseline separator rounds is at most

$$1 + \sum_m q_m \leq 1 + \frac{D}{2} \leq D \leq \frac{T}{2}.$$

Here we used $K_m \geq 1$ and $D \geq K \geq 2$. Thus all blocks and separators fit within the game length T.

Independently for every class-m block, sample the mean assignment from Lemma C.1 with the K_m arms of that class as candidates. If $K_m \geq 2$, one uniformly chosen arm is elevated above the baseline. If $K_m = 1$, its unique arm receives a positive or negative shift; an outside baseline arm exists because $K \geq 2$.

Every class-m block creates exactly two changes in class m, and no changes in any other class. Therefore class m uses at most

$$2q_m = 2 \left\lfloor \frac{L_m}{2} \right\rfloor \leq L_m$$

changes, so every instance in the prior is admissible.

It remains to lower-bound the total weight of the blocks. For every nonnegative integer L_m,

$$L_m + 1 \leq 2 + 2 \left\lfloor \frac{L_m}{2} \right\rfloor.$$

Multiplying by K_m and summing gives

$$\sum_m q_m K_m \geq \frac{D}{2} - K.$$

Since $D > 4K$,

$$\sum_m q_m K_m \geq \frac{D}{4}.$$

Finally, condition on the complete transcript before each learning block. Its random mean assignment is independent of that transcript, so Lemma C.1 applies conditionally. By the tower property,

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}_T] &\geq \frac{1}{32\sqrt{2}} \sum_m q_m \sqrt{K_m b_m} \\ &\geq \frac{1}{64} \sqrt{\frac{T}{D}} \sum_m q_m K_m \\ &\geq \frac{1}{256} \sqrt{TD}.\end{aligned}$$

Averaging over the product prior yields a fixed admissible mean sequence with at least this expected regret. □

We conclude the quantity $D = \sum_m (L_m + 1)K_m$ is therefore unavoidable. Each class contributes its own relearning cost, and observations collected from other classes cannot remove this cost. Consequently every algorithm suffers $\Omega\Big(\sqrt{TD}\Big)$ dynamic regret in general.

The direct-sum construction does not, however, determine whether the additional dependence on M can be necessary when the number of classes grows. The next lower bound isolates this cross-class search difficulty.

Theorem C.4 (Restatement of Theorem 5.2)**.** *Let K be divisible by M, and suppose all classes are balanced with $K_m = K/M$ with change budgets $L_m = 2$ for all $m \in [M]$. If $T \geq 6MK$, then for any policy, there exists an admissible environment such that*

$$\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{512} \sqrt{MKT}.$$

Proof. Apply Lemma C.2 with

$$n = M, \qquad k = \frac{K}{M}, \qquad h = T.$$

The requirement $nk \geq 2$ is exactly the assumption $K \geq 2$, and the horizon requirement becomes

$$T \geq 6n^2 k = 6MK.$$

The lemma therefore gives

$$\mathbb{E}[\mathrm{Regret}_T] \geq \frac{1}{512} M \sqrt{\frac{K}{M} T} = \frac{1}{512} \sqrt{MKT}.$$

Each class is active in at most one block in the construction and therefore undergoes at most two changes. Hence the instance is admissible when $L_m = 2$ for every $m \in [M]$. □

Note that for this balanced family,

$$K_m = \frac{K}{M}, \qquad L_m = 2,$$

and therefore

$$D = \sum_{m=1}^{M} 3\frac{K}{M} = 3K.$$

Thus the direct-sum difficulty is only of order K, yet every algorithm must incur

$$\Omega\Big(\sqrt{MKT}\Big) = \Omega\Big(\sqrt{MTD}\Big)$$

dynamic regret.

Hence the additional factor $\sqrt{M}$ therefore cannot in general be removed. Even with only two changes per class, the learner must repeatedly determine which class contains the improved arm. The $\sqrt{M}$ dependence in our upper bound is thus an unavoidable class-search cost on this balanced family.

The two lower bounds capture different difficulties: the direct-sum bound measures the cumulative cost of relearning arms within classes, while the class-search bound measures the additional cost of locating the relevant class. The following corollary records both statements on the balanced two-change family.

Corollary C.5 (Combined lower bound)**.** *Suppose $K \geq 2$ and K is divisible by M. If*

$$K_m = \frac{K}{M}, \qquad L_m = 2$$

for every $m \in [M]$, and $T \geq 6MK$, then every algorithm has an admissible instance on which

$$\mathbb{E}[\mathrm{Regret}_T] \geq \max\left\{ \frac{1}{512}\sqrt{MKT}, \frac{1}{256}\sqrt{TD} \right\}.$$

Proof. For these parameters,

$$D = 3K.$$

Moreover,

$$T \geq 6MK \geq 6K = 2D.$$

Hence both Theorems 5.1 and 5.2 apply to the same class sizes and budgets.

Fix an algorithm. Each theorem supplies an admissible hard instance, possibly a different one. Choose the instance supplied by whichever theorem gives the larger of the two displayed lower bounds. □

D Extension to Unknown Budgets

In this section, we remove the assumption that the class change budgets $L_1, \ldots, L_M$ are known. We introduce Adaptive Exponential Backfill for Heterogeneous Arm Classes (`AEBF-H`), which receives only the horizon T and the class partition and achieves the same leading regret rate as `EBF-H`.

The main challenge is that the backfill probability of `EBF-H` depends explicitly on L_m. A natural alternative is to replace this budget by the number of changes detected so far in class m. Doing so introduces a new feedback loop: monitoring affects the observations, these observations affect classification and resets, and these in turn determine future monitoring rates. `AEBF-H` breaks this feedback by separating classification observations from additional monitoring observations and by servicing monitoring requests through an independent monitoring branch.

Apart from these adaptations, the algorithm preserves the phase-based classification mechanism of `EBF-H`: it uses the same `ClassifyArms` procedure, blocking arms, dominance gaps and scales, and the same class–scale priority rule from Section 3.

D.1 The `AEBF-H` Algorithm

We retain all notation from Section 3. In particular,

$$J_{\max} := \lceil \log_2 T \rceil, \qquad \Lambda := \lceil \log(2TK) \rceil.$$

Classification and detection statistics. Unlike `EBF-H`, `AEBF-H` separates the observations used for *arm classification* from those used for *change detection*. For every arm a, it maintains two histories: $\mathcal{O}_a^{\mathrm{cls}} \subseteq \mathcal{O}_a^{\mathrm{det}}$.

The *classification history* $\mathcal{O}_a^{\mathrm{cls}}$ contains only rewards collected on the classification branch. These observations are used by `ClassifyArms` to form certificates, classify arms as good or bad, compute dominance gaps and scales, and determine the class–scale priority rule. Thus, with respect to this history, the classification mechanism is exactly the one from Section 3.

The *detection history* $\mathcal{O}_a^{\mathrm{det}}$ contains all classification observations together with the additional observations collected while servicing monitoring obligations. It is used only for detecting distribution changes and deciding when a class should be reset.

This separation is necessary in the unknown-budget setting. Monitoring obligations are launched randomly at rates that depend on the previously detected changes. If the resulting monitoring

observations were also used for classification, the random launch decisions could change future certificates and hence determine which obligations remain relevant. This would create a feedback loop between launching a monitoring request and the future classification trajectory. On the other hand, ignoring the monitoring observations altogether would defeat their purpose: they are precisely the additional samples used to reveal that a previously bad arm has changed. Using two histories resolves both issues: monitoring observations contribute to change detection without directly affecting classification.

For $H \in \{\mathrm{cls}, \mathrm{det}\}$ and an interval I, let

$$N_a^H(I), \qquad \hat{\mu}_a^H(I), \qquad \mathrm{CB}_a^H(I), \qquad \hat{\mu}_a^{H,(\pm q)}(I)$$

denote the pull count, empirical mean, confidence radius, and empirical confidence bounds of Section 3, computed using only observations from history H. Throughout this appendix, unmarked statistics refer to the classification history, while detection statistics retain the det superscript. In particular, `ClassifyArms` from Algorithm 2 can be applied without modification.

Mechanism 1: Adaptive class-isolated change detection. As in `EBF-H`, every class has its own phase start ϕ_m, and resets are performed independently across classes. The difference is that change detection may now use either observation history.

Let $\widehat{L}_m(t)$ be the number of resets of class m completed through the reset step at the beginning of round t, with $\widehat{L}_m(0) = 0$. At round t, class m is reset if there exist $a \in \mathcal{A}_m$, $H, H' \in \{\mathrm{cls}, \mathrm{det}\}$, and intervals $I_1, I_2 \subseteq [\phi_m, t-1]$ such that

$$\hat{\mu}_a^{H,(-1)}(I_1) > \hat{\mu}_a^{H',(+1)}(I_2). \tag{14}$$

This is exactly the change-detection test of Section 3, except that its two intervals may use either history.

The intuition behind the design is that the two sides of the reset test need not use the same observation history. Indeed, suppose arm a has not changed during the current phase and has constant mean μ. On the concentration event, for either $H \in \{\mathrm{cls}, \mathrm{det}\}$ and every stationary interval I,

$$\hat{\mu}_a^{H,(-1)}(I) \le \mu \le \hat{\mu}_a^{H,(+1)}(I).$$

Consequently, for any two histories H, H' and any two phase intervals I_1, I_2,

$$\hat{\mu}_a^{H,(-1)}(I_1) \le \mu \le \hat{\mu}_a^{H',(+1)}(I_2),$$

so the strict inequality in the reset test cannot hold without a true change.

Allowing cross-history comparisons is essential because the two histories play different roles. Classification observations determine certificates and therefore the future classification trajectory, whereas additional monitoring observations are recorded only in the detection history. Thus, after a bad arm changes, fresh monitoring observations can eventually become inconsistent with its stale classification certificate, e.g.,

$$\hat{\mu}_a^{\mathrm{det},(-1)}(I_{\mathrm{new}}) > \hat{\mu}_a^{\mathrm{cls},(+1)}(I_{\mathrm{old}}),$$

triggering a reset without allowing monitoring samples to directly affect classification. The four choices $(\mathrm{cls}, \mathrm{cls})$, $(\mathrm{cls}, \mathrm{det})$, $(\mathrm{det}, \mathrm{cls})$, and $(\mathrm{det}, \mathrm{det})$ therefore preserve the original detector while allowing fresh monitoring data to detect changes in either direction.

When a reset occurs, we set

$$\phi_m \leftarrow t, \qquad \widehat{L}_m(t) \leftarrow \widehat{L}_m(t) + 1,$$

and discard all outstanding monitoring requests for arms in $\mathcal{A}_m$. Observations need not be deleted, since all subsequent statistics are restricted to the new phase.

As in the known-budget algorithm, the reset rule is conservative: Lemma D.3 shows that, under simultaneous concentration, every reset can be charged to a true change. In particular,

$$\widehat{L}_m(t) \le L_m.$$

Thus, the detected reset count can therefore serve as an online proxy for the unknown budget.

Mechanism 2: Arm classification and dominance gaps. Classification is unchanged from EBF−H. After the reset step, we run $\texttt{ClassifyArms}(t, \phi_1, \ldots, \phi_M)$ from Algorithm 2, using the classification observations only. This produces exactly the same objects as in Section 3: $\mathcal{G}_t$, $\mathcal{B}_t$, blocking arms, certificate intervals, dominance gaps $\Delta_t(a)$, and dominance scales $i_t(a)$. The structural property that blocking arms are good also continues to hold.

Mechanism 3: Budget-free adaptive backfill. The main difference from EBF−H is the monitoring schedule. At scale j, a monitoring request requires $n_j := 2^{2j+7}\Lambda$ observations. Instead of using the unknown budget L_m, class m launches monitoring requests at rate

$$q_{m,j}(t) := \min\left\{1,\ 2^{-j}\sqrt{\frac{\widehat{L}_m(t)+1}{TK_m\Lambda}}\right\}. \tag{15}$$

This rule has the same qualitative behavior as the backfill rule of EBF−H. Finer scales require more samples and are therefore launched less frequently, while classes in which more changes have already been detected are monitored more aggressively. The crucial difference is that the rate is now completely data-driven.

A successful launch creates a *monitoring obligation* (a, j, s, r), where a is the arm, j is the scale, s is the launch round, and r is the number of required monitoring observations remaining. A scale-j launch at round s creates (a, j, s, n_j). Let $\mathcal{Z}$ denote the set of live obligations.

Unlike the single counters of EBF−H, different launches are kept separately rather than overwriting one another. This is important in the adaptive analysis because each launch has its own independent random seed and can be charged separately.

For every bad arm $a \in \mathcal{A}_m \cap \mathcal{B}_t$ and scale $j \in \{0, \ldots, \min\{i_t(a), J_{\max}\}\}$, an obligation is launched with probability $q_{m,j}(t)$. All launch decisions use mutually independent uniform random seeds, indexed by arm, scale, and round. Conditional on the history before round t, the launch decisions at that round are independent with their stated history-dependent probabilities.

Mechanism 4: Independent monitoring. Once a monitoring obligation is created, the algorithm must collect the additional samples it requests. The difficulty is that these samples cannot be scheduled using the obligation set itself: whether an obligation exists depends on the random coin used to launch it, so choosing future actions based directly on the outstanding obligations would make those launch coins influence the subsequent sampling process and, potentially, the classification trajectory.

To avoid this dependence, AEBF−H separates ordinary classification from monitoring. At every round, it independently chooses between a *classification branch*, which continues the usual EBF−H classification process, and a *monitoring branch*, which provides opportunities to service outstanding obligations. On the monitoring branch, the candidate arm is sampled independently of which obligations are currently active. Thus an obligation determines only whether a monitoring opportunity is used, not which arm receives that opportunity.

An obligation (a, j, s, r) is *serviceable* at round t if either a is currently good, or a is bad and its scale is still relevant:

$$\mathcal{Z}_t(a) := \left\{(a, j, s, r) \in \mathcal{Z} : a \in \mathcal{G}_t \text{ or } \left(a \in \mathcal{B}_t, j \le i_t(a)\right)\right\}.$$

An obligation that is temporarily not serviceable remains live.

At every round, the algorithm first constructs a *classification proposal*. It draws a class uniformly among those containing a good arm and a scale uniformly from $\{0, \ldots, J_{\max}\}$, and then applies the same recency rule as EBF−H, restricted to good arms and classification observations.

The algorithm then chooses uniformly between two branches:

1. *Classification branch.* Play the classification proposal and add its reward to both the classification and detection histories.
2. *Monitoring branch.* Draw a class uniformly from $[M]$, and then draw an arm uniformly from that class, independently of the outstanding obligations. If the selected arm has a

serviceable obligation, play it, add its reward only to the detection history, and decrease the remaining requirement of every serviceable obligation of that arm by one. Otherwise, play the classification proposal but do not add this reward to either history.

The last convention is deliberate. Whether the monitoring candidate has a serviceable obligation depends on previous launch coins. If a fallback reward were added to the classification history, these launch coins could change the future classification trajectory. Discarding the fallback observation prevents this feedback.

Consequently, every arm $a \in \mathcal{A}_m$ receives an independent monitoring opportunity with probability

$$p_m^{\mathrm{M}} = \frac{1}{2MK_m}$$

at every round. This predictable service rate is the key property used in the adaptive analysis.

Run-time complexity. As for `EBF-H`, the exact implementation is dominated by the all-interval classification step. At round t, `ClassifyArms` requires $O\big(K^2t^4\big)$ time. The two-history change-detection test considers only four ordered history pairs and requires $O\big(Kt^4\big)$ time, which is lower order.

At most one obligation is launched for each arm, scale, and round, so there are at most $O(KT(1+\log T))$ live obligations. Storing them in per-arm lists makes their total scanning cost $O\big(KT^2(1+\log T)\big)$, which is again lower order. Hence the total running time is $O\big(K^2T^5\big)$. Using prefix statistics for both observation histories, the total memory complexity is $O(KT(1+\log T))$, hence the algorithm is efficient.

The next theorem states the expected dynamic regret of Algorithm 3.

Theorem D.1 (Expected Dynamic Regret for Unknown Class Change Budgets)**.** *For every admissible nonstationary environment, the expected dynamic regret of `AEBF-H` satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le 16384\Lambda^2\sqrt{MTD} + 3 = \widetilde{O}\Big(\sqrt{MTD}\Big)\,.$$

We emphasize that the class change budgets are not supplied to the algorithm and appear only through D in the guarantee.

Thus prior knowledge of the class change budgets can be removed without changing the leading regret rate of `EBF-H`. The proof of Theorem D.1 is given next.

D.2 Regret Analysis

Our goal is to prove Theorem D.1. The proof follows the same structure as the known-budget analysis in Section 4 and Appendix B.1. We first establish simultaneous concentration and show that every algorithmic reset can be charged to a true change. This controls the number of maximal static intervals (MSIs) and, crucially, implies that the number of detected changes never exceeds the unknown budget.

We then prove the confidence-sum bounds needed in the regret decomposition. As in the known-budget analysis, every bad arm has a good blocking arm, so both the played and optimal arms can be represented by good arms.

Finally, we analyze the three terms in equation 4 separately. Terms (i) and (ii) are close to their known-budget counterparts. The main new argument is for Term (iii), delayed change detection. Between two detected resets of a class, the adaptive launch probability is fixed. We use this fact to bound the waiting time until a useful monitoring obligation is launched and then show that sufficiently servicing such an obligation forces a reset. This replaces the fixed-budget waiting-time argument used for `EBF-H`.

Throughout the analysis, $\mathcal{F}_{t-1}$ denotes the information revealed before the fresh random draws at round t.

Algorithm 3: `AEBF-H`: Adaptive Exponential Backfill for Heterogeneous Arm Classes

Input: Horizon T and classes $\mathcal{A}_1, \dots, \mathcal{A}_M$; no change budgets.

Set $J_{\max} \leftarrow \lceil \log_2 T \rceil$, $\Lambda \leftarrow \lceil \log(2TK) \rceil$, and $n_j \leftarrow 2^{2j+7}\Lambda$ for $j \in \{0, \dots, J_{\max}\}$

Initialize $\phi_m \leftarrow 1$ and $\widehat{L}_m(0) \leftarrow 0$ for every class m; initialize both observation histories and $\mathcal{Z}$ to $\emptyset$

for $t = 1, \dots, T$ **do**

  `// 1. Adaptive class-isolated change detection`

  foreach $m \in [M]$ **do**

    $\widehat{L}_m(t) \leftarrow \widehat{L}_m(t-1)$

    if $\exists a \in \mathcal{A}_m$, $H, H' \in \{\mathrm{cls}, \mathrm{det}\}$, $I_1, I_2 \subseteq [\phi_m, t-1]$ *such that*

$$\hat{\mu}_a^{H,(-1)}(I_1) > \hat{\mu}_a^{H',(+1)}(I_2)$$

    then

      $\phi_m \leftarrow t$

      $\widehat{L}_m(t) \leftarrow \widehat{L}_m(t) + 1$

      Delete all obligations whose arm belongs to $\mathcal{A}_m$

    end

  end

  `// 2. Arm classification`

  Run `ClassifyArms`$(t, \phi_1, \dots, \phi_M)$ from Algorithm 2 using classification observations only

  `// 3. Adaptive backfill launches`

  foreach $a \in \mathcal{B}_t$ **do**

    $m \leftarrow \mathrm{class}(a)$

    foreach $j = 0, \dots, \min\{i_t(a), J_{\max}\}$ **do**

      With probability $q_{m,j}(t)$, add (a, j, t, n_j) to $\mathcal{Z}$

    end

  end

  `// 4. Classification proposal`

  Draw m_t uniformly from $\{m : \mathcal{A}_m \cap \mathcal{G}_t \neq \emptyset\}$ and j_t uniformly from $\{0, \dots, J_{\max}\}$

  Set

$$a_t^{\mathrm{cls}} \leftarrow \operatorname*{arg\,min}_{a \in \mathcal{A}_{m_t} \cap \mathcal{G}_t} \tau_a^{j_t},$$

  using classification observations in the current phase

  `// 5. Classification or monitoring`

  Draw χ_t uniformly from $\{\mathrm{cls}, \mathrm{mon}\}$

  if $\chi_t = \mathrm{cls}$ **then**

    Play $a_t \leftarrow a_t^{\mathrm{cls}}$ and observe $R_{a_t,t}$

    Add t to both histories of a_t

  else

    Draw m_t^{M} uniformly from $[M]$, and then a_t^{M} uniformly from $\mathcal{A}_{m_t^{\mathrm{M}}}$

    if $\mathcal{Z}_t(a_t^{\mathrm{M}}) \neq \emptyset$ **then**

      Play $a_t \leftarrow a_t^{\mathrm{M}}$ and observe $R_{a_t,t}$

      Add t only to the detection history of a_t

      foreach *serviceable obligation* (a_t, j, s, r) **do**

        Decrease r by one and delete the obligation if $r = 0$

      end

    else

      Play $a_t \leftarrow a_t^{\mathrm{cls}}$ and observe its reward, but do not add t to either history

    end

  end

end

D.2.1 Concentration, Reset Counting, and MSIs

We first extend the concentration and reset-counting arguments of the known-budget analysis to the two histories maintained by `AEBF-H`.

Lemma D.2 (Two-history concentration)**.** *Let $\mathcal{C}'$ be the event that the following holds simultaneously for both $H \in \{\mathrm{cls}, \mathrm{det}\}$, every arm a, and every nonempty interval $I = [s, v]$ contained in one*

stationary segment of that arm:

$$\left|\hat{\mu}_a^H(I) - \mu_a(s)\right| \le \mathrm{CB}_a^H(I).$$

Then

$$\mathbb{P}(\mathcal{C}') \ge 1 - T^{-1}.$$

Proof intuition. The observation times in both histories are adaptive, but whether a reward is recorded is determined before that reward is revealed. The recorded centered rewards therefore form a bounded martingale-difference sequence, so concentration remains valid at these adaptive observation times.

Proof. Fix $a \in \mathcal{A}$, $H \in \{\mathrm{cls}, \mathrm{det}\}$, and $s \in [T]$. For $u \in [T]$, the event $u \in \mathcal{O}_a^H$ means that round u records a reward of arm a in history H.

Let $\mathcal{F}_u^{\mathrm{pre}}$ additionally contain all internal random draws at round u that determine the played arm and whether its reward is recorded, but not the reward itself. Then

$$\mathcal{F}_{u-1} \subseteq \mathcal{F}_u^{\mathrm{pre}} \subseteq \mathcal{F}_{u+1-1},$$

and $\mathbf{1}\{u \in \mathcal{O}_a^H\}$ is $\mathcal{F}_u^{\mathrm{pre}}$-measurable.

Because the reward is sampled only after these decisions, on a stationary segment of arm a with mean μ,

$$\mathbb{E}\big[\mathbf{1}\{u \in \mathcal{O}_a^H\}\big(R_{a,u} - \mu\big) \,\big|\, \mathcal{F}_u^{\mathrm{pre}}\big] = 0.$$

Thus the recorded centered rewards form a bounded martingale-difference sequence.

Let v be the last round in the stationary segment of arm a containing s. For $n \in [T]$, let u_n be the round of the n-th reward recorded in history H during $[s, v]$, with $u_n = \infty$ if there are fewer than n such rewards. Let $\mu := \mu_a(s)$. For every fixed $\lambda \in \mathbb{R}$, the conditional form of Hoeffding's lemma (Lattimore & Szepesvári, 2020, Section 5.4, Note 4) gives the nonnegative supermartingale

$$\exp\left(\lambda \sum_{u=s}^{t} \mathbf{1}\{u \in \mathcal{O}_a^H\}(R_{a,u} - \mu) - \frac{\lambda^2}{8} \sum_{u=s}^{t} \mathbf{1}\{u \in \mathcal{O}_a^H\}\right), \qquad s \le t \le v.$$

Stop it at $\min\{u_n, v\}$, using bounded optional stopping (Lattimore & Szepesvári, 2020, Theorem 3.8). Optimizing λ for each of the two tails gives

$$\mathbb{P}\left(u_n \le v, \quad \left|\hat{\mu}_a^H([s, u_n]) - \mu\right| > \sqrt{\frac{8\log(2TK)}{n}}\right) \le 2\exp\big(-16\log(2TK)\big).$$

The empirical statistic in this event is evaluated only when u_n is finite. In particular, the bound is on the joint event of observing n rewards and violating concentration; it does not condition on the random observation count.

Now take any interval $I = [s, t]$ contained in that stationary segment. If I contains $n \ge 1$ rewards of arm a recorded in history H, and $u_n \le t$ is the round of the last such reward, then

$$N_a^H(I) = N_a^H([s, u_n]) = n,$$

and

$$\hat{\mu}_a^H(I) = \hat{\mu}_a^H([s, u_n]).$$

The two intervals also have the same constant mean. Hence every nonzero-count interval statistic is determined by its arm, history, starting round, and final tagged-observation count.

A union bound over at most $2KT^2$ such choices has failure probability at most

$$4KT^2 \exp\big(-16\log(2TK)\big) \le T^{-1}.$$

When an interval contains no reward recorded in history H, its empirical mean is zero, its true mean lies in $[0, 1]$, and its confidence radius is at least one. This proves simultaneous concentration. □

We next show that concentration prevents spurious resets.

Lemma D.3 (Reset counting)**.** *On $\mathcal{C}'$, if (a, H, H', I_1, I_2) witnesses a class-m reset at round t through equation 14, then arm a has a mean change strictly after the phase start immediately preceding that reset and before t. Consequently, for every class m and round t,*

$$\widehat{L}_m(t) \le \left|\{s \in \mathcal{L}_m : s < t\}\right| \le |\mathcal{L}_m| \le L_m.$$

Proof intuition. Once both histories concentrate, a reset cannot occur unless a genuine change has occurred since the beginning of the current phase: otherwise all confidence intervals for the same arm contain the same constant mean. Moreover, after a reset the phase start moves forward, so the same true change cannot be used to justify a later reset.

Proof. Fix $m \in [M]$ and consider a reset of class m at round t. Let ϕ_m denote its previous phase start. By the reset rule, there are $a \in \mathcal{A}_m$, $H, H' \in \{\mathrm{cls}, \mathrm{det}\}$, and intervals

$$I_1, I_2 \subseteq [\phi_m, t-1]$$

satisfying equation 14.

Suppose, toward a contradiction, that arm a has no mean change in rounds

$$\phi_m + 1, \ldots, t-1.$$

Its mean is then $\mu_a(\phi_m)$ throughout both witness intervals.

The two witness intervals each contain at least one observation in their respective histories. With zero observations, the tagged lower bound is negative and the tagged upper bound exceeds one, so a zero-observation interval cannot occur on either side of the strict reset inequality.

On $\mathcal{C}'$,

$$\hat{\mu}_a^{H,(-1)}(I_1) \le \mu_a(\phi_m) \le \hat{\mu}_a^{H',(+1)}(I_2),$$

contradicting equation 14. Hence every reset can be charged to a class change strictly after the preceding phase start.

After the reset, the new phase starts at t, so a later reset must be charged to a later change. Therefore, for every round u,

$$\widehat{L}_m(u) \le \big|\{v \in \mathcal{L}_m : v < u\}\big| \le |\mathcal{L}_m| \le L_m.$$

□

The reset-counting bound immediately controls the number of MSIs.

Lemma D.4 (Number of MSIs)**.** *On $\mathcal{C}'$, the number of class-m MSIs is at most*

$$|\mathcal{L}_m| + \widehat{L}_m(T) + 1 \le 2L_m + 1.$$

Proof. The first class-m MSI starts at round 1, and every subsequent MSI starts at either a true class-m change or an algorithmic reset of class m. Hence the number of class-m MSIs is at most

$$|\mathcal{L}_m| + \widehat{L}_m(T) + 1.$$

By Lemma D.3,

$$\widehat{L}_m(T) \le L_m,$$

while by definition

$$|\mathcal{L}_m| \le L_m.$$

Therefore

$$|\mathcal{L}_m| + \widehat{L}_m(T) + 1 \le 2L_m + 1.$$

□

The stopped process. To avoid conditioning fresh random draws on the future event $\mathcal{C}'$, let τ_{A} be the first round in $[T]$ at whose start one of the concentration inequalities in Lemma D.2 fails for either history on an arm-stationary interval of preceding rounds. Set $\tau_{\mathrm{A}} = T + 1$ if no such failure occurs, and define $\mathcal{T}_{\mathrm{A}} := \{t \in [T] : t < \tau_{\mathrm{A}}\}$. We call the execution on $\mathcal{T}_{\mathrm{A}}$ the *stopped process*. This stopping rule is used only in the analysis.

For every $t \in \mathcal{T}_{\mathrm{A}}$, all classification- and detection-history statistics formed from observations available at the start of round t satisfy the concentration inequalities of Lemma D.2. Moreover, whether $t \in \mathcal{T}_{\mathrm{A}}$ is determined before the fresh random draws and reward at round t. Consequently, all algorithmic random variables retain their stated conditional distributions on the stopped process.

Applying the reset-counting argument of Lemma D.3 to the concentrated observations available before each stopped round gives

$$\widehat{L}_m(t) \le L_m, \qquad m \in [M], \quad t \in \mathcal{T}_{\mathrm{A}}.$$

Finally, the omitted suffix $[T] \setminus \mathcal{T}_{\mathrm{A}}$ is nonempty only if $\mathcal{C}'$ fails. By Lemma D.2,

$$\mathbb{P}\big((\mathcal{C}')^c\big) \le T^{-1}.$$

Since the regret over the omitted suffix is at most T, its expected contribution is at most one.

Convention for the remainder of the proof. From this point onward, all regret bounds are proved for the stopped process. Thus, whenever $t \in \mathcal{T}_{\mathrm{A}}$, every classification- or detection-history statistic formed from observations available at the start of round t satisfies the concentration bounds of Lemma D.2. We therefore invoke concentration freely on stopped rounds without conditioning on the global event $\mathcal{C}'$. This distinction is useful because membership in $\mathcal{T}_{\mathrm{A}}$ is determined before the fresh random draws at round t. Hence the launch, branch, class, and arm-selection coins retain their stated conditional distributions throughout the stopped-process analysis.

D.2.2 Blocking Arms Are Good

The classification mechanism is unchanged from `EBF-H`, except that it uses only the classification history. The blocking-arm property therefore continues to hold.

Proposition D.5 (Adaptive blocking arms are good)**.** *For every round t and arm $a \in \mathcal{B}_t$, $\bar{a}_t(a) \in \mathcal{G}_t$. Consequently, $\mathcal{G}_t \neq \emptyset$.*

Proof. Fix a round t and an arm $a \in \mathcal{B}_t$, and write $b := \bar{a}_t(a)$. Suppose, toward a contradiction, that $b \in \mathcal{B}_t$. Let the selected certificate for a be $\big(b,\, I_t(a),\, \overline{I}_t(a)\big)$, where $I_t(a)$ is the observation interval for a, and $\overline{I}_t(a)$ is the observation interval for b. Hence

$$\hat{\mu}_b^{(-3)}(\overline{I}_t(a)) > \hat{\mu}_a^{(+3)}(I_t(a)). \tag{16}$$

Since b is also bad, let $c := \bar{a}_t(b)$. Let its selected certificate be $\big(c,\, I_t(b),\, \overline{I}_t(b)\big)$, and therefore

$$\hat{\mu}_c^{(-3)}(\overline{I}_t(b)) > \hat{\mu}_b^{(+3)}(I_t(b)). \tag{17}$$

Step 1: Construct a second certificate for a. We claim that $\big(c,\, I_t(a),\, \overline{I}_t(b)\big)$ is also a valid certificate for a.

Because b is bad after the reset step at round t, its class was not reset at that round. Indeed, immediately after a reset the new phase contains no classification observations, so no three-confidence certificate for b can exist.

Thus the change-detection test did not fire for the class of b. Applying its $(\mathrm{cls}, \mathrm{cls})$ specialization to the two classification intervals of b, $\overline{I}_t(a)$ and $I_t(b)$, gives

$$\hat{\mu}_b^{(-1)}(\overline{I}_t(a)) \le \hat{\mu}_b^{(+1)}(I_t(b)). \tag{18}$$

Equivalently,

$$\hat{\mu}_b(I_t(b)) \ge \hat{\mu}_b(\overline{I}_t(a)) - \mathrm{CB}_b(\overline{I}_t(a)) - \mathrm{CB}_b(I_t(b)).$$

Combining this inequality with equation 16 and equation 17 yields

$$\begin{aligned}
\hat{\mu}_c^{(-3)}(\overline{I}_t(b)) &> \hat{\mu}_b^{(+3)}(I_t(b)) \\
&\ge \hat{\mu}_b(\overline{I}_t(a)) - \mathrm{CB}_b(\overline{I}_t(a)) + 2\mathrm{CB}_b(I_t(b)) \\
&\ge \hat{\mu}_b^{(-1)}(\overline{I}_t(a)) \\
&= \hat{\mu}_b^{(-3)}(\overline{I}_t(a)) + 2\mathrm{CB}_b(\overline{I}_t(a)) \\
&> \hat{\mu}_a^{(+3)}(I_t(a)).
\end{aligned}$$

Hence c, together with $I_t(a)$ and $\overline{I}_t(b)$, satisfies the three-confidence certificate inequality for a.

It remains to check that $c \neq a$. If $c = a$, the preceding strict inequality implies

$$\hat{\mu}_a^{(-3)}(\overline{I}_t(b)) > \hat{\mu}_a^{(+3)}(I_t(a)),$$

and therefore

$$\hat{\mu}_a^{(-1)}(\overline{I}_t(b)) > \hat{\mu}_a^{(+1)}(I_t(a)).$$

Both intervals belong to the current phase of a, so this would have triggered the change-detection test before classification. Thus $c \neq a$, and the displayed tuple is a valid candidate certificate for a.

Step 2: Compare the certificate separations. The original selected certificate for a and the candidate constructed above use the same interval $I_t(a)$ for the blocked arm a. Therefore, to compare their one-confidence separations, it suffices to compare the lower confidence bounds of their blocking arms.

The inequalities from Step 1 imply

$$\hat{\mu}_c^{(-3)}(\overline{I}_t(b)) > \hat{\mu}_b^{(-1)}(\overline{I}_t(a)),$$

and hence

$$\hat{\mu}_c^{(-1)}(\overline{I}_t(b)) > \hat{\mu}_b^{(-1)}(\overline{I}_t(a)).$$

Consequently,

$$\hat{\mu}_c^{(-1)}(\overline{I}_t(b)) - \hat{\mu}_a^{(+1)}(I_t(a)) > \hat{\mu}_b^{(-1)}(\overline{I}_t(a)) - \hat{\mu}_a^{(+1)}(I_t(a)).$$

Thus the candidate certificate with blocking arm c has strictly larger one-confidence separation than the selected certificate with blocking arm $b = \bar{a}_t(a)$. This contradicts the maximizing definition of $\bar{a}_t(a)$.

Therefore $\bar{a}_t(a) \in \mathcal{G}_t$. Finally, if some arm is bad, its blocking arm is good; if no arm is bad, every arm is good. Hence $\mathcal{G}_t \neq \emptyset$. □

D.2.3 Confidence-Sum Bounds

We use the current-MSI statistics of Section B.1 for both histories. Thus, if τ is the first round of the current MSI of arm a's class, then $\mathrm{CB}_a(t) = \mathrm{CB}_a([\tau, t-1])$, and analogously for the detection-history statistics.

We first bound the simpler confidence sum over serviced monitoring observations. Define

$$\mathcal{T}_{\mathrm{M}} := \left\{ t \in \mathcal{T}_{\mathrm{A}} : \chi_t = \mathrm{mon},\ \mathcal{Z}_t(a_t^{\mathrm{M}}) \neq \emptyset \right\}.$$

On every $t \in \mathcal{T}_{\mathrm{M}}$, the algorithm plays $a_t = a_t^{\mathrm{M}}$ and adds its reward to the detection history.

Lemma D.6 (Detection-history confidence sum)**.** *For the stopped process,*

$$\sum_{t \in \mathcal{T}_{\mathrm{M}}} \mathrm{CB}_{a_t}^{\mathrm{det}}(t) \leq 8\sqrt{\Lambda T D}. \tag{19}$$

Proof intuition. Within an MSI, every serviced monitoring observation increases the detection-history count of the played arm. We can therefore sum the confidence radii separately for every arm–MSI pair and then combine the pairs using Cauchy–Schwarz, exactly as for the played-arm confidence sum in the known-budget analysis.

Proof. By Lemma D.4, class m has at most $2L_m + 1$ MSIs intersecting $\mathcal{T}_{\mathrm{A}}$. Hence the number of relevant arm–MSI pairs is at most

$$\sum_{m=1}^{M} (2L_m + 1)K_m \leq 2D.$$

Fix one such arm–MSI pair. If it contains no rounds from $\mathcal{T}_M$, its confidence contribution is zero. Otherwise, let $n \geq 1$ be the number of rounds from $\mathcal{T}_M$ in this pair.

Before its r-th serviced monitoring observation, the detection history contains at least $r-1$ observations of that arm in the MSI. Its total confidence contribution is therefore at most

$$\sqrt{8\Lambda}\left(1+\sum_{r=1}^{n-1} r^{-1/2}\right) \le 2\sqrt{8\Lambda n}.$$

Thus the confidence contribution is at most $2\sqrt{8\Lambda n}$ for every $n \ge 0$, using the zero contribution when $n = 0$.

Let $n_{a,I}$ denote the number of serviced monitoring rounds assigned to arm–MSI pair (a, I). Summing over all pairs and applying Cauchy–Schwarz gives

$$\sum_{t\in\mathcal{T}_{\mathrm{M}}} \mathrm{CB}^{\mathrm{det}}_{a_t}(t) \le 2\sqrt{8\Lambda} \sum_{(a,I)} \sqrt{n_{a,I}} \le 2\sqrt{8\Lambda}\sqrt{|\{(a,I)\}| \sum_{(a,I)} n_{a,I}}.$$

There are at most $2D$ pairs and at most T serviced monitoring rounds. Therefore

$$\sum_{t\in\mathcal{T}_{\mathrm{M}}} \mathrm{CB}^{\mathrm{det}}_{a_t}(t) \le 8\sqrt{\Lambda T D}.$$

□

The preceding argument applies only to observations that are actually collected. We also need to control the confidence radius of a good arm even when that arm is not selected. For this we use the randomized class–scale selection rule, as in the eligible-arm confidence bound of the known-budget analysis.

However, we note that the corresponding bound in the known-budget analysis is stated over all eligible arms, because `EBF-H` uses a single selection rule in which bad arms with active monitoring counters remain eligible. The adaptive algorithm separates these two roles. The classification branch of `AEBF-H` selects only from $\mathcal{G}_t$, while bad arms are sampled separately by the monitoring branch and are controlled through their detection-history confidence radii. Moreover, all representatives appearing in the regret decomposition are good by Proposition D.5. It is therefore enough here to control the largest classification confidence radius among good arms.

Lemma D.7 (Good-arm confidence sum)**.** *For the stopped process,*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{A}}} \max_{a\in\mathcal{G}_t} \mathrm{CB}_a(t)\right] \le 128\sqrt{M(J_{\max}+1)\Lambda T D}. \tag{20}$$

Proof intuition. A good arm with a large confidence radius need not itself be selected. At each round, we therefore compare the widest good arm with the arm that the classification rule would choose from the same class at an appropriate scale. This comparison arm has a confidence radius within a factor two of the widest arm and is selected with probability at least $1/[2M(J_{\max}+1)]$. Each time it is selected, its classification count increases. Counting how often each observation count can occur then bounds the cumulative confidence radius.

Proof. Write $\ell := \log(2TK)$. For each stopped round $t \in \mathcal{T}_A$, choose

$$x_t \in \arg\max_{a\in\mathcal{G}_t} \mathrm{CB}_a(t), \qquad k_t := N_{x_t}(t), \qquad h_t := \lceil \log_2(k_t+1) \rceil,$$

using the fixed tie-breaking rule. Define

$$b_t \in \arg\min_{a\in\mathcal{G}_t\cap\mathcal{A}_{\mathrm{class}(x_t)}} \tau_a^{h_t},$$

with the same tie-breaking rule as the classification proposal. The arm x_t has the largest classification confidence radius, whereas b_t is the arm the classification rule would select from its class at scale h_t. The symbol j_t continues to denote the algorithm's fresh, uniformly sampled scale; it is not redefined in this proof.

The quantities x_t, k_t, h_t, b_t are determined before the fresh round-t draws, and $0 \le h_t \le J_{\max}$. Let u_t be the common current-MSI start of x_t and b_t. Since $k_t < 2^{h_t} \le 2(k_t+1)$, the 2^{h_t}-th most recent

classification observation of x_t in its current phase precedes u_t, if it exists. The recency rule therefore gives

$$\tau_{b_t}^{h_t} \le \tau_{x_t}^{h_t} < u_t, \qquad N_{b_t}(t) < 2^{h_t} \le 2(k_t+1).$$

If $k_t = 0$, then $h_t = 0$ and $N_{b_t}(t) = 0$; if $k_t \ge 1$, then $2(k_t+1) \le 4k_t$. Thus

$$\max\{N_{b_t}(t), 1\} \le 4\max\{k_t, 1\}, \qquad \max_{a\in\mathcal{G}_t} \mathrm{CB}_a(t) = \mathrm{CB}_{x_t}(t) \le 2\,\mathrm{CB}_{b_t}(t).$$

Conditional on $\mathcal{F}_{t-1}$, the event

$$A_t := \{\chi_t = \mathrm{cls},\ m_t = \mathrm{class}(x_t),\ j_t = h_t\}$$

has probability at least

$$p_{\mathrm{cls}} := \frac{1}{2M(J_{\max}+1)}.$$

On A_t, the algorithm plays b_t and records a classification observation of that arm. For $n \in \{0, \dots, T-1\}$, set

$$V_n := \sum_{t\in\mathcal{T}_A} \mathbf{1}\{N_{b_t}(t) = n\}, \qquad S_n := \sum_{t\in\mathcal{T}_A} \mathbf{1}\{N_{b_t}(t) = n\}\mathbf{1}\{A_t\}.$$

Every round counted by S_n supplies the $(n+1)$-st classification observation of a particular arm in a particular MSI. No arm–MSI pair can do this twice. The stopped reset-counting argument bounds the number of such pairs by $2D$, so $S_n \le 2D$ pathwise. Since membership in $\mathcal{T}_A$ and the condition $N_{b_t}(t) = n$ are determined before the fresh draws,

$$p_{\mathrm{cls}}\mathbb{E}[V_n] \le \mathbb{E}[S_n] \le 2D.$$

The exact confidence-sum identity is

$$\sum_{t\in\mathcal{T}_A} \mathrm{CB}_{b_t}(t) = \sqrt{8\ell}\sum_{n=0}^{T-1} \frac{V_n}{\sqrt{\max\{n,1\}}}.$$

Set $\beta := Tp_{\mathrm{cls}}/(2D)$. If $\beta < 1$, then

$$\sum_{t\in\mathcal{T}_A} \max_{a\in\mathcal{G}_t} \mathrm{CB}_a(t) \le T\sqrt{8\ell} \le 4\sqrt{\frac{\ell TD}{p_{\mathrm{cls}}}}.$$

For $\beta \ge 1$, the inequalities $\mathbb{E}[V_n] \le 2D/p_{\mathrm{cls}}$ and $\sum_{0\le n<\beta} 1/\sqrt{\max\{n,1\}} \le 2\sqrt{\beta}$ bound the small-count contribution by $8\sqrt{\ell TD/p_{\mathrm{cls}}}$. On the remaining rounds, each radius is at most $\sqrt{8\ell/\beta}$, so their contribution is at most $4\sqrt{\ell TD/p_{\mathrm{cls}}}$. Therefore

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_A} \max_{a\in\mathcal{G}_t} \mathrm{CB}_a(t)\right] \le 24\sqrt{\frac{\ell TD}{p_{\mathrm{cls}}}} = 24\sqrt{2M(J_{\max}+1)\ell TD} \le 128\sqrt{M(J_{\max}+1)\Lambda TD},$$

where the final step uses $\ell \le \Lambda$ and $24\sqrt{2} \le 128$. This also covers $\beta < 1$ and proves the stated bound. □

D.2.4 Regret Decomposition

We begin by defining the two good arms that will serve as reference points in the regret decomposition. For the played arm, define

$$\hat{a}_t := \begin{cases} a_t, & a_t \in \mathcal{G}_t, \\ \bar{a}_t(a_t), & a_t \in \mathcal{B}_t. \end{cases}$$

Thus, if the played arm is good, it represents itself; if it is bad, we replace it by the blocking arm selected by its classification certificate.

Similarly, for the optimal arm define

$$\hat{a}_t^* := \begin{cases} a_t^*, & a_t^* \in \mathcal{G}_t, \\ \bar{a}_t(a_t^*), & a_t^* \in \mathcal{B}_t. \end{cases}$$

Hence the optimal arm also represents itself whenever it is good, while a bad optimal arm is represented by the arm that currently certifies it as bad.

By Proposition D.5, the blocking arm of every bad arm is good. Consequently, $\hat{a}_t \in \mathcal{G}_t$ and $\hat{a}_t^* \in \mathcal{G}_t$ at every round. This is the purpose of the representatives: they replace any bad arm appearing in the regret comparison by a good arm whose current confidence radius can be controlled. Importantly, when an arm is bad, its representative is chosen using its historical classification certificate; we do not assume that the representative's current mean is larger than the mean of the arm it represents.

Adding and subtracting the means of these two representatives gives

$$\mu_{a_t^*}(t) - \mu_{a_t}(t) = \underbrace{\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)}_{\text{Term (i)}} + \underbrace{\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)}_{\text{Term (ii)}} + \underbrace{\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)}_{\text{Term (iii)}}.$$

The three terms isolate the same statistical effects as in the known-budget analysis. Term (i) is nonzero only when the algorithm plays a bad arm and measures the cost of monitoring that arm relative to its good representative. Term (ii) compares two good arms and is therefore controlled by their current confidence radii. Term (iii) is nonzero only when the optimal arm is bad and measures the loss caused by delayed detection of a stale optimal-arm certificate.

Define the two representatives as in Section 4.2, using the adaptive good and bad sets. By Proposition D.5, both representatives are good. Hence the algebraic decomposition equation 4 applies unchanged:

$$\mu_{a_t^*}(t) - \mu_{a_t}(t) = \underbrace{\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)}_{\text{Term (i)}} + \underbrace{\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)}_{\text{Term (ii)}} + \underbrace{\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)}_{\text{Term (iii)}}.$$

As in the known-budget analysis, Term (i) is the cost of playing a bad arm for monitoring, Term (ii) compares the two good representatives, and Term (iii) captures the delay in recognizing an optimal arm whose certificate has become stale. We bound the three terms separately.

Term (i): monitoring bad arms. Term (i) vanishes whenever the played arm is good. If a bad arm is played, it must be servicing a monitoring obligation. We charge the resulting dominance-gap contribution to that obligation.

Lemma D.8 (Bounding Term (i))**.** *For the stopped process,*

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} (\mu_{\hat{a}_t}(t) - \mu_{a_t}(t))\right] \le 4096\Lambda^2\sqrt{MTD}.$$

Proof intuition. When a bad arm is played, its selected certificate compares it with a good blocking arm. The no-reset condition lets us compare the historical certificate with the current classification and detection statistics. This bounds the regret by two confidence terms and the arm's dominance gap. The confidence terms are controlled by Lemmas D.6 and D.7. For the dominance gap, a scale-j obligation is charged at most n_j times, and the expected number of launched obligations is controlled by the adaptive launch probability.

Proof. Fix $t \in \mathcal{T}_{\mathrm{A}}$. The term is zero if $a_t \in \mathcal{G}_t$. Otherwise, let $a := a_t$ and $b := \hat{a}_t = \bar{a}_t(a)$.

The selected three-confidence certificate for a implies

$$\Delta_t(a) \ge 2\mathrm{CB}_b(\overline{I}_t(a)) + 2\mathrm{CB}_a(I_t(a)).$$

Neither certificate interval can be empty: an empty-history three-confidence lower bound is negative, whereas an empty-history three-confidence upper bound exceeds one.

Since these certificates exist after the round-t reset step, the corresponding classes were not reset at round t. Hence the reset test did not invoke on the relevant intervals. For the blocking arm b, its $(\mathrm{cls}, \mathrm{cls})$ specialization gives

$$\hat{\mu}_b^{(-1)}(t) \le \hat{\mu}_b^{(+1)}(\overline{I}_t(a)),$$

while for a, its $(\mathrm{cls}, \mathrm{det})$ specialization gives

$$\hat{\mu}_a^{(-1)}(I_t(a)) \le \hat{\mu}_a^{\mathrm{det},(+1)}(t).$$

If either current-MSI interval is empty, the corresponding comparison follows directly from the empty-history convention.

Since $t \in \mathcal{T}_{\mathrm{A}}$, all statistics available at the start of round t satisfy the concentration inequalities of Lemma D.2. Combining concentration with the preceding no-reset comparisons gives

$$\mu_b(t) - \left[\hat{\mu}_b^{(-1)}(\overline{I}_t(a)) + \Delta_t(a)\right] \le 2\mathrm{CB}_b(t),$$

and

$$\left[\hat{\mu}_a^{(+1)}(I_t(a)) - \Delta_t(a)\right] - \mu_a(t) \le 2\mathrm{CB}_a^{\mathrm{det}}(t).$$

By definition,

$$\hat{\mu}_b^{(-1)}(\overline{I}_t(a)) - \hat{\mu}_a^{(+1)}(I_t(a)) = \Delta_t(a).$$

Adding the three contributions yields

$$\mu_{\hat{a}_t}(t) - \mu_{a_t}(t) \le 2\mathrm{CB}_{\hat{a}_t}(t) + 3\Delta_t(a_t) + 2\mathrm{CB}_{a_t}^{\mathrm{det}}(t). \tag{21}$$

We next bound the dominance-gap contribution. Every bad arm played by `AEBF-H` has a serviceable obligation. Assign $3\Delta_t(a_t)$ to the first such obligation under a fixed deterministic ordering, and let j be its scale. Since $j \le i_t(a_t)$, we have $\Delta_t(a_t) \le 2^{1-j}$. Thus the obligation receives a charge of at most $6 \cdot 2^{-j}$ at this round.

Each service decreases the obligation's remaining count by one, so a scale-j obligation receives at most n_j charges before completion. Its total charge is therefore at most $6 \cdot 2^{-j} n_j$.

Fix $m \in [M]$ and $j \in \{0, \dots, J_{\max}\}$. For $a \in \mathcal{A}_m$, let $B_{a,j,t}$ denote the indicator that a scale-j obligation is launched for a at round t. Since the classification and launch probability are determined before the launch seed is examined,

$$\mathbb{E}[B_{a,j,t} \mid \mathcal{F}_{t-1}] = \mathbf{1}\{a \in \mathcal{B}_t,\, j \le \min\{i_t(a), J_{\max}\}\} q_{m,j}(t) \le q_{m,j}(t).$$

On every stopped round, $\widehat{L}_m(t) + 1 \le L_m + 1$. Moreover, $n_j = 2^{2j+7}\Lambda \le 129\Lambda 2^{2j}$. Membership in $\mathcal{T}_{\mathrm{A}}$ is $\mathcal{F}_{t-1}$-measurable, so the tower property gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} \sum_{a \in \mathcal{A}_m} 6 \cdot 2^{-j} n_j B_{a,j,t}\right] \\
&\quad \le \mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} \sum_{a \in \mathcal{A}_m} 6 \cdot 2^{-j} n_j q_{m,j}(t)\right] \\
&\quad \le 774\sqrt{\Lambda T (L_m + 1) K_m}.
\end{aligned}$$

Summing over the $J_{\max} + 1$ scales and applying Cauchy–Schwarz over classes bounds all dominance-gap charges by

$$774(J_{\max} + 1)\sqrt{\Lambda M T D}.$$

Finally, the first confidence term in equation 21 is controlled by equation 20, since the representative is good, and the detection-history term is controlled by equation 19. Using $J_{\max} + 1 \le 3\Lambda$, $\Lambda \ge 1$, and $M \ge 1$ proves the lemma. □

Term (ii): comparing good representatives. Both representatives are good, so this term is controlled entirely by their current classification confidence radii.

Lemma D.9 (Bounding Term (ii))**.** *For the stopped process,*

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} \left(\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)\right)\right] \le 4096\Lambda^2\sqrt{MTD}.$$

Proof intuition. Since both representatives are good, their current-MSI intervals cannot form a three-confidence certificate. Concentration converts this empirical comparison into a bound by their confidence radii, whose cumulative sum is controlled by Lemma D.7.

Proof. Fix $t \in \mathcal{T}_{\mathrm{A}}$, and let $a' := \hat{a}_t^*$ and $a := \hat{a}_t$. Both arms belong to $\mathcal{G}_t$.

If $a = a'$, the contribution is zero. Otherwise, the classification rule does not declare a bad using a' and their two current-MSI intervals. Hence

$$\hat{\mu}_{a'}(t) - \hat{\mu}_a(t) \le 3\mathrm{CB}_{a'}(t) + 3\mathrm{CB}_a(t).$$

If either current-MSI history is empty, the same inequality follows directly from the empty-history convention.

Since $t \in \mathcal{T}_{\mathrm{A}}$, concentration holds for both current-MSI statistics. Therefore

$$\mu_{a'}(t) - \mu_a(t) \le 4\mathrm{CB}_{a'}(t) + 4\mathrm{CB}_a(t).$$

Both radii are at most $\max_{u \in \mathcal{G}_t} \mathrm{CB}_u(t)$. Applying equation 20 gives

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} \left(\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)\right)\right] \le 1024\sqrt{M(J_{\max}+1)\Lambda TD}.$$

Using $J_{\max} + 1 \le 3\Lambda$ and $\Lambda \ge 1$ proves the lemma. □

Term (iii): delayed detection. Term (iii) is nonzero only when the optimal arm is classified as bad. In this case its classification certificate may contain observations from before a recent change, so the historical certificate need not reflect the arm's current mean.

Lemma D.10 (Bounding Term (iii))**.** *For the stopped process,*

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{A}}} \left(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\right)\right] \le 8192\Lambda^2\sqrt{MTD} + 2.$$

Proof intuition. Suppose the optimal arm is currently classified as bad. Its representative is its good blocking arm. We compare the current mean of the optimal arm with the historical upper confidence bound that caused it to be classified as bad. The difference between these two quantities is the *certificate residual*.

Small residuals can be absorbed directly into confidence and class-dependent threshold terms. For a large residual at scale j, the algorithm launches scale-j monitoring obligations for the optimal arm. While class m has detected $e-1$ changes, the launch probability is fixed and gives a geometric waiting time of order $e^{-1/2}$. Once a useful obligation is launched, sufficiently many fresh monitoring samples make the detection interval incompatible with the stale classification certificate and force a reset.

We first introduce the notation used in this argument and then prove the three ingredients: completion of an obligation forces detection, a useful obligation is launched quickly enough, and the resulting large residuals have the desired cumulative bound.

Residuals and hard rounds. Let $\mathcal{U} := \{t \in \mathcal{T}_{\mathrm{A}} : a_t^* \in \mathcal{B}_t\}$. Thus $\mathcal{U}$ contains exactly the stopped rounds on which Term (iii) may be nonzero.

Fix $t \in \mathcal{U}$, and write $a^\star := a_t^*$. Since $a^\star$ is bad, its selected classification certificate contains an interval $I_t(a^\star)$ for $a^\star$. Define its *certificate residual* by

$$r_t := \mu_{a_t^*}(t) - \widehat{\mu}_{a_t^*}^{(+1)}\big(I_t(a_t^*)\big), \qquad t \in \mathcal{U}. \tag{22}$$

Hence a positive residual means that the current mean of the optimal arm lies above the historical upper confidence bound used in its certificate.

We also introduce a class-dependent threshold for separating residuals that can be summed directly from those that require monitoring. For each class $m \in [M]$, let

$$T_m := |\{t \in [T] : a_t^* \in \mathcal{A}_m\}|$$

be the number of rounds in the full horizon whose optimal arm belongs to class m. The fixed tie-breaking rule makes the optimal arm unique, so $\sum_{m=1}^{M} T_m = T$.

Define

$$\zeta_m := \begin{cases} \Lambda\sqrt{\dfrac{MK_m(L_m+1)}{T_m}}, & T_m > 0, \\ 0, & T_m = 0. \end{cases}$$

These thresholds are used only in the analysis; the algorithm does not know the change budgets.

We call a stopped round a *large-residual round* if the optimal arm is bad and its certificate residual exceeds both its dominance gap and its class threshold:

$$\mathcal{U}_{\text{large}} := \left\{ t \in \mathcal{U} : \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) > \max\left\{ \Delta_t(a_t^*), \zeta_{\text{class}(a_t^*)} \right\} \right\}.$$

To match such residuals with the monitoring scales, for $m \in [M]$ and $j \in \{1, \dots, J_{\max}\}$, define

$$\mathcal{U}_{m,j} := \left\{ t \in \mathcal{U}_{\text{large}} : a_t^* \in \mathcal{A}_m,\ 2^{-j} < \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le 2^{1-j} \right\}.$$

Thus $\mathcal{U}_{m,j}$ consists of the large-residual rounds for class m whose residual is at resolution 2^{-j}.

Residuals in $\mathcal{U}_{\text{large}}$ that are at most $2^{-J_{\max}}$ are not assigned to any of these scales. Their total contribution is at most $T2^{-J_{\max}} \le 1$.

First reduction of Term (iii). Fix $t \in \mathcal{U}$, write $a^\star := a_t^*$, and let $b := \hat{a}_t^* = \bar{a}_t(a^\star)$. By Proposition D.5, $b \in \mathcal{G}_t$.

Let $\overline{I}_t(a^\star)$ be the interval for b in the selected certificate of $a^\star$. By the definition of the dominance gap,

$$\Delta_t(a^\star) = \hat{\mu}_b^{(-1)}(\overline{I}_t(a^\star)) - \hat{\mu}_{a^\star}^{(+1)}(I_t(a^\star)).$$

Using equation 22, we therefore have the exact identity

$$\begin{aligned} \mu_{a^\star}(t) - \mu_b(t) = \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a^\star) \\ + \hat{\mu}_b^{(-1)}(\overline{I}_t(a^\star)) - \mu_b(t). \end{aligned}$$

The class of b was not reset at the start of round t, because the certificate interval displayed above is still available. Since the reset test did not fire, comparing this certificate interval with the current classification history of b gives

$$\hat{\mu}_b^{(-1)}(\overline{I}_t(a^\star)) \le \hat{\mu}_b^{(+1)}(t).$$

Since $t \in \mathcal{T}_{\text{A}}$, concentration holds for the current MSI statistic of b. Hence

$$\hat{\mu}_b^{(-1)}(\overline{I}_t(a^\star)) - \mu_b(t) \le 2\text{CB}_b(t).$$

Consequently,

$$\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \le \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a_t^*) + 2\text{CB}_{\hat{a}_t^*}(t). \tag{23}$$

If $\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le \Delta_t(a_t^*)$, the first two terms on the right are nonpositive. If the residual exceeds the dominance gap but is at most $\zeta_{\text{class}(a_t^*)}$, their difference is at most that threshold. All remaining rounds belong to $\mathcal{U}_{\text{large}}$. Summing equation 23 therefore gives

$$\begin{aligned} &\sum_{t \in \mathcal{T}_{\text{A}}} \left[\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \right] \\ &\le 2 \sum_{t \in \mathcal{T}_{\text{A}}} \text{CB}_{\hat{a}_t^*}(t) + \sum_{m=1}^{M} T_m \zeta_m + \sum_{t \in \mathcal{U}_{\text{large}}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)). \end{aligned} \tag{24}$$

The first sum is controlled by Lemma D.7, because every representative is good. For the threshold contribution, the definition above and Cauchy–Schwarz give

$$\begin{aligned}\sum_{m=1}^{M} T_m\zeta_m &= \Lambda\sqrt{M}\sum_{\substack{m\in[M]:\\ T_m>0}}\sqrt{T_m K_m(L_m+1)}\\ &\le \Lambda\sqrt{M}\sqrt{\sum_{m=1}^{M} T_m}\sqrt{\sum_{m=1}^{M} K_m(L_m+1)}\\ &= \Lambda\sqrt{MTD}. \end{aligned} \tag{25}$$

It remains to control the large-residual sum.

Monitoring opportunities and completion. Recall that a scale-j monitoring obligation requires

$$n_j = 2^{2j+7}\Lambda$$

services before it is completed.

Fix a class m and an arm $a \in \mathcal{A}_m$. We call round t a *monitoring slot for* a if $\chi_t = \mathrm{mon}$ and $a_t^{\mathrm{M}} = a$. Conditional on the information available before these two fresh random draws, the probability of such a slot is

$$p_m^{\mathrm{M}} := \frac{1}{2MK_m}.$$

If a has a serviceable obligation, a monitoring slot supplies one new detection-history observation of a and advances every serviceable obligation of that arm.

For $m \in [M]$ and $j \in \{1, \dots, J_{\max}\}$, define

$$W_{m,j} := 8MK_m n_j. \tag{26}$$

Using the definition of n_j,

$$W_{m,j} \le 1033MK_m\Lambda 2^{2j}.$$

Moreover,

$$W_{m,j}p_m^{\mathrm{M}} = 4n_j.$$

Thus a block of $W_{m,j}$ hard opportunities contains four times the required number of monitoring slots in expectation.

Lemma D.11 (Completion within a stationary segment)**.** *Fix $m \in [M]$, $j \in \{1, \dots, J_{\max}\}$, and $s \in \mathcal{U}_{m,j}$. Let $a := a_s^*$. Suppose a scale-j obligation for a is launched at round s and receives n_j services before the next true class-m change. Then, at the start of the round immediately following completion, class m is reset or that round does not belong to $\mathcal{T}_{\mathrm{A}}$.*

Proof intuition. At the launch round s, the residual is larger than 2^{-j}, so the current mean of a lies noticeably above its stale classification upper bound. After n_j fresh detection observations, the detection confidence radius is at most 2^{-j-2}. If the class has not changed again, the fresh detection lower bound must therefore exceed the stale classification upper bound, which triggers the cross-history reset test.

Proof. Fix m, j, s as in the lemma and let $a := a_s^*$. If the stopped process ends before the obligation completes, the second alternative in the statement already holds.

Otherwise, let u be the round in which the obligation receives its n_j-th service. By assumption, no true class-m change occurs after s and through u, so $[s, u]$ lies inside one true class-m stationary segment.

A class-m reset deletes every live obligation of class m. Therefore, because the obligation launched at s survives until its completion at u, no class-m reset occurs between s and u. Hence both the detection interval $[s, u]$ and the original classification interval $I_s(a)$ belong to the same class-m phase at the start of round $u + 1$.

The interval $[s, u]$ contains at least n_j detection observations of a. Consequently,

$$\mathrm{CB}_a^{\mathrm{det}}([s,u]) \le \sqrt{\frac{8\Lambda}{2^{2j+7}\Lambda}} = 2^{-j-2}.$$

If the observation at round u creates the first failure of the concentration inequalities, then $u+1 = \tau_{\mathrm{A}}$, so the following round is outside the stopped process.

Otherwise, concentration holds for the detection statistic on $[s, u]$. Because $s \in \mathcal{U}_{m,j}$, $\mu_{a_s^*}(s) - \hat{\mu}_{a_s^*}^{(+1)}(I_s(a_s^*)) > 2^{-j}$. By the definition of the residual,

$$\mu_a(s) > \hat{\mu}_a^{(+1)}(I_s(a)) + 2^{-j}.$$

Since the mean of a is constant on $[s, u]$, concentration gives

$$\begin{aligned}\hat{\mu}_a^{\mathrm{det},(-1)}([s,u]) &\ge \mu_a(s) - 2\mathrm{CB}_a^{\mathrm{det}}([s,u]) \\ &> \hat{\mu}_a^{(+1)}(I_s(a)) + 2^{-j} - 2^{-j-1} \\ &> \hat{\mu}_a^{(+1)}(I_s(a)).\end{aligned}$$

If $u = T$, there is no following round in the stopped process. Otherwise, the last inequality is precisely a $(\mathrm{det}, \mathrm{cls})$ witness for the reset condition equation 14. Thus class m resets at the start of round $u + 1$. □

Detected episodes and their launch probabilities. We next control how long the process waits before such a useful obligation is launched.

For a class $m \in [M]$ and an integer $e \ge 1$, define its e-th *detected episode* by

$$\mathcal{P}_{m,e} := \left\{ t \in \mathcal{T}_{\mathrm{A}} : \widehat{L}_m(t) + 1 = e \right\}.$$

If nonempty, this is the consecutive set of stopped rounds after the $(e-1)$-st class-m reset and before the next class-m reset, or before the stopped process ends. Throughout this episode, $\widehat{L}_m(t) = e - 1$.

Fix a scale j. Since the detected-change count is constant throughout the episode, the class-m, scale-j launch probability is also constant. Denote this probability by

$$q_{m,j,e} := 2^{-j}\sqrt{\frac{e}{TK_m\Lambda}}.$$

For every $t \in \mathcal{P}_{m,e}$,

$$q_{m,j}(t) = q_{m,j,e}.$$

To see that the cap in equation 15 is inactive, note that on a nonempty episode $e = \widehat{L}_m(t)+1 \le t \le T$, while $K_m \ge 1$, $\Lambda \ge 1$, and $2^{-j} \le 1$.

An auxiliary hard-round sequence. A subtlety is that a successful launch may later cause a reset, which changes future classifications and hence changes which future rounds are hard. We therefore define an auxiliary trajectory whose hard rounds can be fixed without looking at the class-m, scale-j launch seeds.

Fix $m \in [M]$, $j \in \{1, \dots, J_{\max}\}$, and $e \ge 1$, and suppose first that $\mathcal{P}_{m,e} \ne \emptyset$. Let $s_e := \min \mathcal{P}_{m,e}$ be the first round of the episode.

Starting from the state immediately after the reset step at round s_e, construct an auxiliary execution using the same round-indexed rewards and the same algorithmic random seeds as the actual process, except that every class-m reset update is suppressed. Thus, when the class-m reset test fires in the auxiliary execution, we keep ϕ_m fixed, do not increment $\widehat{L}_m(t)$, and do not delete class-m obligations. Resets of all other classes are processed normally.

The auxiliary execution continues through round T without imposing any concentration stopping rule. Neither membership in $\mathcal{T}_A$ nor an auxiliary counterpart of that condition is used to define its frozen hard sequence. Actual stopping enters only through the definition of reached rounds below.

A suppressed reset may make the auxiliary good set empty. To keep the auxiliary recursion well defined in that case, define its classification proposal to be the first arm under the fixed arm ordering. This convention is needed only after the auxiliary and actual executions have diverged; it is never invoked while the actual detected episode is still active.

We now define three auxiliary objects.

Frozen hard sequence. Use the superscript aux for classifications and certificate statistics computed along the auxiliary execution. Whenever $a_t^* \in \mathcal{B}_t^{\mathrm{aux}}$, define

$$r_t^{\mathrm{aux}} := \mu_{a_t^*}(t) - \widehat{\mu}_{a_t^*}^{\mathrm{aux},(+1)}\big(I_t^{\mathrm{aux}}(a_t^*)\big).$$

Let $\mathcal{H}_{m,j,e}^{\mathrm{aux}}$ be the ordered sequence of all rounds $t \in \{s_e, \dots, T\}$ satisfying

$$\begin{aligned} a_t^* &\in \mathcal{A}_m \cap \mathcal{B}_t^{\mathrm{aux}}, \\ r_t^{\mathrm{aux}} &> \max\{\Delta_t^{\mathrm{aux}}(a_t^*), \zeta_m\}, \\ 2^{-j} &< r_t^{\mathrm{aux}} \le 2^{1-j}. \end{aligned}$$

In particular, this definition does not require $t \in \mathcal{T}_A$.

We say that the actual execution *reaches* an element $t \in \mathcal{H}_{m,j,e}^{\mathrm{aux}}$ if $t \in \mathcal{P}_{m,e}$. While the detected episode is active, no class-m reset has yet separated the actual and auxiliary trajectories, so their classification histories agree. Hence the reached elements are exactly

$$\mathcal{U}_{m,j} \cap \mathcal{P}_{m,e}.$$

Core opportunity sequence. Partition $\mathcal{H}_{m,j,e}^{\mathrm{aux}}$ according to the true class-m stationary segments. In each part, remove its final $W_{m,j}$ elements, or remove the whole part if it contains fewer than this many elements. Concatenate the remaining elements in their original order and call the resulting sequence $\mathcal{H}_{m,j,e}^{\circ}$.

The purpose of this trimming is that every element of $\mathcal{H}_{m,j,e}^{\circ}$ has at least $W_{m,j}$ later frozen hard rounds in the same true stationary segment. Thus a useful obligation launched at a core round has enough future hard opportunities to be completed before the next true class change.

Core waiting count. Recall that $U_{a,j,t} \sim \mathrm{Unif}[0,1]$ is the latent uniform seed used to decide whether a scale-j obligation for arm a is launched at round t.

Let $G_{m,j,e}$ be the number of reached elements of $\mathcal{H}_{m,j,e}^{\circ}$, through and including the first reached core round t satisfying

$$U_{a_t^*,j,t} \le q_{m,j,e}.$$

If no such successful launch occurs before the detected episode ends, $G_{m,j,e}$ is the total number of reached core elements.

Finally, if $\mathcal{P}_{m,e} = \emptyset$, we use the convention

$$\mathcal{H}_{m,j,e}^{\mathrm{aux}} = \mathcal{H}_{m,j,e}^{\circ} = \emptyset, \qquad G_{m,j,e} = 0.$$

Lemma D.12 (Waiting time on hard class–scale rounds)**.** *Fix $m \in [M]$, $j \in \{1, \dots, J_{\max}\}$, and $e \ge 1$. Then*

$$\mathbb{E}[G_{m,j,e}] \le q_{m,j,e}^{-1}.$$

Moreover, suppose a reached core round $s \in \mathcal{H}_{m,j,e}^{\circ}$ successfully launches a scale-j obligation for $a := a_s^*$. If the actual process reaches the next $W_{m,j}$ frozen hard rounds in the same true class-m stationary segment and at least n_j of them are monitoring slots for a, then class m resets no later than the start of the following round, or the stopped process has ended by then.

Proof intuition. We expose all randomness except the class-m, scale-j launch seeds. The auxiliary hard sequence is then already fixed. Consequently, each reached core round reveals a fresh independent launch coin with the same success probability. A reset can only stop the sequence early, so it cannot increase the waiting time.

Proof. Fix m, j, e. The result is immediate when $\mathcal{P}_{m,e} = \emptyset$, so assume the episode is nonempty.

Let $s \in [T]$ be a possible deterministic episode start. The event

$$\{\min \mathcal{P}_{m,e} = s\}$$

is $\mathcal{F}_{s-1}$-measurable: whether s belongs to the stopped process, whether a class-m reset occurs at its start, and the detected-change count after that reset step are all determined before the fresh random draws of round s.

Recall the latent-seed representation of the launch rule: before the game, sample mutually independent $U_{a,h,u} \sim \mathrm{Unif}[0,1]$ for every arm a, scale h, and round u. Whenever the algorithm considers a scale-h obligation for a at round u, it launches the obligation exactly when

$$U_{a,h,u} \le q_{\mathrm{class}(a),h}(u).$$

On the event $\{\min \mathcal{P}_{m,e} = s\}$, expose $\mathcal{F}_{s-1}$, the entire round-indexed reward array $\{R_{a,u}\}_{a,u}$, and all future algorithmic random seeds except

$$\{U_{a,j,u} : a \in \mathcal{A}_m,\ u \in [s,T]\}.$$

The index set of these omitted seeds is deterministic once m, j, s are fixed and does not depend on the future hard sequence.

To make the independence explicit, construct the auxiliary execution described above while additionally suppressing every class-m, scale-j obligation whose launch seed belongs to the omitted collection. We claim that this additional suppression does not change the auxiliary classification trajectory.

Indeed, argue inductively over rounds. On a classification branch, identical phase starts and classification histories produce the same good and bad sets, the same class and scale choices, the same classification proposal, and the same recorded classification reward. On a monitoring branch, no classification reward is recorded, including on a fallback play, and the monitoring candidate is sampled independently of the live obligations.

Changing the omitted launch seeds can therefore affect only the class-m, scale-j obligations and the class-m detection history. It does not alter any classification history. Resets of classes other than m are therefore unchanged, while class-m resets are suppressed by construction and hence cannot alter its phase start. Consequently, all classifications, certificates, and residuals in the auxiliary trajectory are unchanged.

This invariance applies to the entire unstopped auxiliary classification trajectory. Conditional on the exposed episode-start history, reward array, and other seeds, both $\mathcal{H}^{\mathrm{aux}}_{m,j,e}$ and $\mathcal{H}^{\circ}_{m,j,e}$ are therefore fixed independently of the omitted launch seeds. Those seeds may change the actual detection history, concentration stopping time, and next class-m reset, but these affect only which frozen rounds are reached, not the frozen sequences themselves.

Until the actual detected episode ends, its classification trajectory agrees with the auxiliary one. At every reached frozen hard round t,

$$\Delta_t(a_t^*) < r_t \le 2^{1-j},$$

so $j \le i_t(a_t^*)$ and the algorithm examines the scale-j launch seed of the optimal arm. Throughout the episode, its success probability is $q := q_{m,j,e}$.

Conditional on the exposed randomness, the core rounds and their associated optimal arms are fixed. If $G_{m,j,e} > r$, then the first r core rounds exist, are reached, and all their launch seeds exceed q. These are distinct independent uniform seeds. Thus the conditional probability of this event is at most $(1-q)^r$, regardless of any earlier termination caused by noncore launches or concentration failure. Averaging over the additionally exposed randomness gives, for every integer $r \ge 0$,

$$\mathbb{P}(\{\min \mathcal{P}_{m,e} = s\} \cap \{G_{m,j,e} > r\} \,|\, \mathcal{F}_{s-1}) \le \mathbf{1}\{\min \mathcal{P}_{m,e} = s\}(1-q)^r.$$

A launch at a noncore round may cause an earlier reset and shorten the episode, but it cannot make the event on the left more likely.

Summing the geometric tail probabilities gives

$$\mathbb{E}[G_{m,j,e}\mathbf{1}\{\min \mathcal{P}_{m,e} = s\}] \le \frac{\mathbb{P}(\min \mathcal{P}_{m,e} = s)}{q}.$$

The possible episode-start events are disjoint, so summing over $s \in [T]$ yields

$$\mathbb{E}[G_{m,j,e}] \le q^{-1}.$$

For the second claim, suppose a reached core round s successfully launches a scale-j obligation for $a = a_s^*$. By the definition of the core sequence, there are at least $W_{m,j}$ later frozen hard rounds in the same true stationary segment.

Unless a class-m reset or τ_{A} occurs first, these rounds are also reached by the actual process. On every such hard round, the global optimal arm belongs to class m. Since the means of all class-m arms and the deterministic tie-breaking rule are unchanged inside the stationary segment, this optimal arm is the same arm a. Moreover, the hard-scale condition gives $j \le i_t(a)$, so the obligation remains serviceable.

Hence n_j monitoring slots for a complete the obligation. Lemma D.11 then implies that class m resets by the following round or the stopped process has already ended. □

Bounding the large residuals. We can now combine the waiting-time and completion arguments.

Lemma D.13 (Regret on hard class–scale rounds)**.** *For the stopped process,*

$$\mathbb{E}\left[\sum_{m=1}^{M}\sum_{j=1}^{J_{\max}}\sum_{t\in\mathcal{U}_{m,j}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))\right] \le 4096\Lambda^2\sqrt{MTD} + 1.$$

Proof intuition. Fix a class m, scale j, and detected episode e. Before the first useful launch, each hard round contributes at most 2^{1-j}, while the expected number of such rounds is at most the inverse launch probability. The factor 2^{-j} in the launch probability cancels the residual scale, leaving a contribution proportional to $e^{-1/2}$.

After a successful launch, a block of $W_{m,j}$ hard rounds contains enough monitoring slots with very high probability. The obligation then completes and forces a reset. We separately charge the few hard rounds removed near stationary-segment boundaries.

Proof. Fix $m \in [M]$ and $j \in \{1, \dots, J_{\max}\}$.

Every $t \in \mathcal{U}_{m,j}$ satisfies

$$\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le 2^{1-j}$$

and

$$\Delta_t(a_t^*) < \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le 2^{1-j}.$$

Hence $j \le i_t(a_t^*)$. In particular, a scale-j obligation for the optimal arm is serviceable on every such round.

Within a true class-m stationary segment, all class-m means are fixed. Since every hard (m, j) round has its global optimum in class m, the deterministic tie-breaking rule selects the same class-m optimal arm on all such rounds in that segment.

Enough monitoring slots after a launch. Fix a deterministic candidate launch round $s \in [T]$, and let $a := a_s^*$. If $a \notin \mathcal{A}_m$, then s cannot be a class-m core launch and there is nothing to consider.

Otherwise, follow the actual execution after s until the first of: the next class-m reset, the next true class-m change, or τ_{A}. Among the rounds before this boundary, mark the successive rounds belonging to $\mathcal{U}_{m,j}$, stopping after $W_{m,j}$ marked rounds have been reached.

Whether a round is marked is determined by $\mathcal{F}_{t-1}$. Conditional on the information available before the fresh branch and monitor-candidate draws at a marked round, the probability that the round is a monitoring slot for a is $p_m := p_m^{\mathrm{M}}$.

Let $X \in \{0,1\}$ be the indicator of such a monitoring slot. For every $\theta > 0$,

$$\mathbb{E}[e^{-\theta X} \mid \text{past}] = 1 - p_m + p_m e^{-\theta} \le \exp\bigl(-p_m(1-e^{-\theta})\bigr).$$

Therefore the exponential of minus θ times the number of monitoring slots, plus $p_m(1-e^{-\theta})$ times the number of reached marked rounds, is a nonnegative supermartingale. Hold both counts fixed after the stopping boundary.

Let $W := W_{m,j}$, $n := n_j$, and $p := p_m^{\mathrm{M}}$. Recall that $Wp = 4n$.

Fix the candidate launch round s and launch arm $a := a_s^*$. Starting after round s, follow the actual execution until the first of the following events: a class-m reset, a true class-m change, τ_{A}, or the W-th subsequent round in $\mathcal{U}_{m,j}$. Let τ denote this stopping round.

For $u > s$, let N_u be the number of marked hard (m,j) rounds reached by time u, capped at W, and let S_u be the number among these reached marked rounds that are monitoring slots for a. Both processes are kept constant after τ.

Consider a reached marked round u. Whether u is marked is $\mathcal{F}_{u-1}$-measurable. Conditional on $\mathcal{F}_{u-1}$, the fresh branch and monitoring-candidate draws are independent of the past, and

$$\mathbb{P}\bigl(\chi_u = \mathrm{mon},\, a_u^{\mathrm{M}} = a \,\big|\, \mathcal{F}_{u-1}\bigr) = p.$$

Thus, if X_u denotes the indicator that the marked round u is a monitoring slot for a, then for every $\theta > 0$,

$$\mathbb{E}\bigl[e^{-\theta X_u} \,\big|\, \mathcal{F}_{u-1}\bigr] = 1 - p + pe^{-\theta} \le \exp\bigl(-p(1-e^{-\theta})\bigr).$$

Consequently,

$$M_u := \exp\bigl(-\theta S_u + p(1-e^{-\theta})N_u\bigr)$$

is a nonnegative supermartingale with $M_s = 1$. Since $\tau \le T$, bounded optional stopping gives $\mathbb{E}[M_\tau] \le 1$.

Let $\mathcal{F}_{s,m,j}$ be the event that all W marked hard rounds are reached before the stopping boundary, but fewer than n of them are monitoring slots for a. On this event, $N_\tau = W$ and $S_\tau < n$. Therefore

$$M_\tau = \exp\bigl(-\theta S_\tau + p(1-e^{-\theta})W\bigr) \ge \exp\bigl(-\theta n + p(1-e^{-\theta})W\bigr).$$

Hence

$$1 \ge \mathbb{E}[M_\tau] \ge \mathbb{P}(\mathcal{F}_{s,m,j}) \exp\bigl(-\theta n + p(1-e^{-\theta})W\bigr),$$

and therefore

$$\mathbb{P}(\mathcal{F}_{s,m,j}) \le \exp\bigl(\theta n - p(1-e^{-\theta})W\bigr).$$

Choosing $\theta = \log 4$, so that $1 - e^{-\theta} = 3/4$, and using $Wp = 4n$, gives

$$\begin{aligned}\mathbb{P}(\mathcal{F}_{s,m,j}) &\le \exp\left(n\log 4 - \frac{3}{4}Wp\right)\\ &= \exp(-(3-\log 4)n)\\ &\le e^{-n} = e^{-n_j},\end{aligned}$$

where the last inequality uses $3 - \log 4 > 1$.

Importantly, this bounds the *joint* event that all W hard opportunities are reached and that fewer than n monitoring slots occur among them. We do not condition on the future event that these W opportunities are reached; their adaptive reachability is handled directly by the stopped counting processes N_u and S_u.

Define $\mathcal{E}_{\mathrm{srv}}$ to be the event that the preceding service condition succeeds for every reached core launch, over all classes, scales, and candidate launch rounds. Since $n_j \ge 128\Lambda$, a union bound gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{srv}}^c) \le MJ_{\max}Te^{-128\Lambda} \le T^{-2}.$$

When $T = 1$, there are no scales $j \in \{1, \dots, J_{\max}\}$, so the event holds trivially.

Waiting before the first core launch. Fix a nonempty detected episode $\mathcal{P}_{m,e}$. By Lemma D.12,

$$\mathbb{E}[G_{m,j,e}] \le q_{m,j,e}^{-1}.$$

Every reached core round in $\mathcal{U}_{m,j}$ contributes at most 2^{1-j} residual. Therefore the expected residual accumulated through the first successful core launch is at most

$$\frac{2^{1-j}}{q_{m,j,e}} = \frac{2^{1-j}}{2^{-j}\sqrt{e/(TK_m\Lambda)}} = 2\sqrt{\frac{TK_m\Lambda}{e}}.$$

By Lemma D.3, the detected-change count never exceeds L_m on the stopped process. Thus there are at most $L_m + 1$ nonempty detected episodes. Summing the preceding bound over them gives

$$2\sqrt{TK_m\Lambda}\sum_{e=1}^{L_m+1} e^{-1/2} \le 4\sqrt{TK_m\Lambda(L_m+1)},$$

where we used $\sum_{e=1}^{L+1} e^{-1/2} \le 2\sqrt{L+1}$.

Post-launch and boundary rounds. On $\mathcal{E}_{\mathrm{srv}}$, once the first successful core launch occurs in an episode, at most $W_{m,j}$ further hard rounds can be reached in the same stationary segment before a class-m reset or τ_{A}. Indeed, if all these opportunities are reached, the service-success event supplies at least n_j monitoring slots, and Lemmas D.11 and D.12 forces the reset.

Moreover, on $\mathcal{E}_{\mathrm{srv}}$, an episode with a successful core launch cannot remain active into the next true class-m stationary segment. Otherwise it would reach all of the $W_{m,j}$ later frozen hard rounds guaranteed by the core definition in the launch segment. The service condition and completion lemma would then end the episode, or the stopped process, no later than the first round of the next segment. Thus the $W_{m,j}$ post-launch charge is paid at most once per detected episode.

There are also hard rounds deliberately removed when constructing the core sequence. Each intersection of a detected episode with a true class-m stationary segment contributes at most $W_{m,j}$ removed rounds.

Before τ_{A}, class m has at most L_m resets by Lemma D.3 and at most L_m true changes. The union of the reset and change boundaries therefore partitions the stopped horizon into at most $2L_m + 1$ detected-episode/stationary-segment intersections.

Consequently, the union of the post-launch hard rounds and the hard rounds removed from the core contains at most

$$3(L_m+1)W_{m,j}$$

rounds. It also contains at most T_m rounds, because every hard (m,j) round has its optimal arm in class m.

Let $\mathcal{R}_{m,j}$ denote the hard (m,j) rounds that are not already charged to the core waiting contribution. By the preceding argument,

$$|\mathcal{R}_{m,j}| \le 3(L_m+1)W_{m,j},$$

while also

$$|\mathcal{R}_{m,j}| \le T_m,$$

because every $t \in \mathcal{R}_{m,j}$ has $a_t^* \in \mathcal{A}_m$.

Moreover, every $t \in \mathcal{R}_{m,j}$ belongs to $\mathcal{U}_{m,j}$, and therefore $\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le 2^{1-j}$. Hence

$$\begin{aligned}\sum_{t\in\mathcal{R}_{m,j}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) &\le 2^{1-j}|\mathcal{R}_{m,j}| \\ &\le 2^{1-j}\min\{T_m,\, 3(L_m+1)W_{m,j}\} \\ &\le 6\cdot 2^{-j}\min\{T_m,\,(L_m+1)W_{m,j}\}.\end{aligned}$$

The last inequality uses $\min\{x, 3y\} \le 3\min\{x,y\}$.

Using $\min\{x,y\} \le \sqrt{xy}$ and equation 26, this is at most

$$256\sqrt{MK_m\Lambda(L_m+1)T_m},$$

where we used $6\sqrt{1033} \le 256$.

Summing over classes and scales. Write $\Gamma := \sqrt{MTD}$ and define

$$H := \sum_{m=1}^{M} \sum_{j=1}^{J_{\max}} \sum_{t \in \mathcal{U}_{m,j}} r_t,$$
$$R_{\text{wait}} := \sum_{m=1}^{M} \sum_{j=1}^{J_{\max}} \sum_{e=1}^{L_m+1} 2^{1-j} G_{m,j,e}.$$

Here R_{wait} is a nonnegative random upper bound on the core waiting contribution, with $G_{m,j,e} = 0$ for empty episodes. The waiting-time bound of Lemma D.12 and Cauchy–Schwarz give the unconditional estimate

$$\mathbb{E}[R_{\text{wait}}] \le 4J_{\max}\sqrt{\Lambda} \sum_{m=1}^{M} \sqrt{TK_m(L_m+1)} \le 4J_{\max}\sqrt{\Lambda}\,\Gamma.$$

The preceding post-launch and boundary calculation gives, pathwise on $\mathcal{E}_{\text{srv}}$,

$$H \le R_{\text{wait}} + 256J_{\max}\sqrt{\Lambda}\,\Gamma.$$

On every realization, the sets $\mathcal{U}_{m,j}$ are disjoint and every residual in them lies in $(0, 1]$, so $0 \le H \le T$. Consequently, on every realization,

$$H \le R_{\text{wait}} + 256J_{\max}\sqrt{\Lambda}\,\Gamma + T\mathbf{1}\{\mathcal{E}_{\text{srv}}^c\}.$$

Taking unconditional expectations and using $\mathbb{P}(\mathcal{E}_{\text{srv}}^c) \le T^{-2}$, $J_{\max} \le 3\Lambda$, and $\Lambda \ge 1$, we obtain

$$\begin{aligned}\mathbb{E}[H] &\le 260J_{\max}\sqrt{\Lambda}\,\Gamma + 1 \\ &\le 780\Lambda^{3/2}\Gamma + 1 \\ &\le 4096\Lambda^2\Gamma + 1.\end{aligned}$$

No independence between R_{wait} and $\mathcal{E}_{\text{srv}}$ is needed. This proves the lemma. □

Proof of Lemma D.10. We first recall why the three contributions in equation 24 cover all of Term (iii).

If $t \notin \mathcal{U}$, then $a_t^* \in \mathcal{G}_t$, so by definition $\hat{a}_t^* = a_t^*$. Hence the Term (iii) summand is zero.

Now fix $t \in \mathcal{U}$. By equation 23,

$$\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t) \le 2\text{CB}_{\hat{a}_t^*}(t) + \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a_t^*).$$

Write $m = \text{class}(a_t^*)$. There are three possibilities for the residual term.

First, if $\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le \Delta_t(a_t^*)$, then

$$\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a_t^*) \le 0,$$

so only the confidence term remains.

Second, suppose

$$\Delta_t(a_t^*) < \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \le \zeta_m.$$

Then

$$\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a_t^*) \le \zeta_m.$$

Summing these contributions over rounds whose optimal arm lies in class m costs at most $T_m\zeta_m$.

Finally, if

$$\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) > \max\{\Delta_t(a_t^*), \zeta_m\},$$

then, by definition, $t \in \mathcal{U}_{\text{large}}$. On these rounds we simply upper-bound

$$\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) - \Delta_t(a_t^*) \le \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)).$$

These three cases partition all rounds in $\mathcal{U}$. Together with the fact that Term (iii) vanishes outside $\mathcal{U}$, they give equation 24:

$$\sum_{t\in\mathcal{T}_{\mathrm{A}}} \big(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\big) \\ \leq 2\sum_{t\in\mathcal{T}_{\mathrm{A}}} \mathrm{CB}_{\hat{a}_t^*}(t) + \sum_{m=1}^{M} T_m\zeta_m + \sum_{t\in\mathcal{U}_{\mathrm{large}}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)).$$

It remains to bound these three sums.

For the first sum, $\hat{a}_t^* \in \mathcal{G}_t$ by Proposition D.5. Hence Lemma D.7 gives

$$2\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{A}}} \mathrm{CB}_{\hat{a}_t^*}(t)\right] \leq 256\sqrt{M(J_{\max}+1)\Lambda TD}.$$

For the second sum, equation 25 gives

$$\sum_{m=1}^{M} T_m\zeta_m \leq \Lambda\sqrt{MTD}.$$

It remains to bound the residual sum over $\mathcal{U}_{\mathrm{large}}$. Split these rounds according to whether their residual is at most $2^{-J_{\max}}$.

The rounds satisfying $\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \leq 2^{-J_{\max}}$ contribute at most

$$\sum_{\substack{t\in\mathcal{U}_{\mathrm{large}}:\\ \mu_{a_t^*}(t)-\hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))\leq 2^{-J_{\max}}}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \leq T\,2^{-J_{\max}} \leq 1.$$

Every remaining large-residual round satisfies $\mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) > 2^{-J_{\max}}$. Such a round has a unique optimal-arm class m, and there is a unique $j \in \{1,\dots,J_{\max}\}$ such that

$$2^{-j} < \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) \leq 2^{1-j}.$$

Hence these rounds are partitioned by the sets $\mathcal{U}_{m,j}$. Therefore

$$\sum_{\substack{t\in\mathcal{U}_{\mathrm{large}}:\\ \mu_{a_t^*}(t)-\hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))>2^{-J_{\max}}}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)) = \sum_{m=1}^{M}\sum_{j=1}^{J_{\max}}\sum_{t\in\mathcal{U}_{m,j}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*)).$$

Applying Lemma D.13 gives

$$\mathbb{E}\left[\sum_{\substack{t\in\mathcal{U}_{\mathrm{large}}:\\ \mu_{a_t^*}(t)-\hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))>2^{-J_{\max}}}} \mu_{a_t^*}(t) - \hat{\mu}_{a_t^*}^{(+1)}(I_t(a_t^*))\right] \leq 4096\Lambda^2\sqrt{MTD} + 1.$$

Combining the confidence, threshold, small-residual, and hard-residual contributions yields

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{A}}} \big(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\big)\right] \\ \leq 256\sqrt{M(J_{\max}+1)\Lambda TD} \\ + \Lambda\sqrt{MTD} \\ + 4096\Lambda^2\sqrt{MTD} + 2.$$

Here one additive 1 comes from the residuals below the smallest maintained scale, and the other comes from the failure probability in Lemma D.13.

Finally, using $J_{\max} + 1 \le 3\Lambda$ and $\Lambda \ge 1$, we obtain

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{A}}} \left(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\right)\right] \le 8192\Lambda^2\sqrt{MTD} + 2.$$

This proves Lemma D.10. □

D.2.5 Proof of the Unknown-Budget Expected Dynamic Regret Theorem

We now combine the three regret contributions and account for the rounds discarded by the stopped-process argument.

Theorem D.14 (Restatement of Theorem D.1)**.** *For every admissible instance,* `AEBF-H` *satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le 16384\Lambda^2\sqrt{MTD} + 3.$$

Equivalently,

$$\mathbb{E}[\mathrm{Regret}_T] = \widetilde{O}\Big(\sqrt{MTD}\Big).$$

Proof. Recall the three-term decomposition from Section D.2.4. For every round t,

$$\begin{aligned}\mu_{a_t^*}(t) - \mu_{a_t}(t) = {} & (\mu_{\hat{a}_t}(t) - \mu_{a_t}(t)) \\ & + \left(\mu_{\hat{a}_t^*}(t) - \mu_{\hat{a}_t}(t)\right) \\ & + \left(\mu_{a_t^*}(t) - \mu_{\hat{a}_t^*}(t)\right).\end{aligned}$$

These are Terms (i), (ii), and (iii), respectively.

By Lemma D.8, the expected stopped contribution of Term (i) is at most

$$4096\Lambda^2\sqrt{MTD}.$$

By Lemma D.9, the expected stopped contribution of Term (ii) is at most

$$4096\Lambda^2\sqrt{MTD}.$$

Finally, by Lemma D.10, the expected stopped contribution of Term (iii) is at most

$$8192\Lambda^2\sqrt{MTD} + 2.$$

Adding the three bounds gives

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{A}}} \left(\mu_{a_t^*}(t) - \mu_{a_t}(t)\right)\right] \\ &\quad \le 16384\Lambda^2\sqrt{MTD} + 2.\end{aligned}$$

It remains to account for the rounds after the stopping time. By Lemma D.2,

$$\mathbb{P}((\mathcal{C}')^c) \le T^{-1}.$$

By construction of τ_{A}, the omitted suffix $[T]\setminus\mathcal{T}_{\mathrm{A}}$ is nonempty only if $\mathcal{C}'$ fails. Since the instantaneous regret is at most one, the regret over this suffix is at most T. Therefore

$$\mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{\mathrm{A}}} \left(\mu_{a_t^*}(t) - \mu_{a_t}(t)\right)\right] \le T\mathbb{P}((\mathcal{C}')^c) \le 1.$$

Combining the stopped and omitted contributions yields

$$\mathbb{E}[\mathrm{Regret}_T] \le 16384\Lambda^2\sqrt{MTD} + 3.$$

By Λ definition, the theorem applies. □

E Additional Related Work

Adversarial online learning and reinforcement learning. Adversarial online learning allows arbitrary loss sequences and measures adaptation through quantities such as comparator switches, interval-wise regret, or comparator path length (Auer et al., 2002; Cutkosky, 2020). In adversarial reinforcement learning, Jin et al. (2020) establish regret against the best fixed policy for episodic Markov decision processes (MDPs) with adversarial losses, unknown transitions, and bandit feedback, while Zhao et al. (2022) study dynamic regret against changing policies in online MDPs with adversarial losses. Related nonstationary reinforcement-learning reductions give dynamic-regret guarantees in terms of global change counts or variation budgets (Wei & Luo, 2021). These works adapt to global temporal or comparator complexity, whereas our model localizes nonstationarity across known classes of arms.

Learning with corrupted feedback. Corruption-robust bandit models start from a stochastic instance but allow an adaptive adversary to alter rewards subject to a total corruption budget; their guarantees degrade with this budget while retaining stochastic behavior when it is zero (Lykouris et al., 2018; Gupta et al., 2019). This framework has been extended to linear bandits and episodic reinforcement learning with corrupted rewards or transitions (Bogunovic et al., 2021; Lykouris et al., 2021). Corruption budgets measure the total magnitude of adversarial alterations around a fixed stochastic instance, whereas our class-specific budgets count persistent mean changes and localize them by class.

Additional related work exploits structure in evolving environments. *Contextual* and *linear* nonstationary bandits use additional information in the form of contexts, policy classes, or linear reward models (Rakhlin & Sridharan, 2016; Luo et al., 2018; Russac et al., 2019). These works exploit structure in the reward model or context space, whereas our focus is on the *location of nonstationarity*: different arm groups may evolve at different rates. Restless bandits provide another model of evolving rewards, with different assumptions on how arm states change over time (Slivkins & Upfal, 2008; Wang et al., 2020; Seznec et al., 2020; Jiang et al., 2023).