MINIMAX-OPTIMAL REGRET FOR HETEROGENEOUS FEDERATED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

In heterogeneous federated bandits, agents on a communication graph face different reward distributions and seek the arm with the highest global mean, that is, the mean reward averaged across agents. We propose Adaptive Federated Elimination (AFE). It estimates global means from all agents' samples and removes clearly suboptimal arms early, so that agents incur little regret while waiting for messages. For N agents, K arms, and T rounds on any connected graph of diameter D, AFE achieves expected regret $O(D + \sqrt{KT/N})$ for every agent. We establish a matching lower bound on every connected graph, showing that AFE is minimax optimal. AFE also adapts to the suboptimality gaps Δ_i and achieves regret $O(\sum_{\Delta_i>0} \log(NT)/(N\Delta_i))$ plus a cost of waiting for messages that does not depend on T. This bound applies the same factor $\log(NT)$ to every arm. We refine it and show that each arm needs only a smaller factor, which depends on the arm's gap and on how many arms have gaps smaller than or close to that gap. When all suboptimal arms have similar gaps, a lower bound shows that AFE is optimal up to additive costs that do not depend on T. In a complete-graph example, AFE's refined bound is smaller than the guarantees of recent algorithms by a factor of order $\log(NT)/\log\log(NT)$.

1 INTRODUCTION

In heterogeneous federated multi-armed bandits, agents on a communication graph face different reward distributions and cooperate to find a common best arm. The target is the arm with the highest global mean, that is, the mean reward averaged across agents. An arm that is best for one agent may be poor for others, so no agent can solve the problem by learning its own best arm. We measure performance by the largest expected regret among agents.

Cooperation lets agents pool their samples, but these samples are spread over the graph. Estimating a global mean requires information from all agents, and messages from distant agents take time to arrive. Agents keep choosing arms and incurring regret while they wait. An efficient algorithm must therefore control both the statistical cost of distinguishing arms and the cost of waiting, namely the regret incurred while information travels.

Early federated bandit algorithms estimate global means through a central server or through gossip (Shi & Shen, 2021; Zhu et al., 2021). On communication graphs, Zhang et al. (2025a) propose DRRB, which combines consensus with elimination, and prove finite-time regret bounds together with asymptotic lower bounds. Xu & Klabjan (2025) establish statistical lower bounds on networks. Recent work also studies bit-efficient communication, anytime guarantees, and reductions to delayed bandits (Zhang et al., 2025b; Hu et al., 2026; Qiu et al., 2026). However, it remains unclear how the optimal worst-case regret depends on the numbers of agents and arms, the horizon, and the graph. We determine this dependence and propose Adaptive Federated Elimination (AFE), an algorithm that attains the minimax optimal regret.

Minimax optimality on every connected graph. For N agents, K arms, and T rounds on a connected graph of diameter D, we show that the minimax regret is $\Theta(D + \sqrt{KT/N})$ whenever $2 \le N, K \le T$, and AFE attains this rate. Our lower bound holds on every connected graph and for every protocol. The term $\sqrt{KT/N}$ is the statistical cost of learning from the pooled samples of

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.

all agents. The term D is the cost of waiting for distant agents, whose observations may change the global best arm. Thus, for fixed N, K, and T, the graph affects the minimax regret only through its diameter, and only additively. AFE matches this lower bound by keeping both costs small. To control the statistical cost, AFE combines the samples of all agents into unbiased estimates of the global means. It also stops increasing the precision once the next level would need more samples than the horizon provides. To control the cost of waiting, AFE first runs a localization stage that removes clearly suboptimal arms, so agents incur little regret while they wait for later messages.

Adaptivity to easier instances. The minimax rate describes the hardest instances, but many instances are easier. Let Δ_i be the gap between the best global mean and the global mean of arm i. AFE adapts to these gaps and achieves regret $O(\sum_{\Delta_i>0} \log(NT)/(N\Delta_i))$ plus a cost of waiting that does not depend on T. This bound grows only logarithmically with T, but it applies the same factor $\log(NT)$ to every arm. We refine this bound and show that, as in single-agent bandits (Lattimore, 2015; 2018), each arm needs only a smaller factor. This factor depends on the arm's gap and on how many arms have gaps smaller than or close to that gap, since arms with much larger gaps are eliminated early. The refined bound adds a localization cost, which also does not depend on T.

Tightness and comparison. To show that this smaller factor is tight, we prove a lower bound on instances where all suboptimal arms have similar gaps. The lower bound shows that AFE is optimal on these instances up to the costs of waiting and localization, even though AFE does not know the scale of the gaps. Unlike the asymptotic lower bounds discussed above, our lower bound holds at every finite horizon and requires no consistency assumption. It therefore applies even when the gaps shrink as the horizon grows, as in the complete-graph example of Section 5. In this example, AFE's refined bound is smaller than the guarantees of four recent algorithms (Zhang et al., 2025a;b; Hu et al., 2026; Qiu et al., 2026) by a factor of order $\log(NT)/\log\log(NT)$.

We discuss further related work in Appendix H.

2 Problem Setup

Throughout the paper, we consider N agents, K arms, and T rounds, where $2 \le N, K \le T$. In each round $t \in [T]$, every agent $v \in [N]$ pulls an arm $A_v(t) \in [K]$ and observes a reward $Y_v(t) \in [0,1]$. The reward of arm i at agent v follows a fixed distribution ν_{iv}, and rewards are independent across agents, arms, and rounds. These distributions may differ arbitrarily across agents, and the agents do not know them.

The agents seek a common best arm defined by the global means. For each arm i and agent v, we write the local mean, the global mean, the best arm, and the gap as

$$\mu_{iv} = \mathbb{E}_{\nu_{iv}} Y, \qquad \mu_i = \frac{1}{N}\sum_{v=1}^{N} \mu_{iv}, \qquad b = \arg\max_i \mu_i, \qquad \Delta_i = \mu_b - \mu_i.$$

We assume that b is unique. For example, consider two agents and two arms. If the local means are $(0.9, 0.6)$ at the first agent and $(0.1, 0.6)$ at the second, then the global means are $(0.5, 0.6)$. Both agents should play arm 2, although the first agent prefers arm 1 locally. Thus the agents observe different rewards but share the same target.

The agents communicate over a connected undirected graph $G = ([N], E)$ with diameter D, so that $1 \le D \le N - 1 < T$. They share a clock and have unique identifiers. Each agent knows N, K, T, D, and its own neighbors, but not the full graph. In each round, an agent may send a message of any size to each of its neighbors. A message crosses at most one edge per round and can affect actions only in later rounds. Hence information from an agent at distance d can influence another agent's actions only after d rounds. A protocol π specifies how each agent chooses arms and messages based on its own past actions, rewards, and received messages. Protocols may use shared or private randomness that does not depend on the instance.

Each pull of arm i incurs regret Δ_i. The regret of agent v and the largest expected regret among agents are

$$R_v(T) = \sum_{t=1}^{T} \Delta_{A_v(t)}, \qquad \mathcal{R}_{\max_\nu}^{\pi}(T) = \max_{v\in[N]} \mathbb{E}_\nu^\pi R_v(T). \tag{1}$$

Every round counts toward $R_v(T)$, including rounds in which agents wait for messages. The maximum is taken over expected regrets, so every agent must have small expected regret. Every trajectory satisfies $R_v(T) \le T\Delta_{\max} \le T$, where $\Delta_{\max} = \max_i \Delta_i$. Our displayed upper bounds omit this trivial cap.

Let $\mathcal{V}(G, K)$ be the class of all reward instances on G with K arms and a unique best arm. We study the minimax regret

$$\mathfrak{R}(G, K, T) = \inf_{\pi} \sup_{\nu \in \mathcal{V}(G,K)} \mathcal{R}_{\max \nu}^{\pi}(T), \tag{2}$$

where the infimum is over all protocols. We write $[n] = \{1, \dots, n\}$ for a positive integer n. Logarithms are natural unless a base is shown.

3 Adaptive Federated Elimination

Adaptive Federated Elimination (AFE) keeps a candidate set of arms that may still be best and shrinks it at successively finer precisions. It first runs a localization stage that removes clearly suboptimal arms, so that agents incur little regret while they later wait for messages. It then runs elimination stages. Each stage selects a reference arm and uses new samples to decide which arms to remove. All sampling lengths and stopping decisions depend only on the known inputs N, K, T, D and on the current candidate set.

3.1 Estimating global means

A global mean weights all agents equally, so its estimate must do the same even when local means differ. To estimate the arms in a common set S, AFE runs a sampling block of u rounds. In each of these rounds, every agent pulls an arm drawn uniformly at random from S and records its reward sum for each arm. The agents then pass these sums through the graph for D rounds. Each agent forwards every sum it receives, labeled by the block and by the agent that produced it, and drops duplicates. After D rounds, every agent holds the sums of all N agents and computes

$$\widehat{\mu}_i = \frac{|S|}{Nu} \sum_{v=1}^{N} \sum_{t=1}^{u} \mathbf{1}\{A_v(t) = i\} Y_v(t), \qquad i \in S. \tag{3}$$

Each pull selects arm i with the known probability $1/|S|$, so this estimate is unbiased, that is, $\mathbb{E}\widehat{\mu}_i = N^{-1} \sum_v \mu_{iv} = \mu_i$. An agent that never pulls arm i simply contributes zero. The estimate therefore needs neither equal pull counts across agents nor identical local means. Lemma C.1 shows that, conditional on the history before the block starts,

$$\Pr\big(|\widehat{\mu}_i - \mu_i| > z \mid \text{start history}\big) \le 2\exp\left(-\frac{Nuz^2}{3|S|}\right), \qquad 0 < z \le 1.$$

Compared with a single agent, pooling the samples of all N agents shortens the sampling time needed for a given accuracy by a factor of N.

Passing the sums takes D rounds, and agents must still pull arms in these rounds. While waiting, every agent pulls the smallest-label arm in the current candidate set. These rewards are discarded, but their regret counts. AFE uses the estimate as it is, even when its value falls outside $[0, 1]$. All agents receive the same statistics and break ties by label, so they keep identical candidate sets and make the same decisions without a central server.

3.2 Localization and elimination

AFE starts with a localization stage, which screens out clearly suboptimal arms and produces the initial candidate set A_0. Localization works at precisions $1, 1/2, 1/4, \dots$ until the precision reaches $\beta = \min\{1, \sqrt{K/[N(D+1)\log(eK)]}\}$. At each precision, it samples all remaining arms and removes the arms whose empirical gap exceeds half of that precision. With high probability, A_0 contains the best arm, and every arm in A_0 has gap at most β. The stage thus narrows the search to a neighborhood of the best arm, which gives the stage its name. Localization uses a high confidence level because its deletions affect all later decisions. The value of β balances the samples spent on

localization against the regret of later waiting rounds. When $\beta = 1$, AFE skips localization and sets $A_0 = [K]$.

After localization, AFE runs elimination stages $r = 0, 1, 2, \ldots$. Stage r works at precision $\varepsilon_r = \beta 2^{-r}$ and updates the candidate set A_r in two steps. The first step selects a reference arm J_r. AFE copies A_r into a temporary set and repeatedly keeps its empirical top half, using new samples for each halving, until one arm remains. We call this procedure MEDIAN. Deletions from the temporary set do not change A_r, so an arm discarded here can still survive the stage. MEDIAN finds a reference whose mean is close to the best mean in A_r, without requiring accurate estimates of all arms at once. When agents wait for messages inside MEDIAN, they still pull an arm from A_r rather than from the temporary set.

The second step samples every arm in A_r again with new data and updates the candidate set by

$$A_{r+1} = \{i \in A_r : \widehat{\mu}_{J_r} - \widehat{\mu}_i \le \varepsilon_r\}. \tag{4}$$

Because this test uses new samples, its estimation errors are independent of the data used to choose the reference. The reference always passes the test, so the candidate set never becomes empty. Arms with large gaps leave in early, coarse stages, while arms close to the best receive more samples in later stages.

To see why the test works, suppose the reference is within ε_r of the best mean in A_r, and the new estimates of the reference and of arm i are both within $\varepsilon_r/4$ of their means. Then arm i passes the test if it is the best arm in A_r, and it fails the test if its mean is more than $3\varepsilon_r$ below the best mean. A poor reference only makes the test more lenient. It may keep some suboptimal arms for another stage, but it cannot cause the best arm in A_r to be removed unless the new estimates themselves are inaccurate. Our analysis therefore handles two failure modes separately. It bounds the probability that a clearly suboptimal arm survives a stage, and it bounds the regret caused by deleting the best remaining arm. It does not require all estimates in all stages to be accurate at the same time.

AFE stops refining when one arm remains or when $|A_r| \ge NT\varepsilon_r^2$. The second condition means that estimating all remaining arms to precision ε_r would need more samples than the NT samples available. If one arm remains, all agents pull it until round T. Otherwise, they pull the previous reference J_{r-1}, which is still in A_r, or the smallest-label arm in A_0 if no stage has run. Algorithm 1 summarizes AFE, and Appendix B gives the full localization schedule and the MEDIAN subroutine.

Algorithm 1 Adaptive Federated Elimination (AFE): main steps

Require: N, K, T, D, agent IDs, neighbor lists, and a common clock.

Set $\beta = \min\{1, \sqrt{K/[N(D+1)\log(eK)]}\}$.
Obtain A_0 by localization if $\beta < 1$; otherwise set $A_0 = [K]$.
$r \leftarrow 0$.
while rounds remain **do**
  Set $a_r = |A_r|$ and $\varepsilon_r = \beta 2^{-r}$.
  if $a_r = 1$ **then**
    Play the sole arm until round T; **return**.
  else if $a_r \ge NT\varepsilon_r^2$ **then**
    Play J_{r-1} until round T, or $\min A_0$ if $r = 0$; **return**.
  end if
  Set $L_r = \log(128NT\varepsilon_r^2/a_r)$ and $p_r = e^{-L_r}$.
  Choose $J_r = \text{MEDIAN}(A_r, \varepsilon_r, p_r)$ using temporary halving sets.
  Sample A_r afresh for $\lceil 48 a_r L_r/(N\varepsilon_r^2) \rceil$ rounds; pass the sums for D rounds.
  Update A_{r+1} by (4); set $r \leftarrow r + 1$.
end while

Every operation stops at round T, including localization and MEDIAN. An interrupted sampling block or message-passing step produces no estimate and removes no arm, but its pulls still count toward regret. AFE therefore uses a statistic only after every agent has received it. Section 4 analyzes the full trajectory of AFE, including rounds spent waiting for messages, the final phase in which agents pull a single arm, and the regret after a wrong deletion.

4 MINIMAX OPTIMALITY

4.1 MAIN THEOREM

AFE controls both the statistical cost of learning the best arm and the cost of waiting for messages. The following theorem shows that the sum of these two costs is the minimax rate on every connected graph.

Theorem 4.1 (Minimax optimality of AFE)**.** *For every connected G and $2 \le N, K \le T$,*

$$\inf_{\pi} \sup_{\nu \in \mathcal{V}(G,K)} \mathcal{R}_{\max}^{\pi}(T) = \Theta\Big(D + \sqrt{KT/N}\Big), \tag{5}$$

and π_{AFE} attains the upper bound.

The term $\sqrt{KT/N}$ is the statistical cost. It has the same order as the minimax regret of a single learner that makes all NT pulls, divided by N, so it remains even if messages arrive instantly. The term D is the cost of waiting for distant agents, whose observations may change the global best arm. The theorem holds for each fixed graph, since its lower bound uses only reward instances on that graph and its upper bound uses AFE with that graph's diameter.

The two terms are equal when $T = ND^2/K$, and the statistical term dominates for larger horizons. On a complete graph, $D = 1$ and the rate is $\Theta(\sqrt{KT/N})$, because $T \ge N$ implies $\sqrt{KT/N} \ge \sqrt{K} > 1$. On a path, $D = N - 1$, and the cost of waiting can dominate even though many agents contribute samples. Reducing the diameter lowers the rate only until the statistical term dominates.

4.2 UPPER BOUND

Pooling samples controls the statistical cost. After u sampling rounds on a set S, the estimate (3) has variance $O(|S|/(Nu))$. Estimating the arms in S to precision ε therefore takes about $|S|/(N\varepsilon^2)$ rounds, times a confidence factor. In elimination stage r, this factor is $L_r = \log(128NT\varepsilon_r^2/|A_r|)$. The analysis of these sampling costs uses no union bound over the candidates. It compares each stage with the precision $\eta = \sqrt{|A_0|/(NT)}$, at which AFE would stop if no arm had been removed. Stages coarser than η use larger confidence factors, but their costs shrink geometrically as the precision grows. Stages finer than η run only after enough arms have been removed, and their costs shrink geometrically as the precision decreases. Summing both parts, the elimination stages incur regret $O(\sqrt{KT/N})$ with no logarithmic factor.

The final phase stays within the same budget. If AFE stops because $|A_r| \ge NT\varepsilon_r^2$, then $\varepsilon_r \le \eta$, and all agents pull the previous reference J_{r-1}. Its expected gap to the best remaining mean is $O(\varepsilon_r)$, so the final phase costs $O(T\eta)$, which is $O(\sqrt{KT/N})$.

Localization controls the cost of waiting. AFE runs many sampling operations, and each needs D more rounds to pass its statistics. If every waiting round cost constant regret, the total would be D times the number of operations, which exceeds the optimal D term by logarithmic factors. After successful localization, every remaining arm has gap at most β, so each waiting round costs at most β. The waiting cost after localization is then at most $C\beta d[\log(eK)]^{3/2}$, where $d = D + 1$. The choice of β balances this cost against the samples spent on localization.

Elimination can also delete the best arm, and such a deletion affects all later rounds. Start counting rounds just after successful localization, which keeps the best arm b. Let M_t be the best remaining mean in round t, and suppose the update after round τ_q lowers this mean by $c_q \ge 0$. Then, on every trajectory and for every number h of remaining rounds,

$$\sum_{t=1}^{h} (\mu_b - \mu_{A_v(t)}) = \sum_{t=1}^{h} (M_t - \mu_{A_v(t)}) + \sum_{q:\tau_q \le h} (h - \tau_q) c_q.$$

The first sum is the regret relative to the best remaining mean, including rounds spent waiting for messages. The second sum is the lasting cost of every wrong deletion, and the validation test bounds the expected size of each drop c_q. Neither sum vanishes after AFE stops sampling, so bounding the survival of suboptimal arms alone would not suffice. A single continuation bound (Lemma C.9)

controls AFE from the start of any elimination stage, using only future samples. It also covers interrupted operations and the final phase.

Combining these steps, including the cost of localization errors, gives

$$\mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le C\left[\sqrt{KT/N} + d + \sqrt{Kd/N}\sqrt{\log(eK)}\log(eNT)\right]. \tag{6}$$

The logarithms appear only in the last term, and Lemma C.10 shows $\sqrt{KD/N} \le T^{-1/6}(D + \sqrt{KT/N})$. Because $NT \le T^2$, this polynomial decay absorbs the logarithms, so the bound is $O(D + \sqrt{KT/N})$ with an absolute constant. Theorem C.3 states the complete bound together with its gap-dependent form, and Appendix C proves it.

4.3 Lower bound

Two families of hard instances on the given graph give the two terms of the lower bound, and both bounds hold for every protocol. Appendix D gives the full proof.

For the statistical term, start from a reference instance in which every arm has Bernoulli mean $1/2$ at every agent. Under this instance, some arm i receives at most NT/K pulls in expectation across all agents. Raising the mean of this arm by α at every agent makes it the unique best arm. We compare the distributions P_0 and P_i of the full transcript under the two instances. The transcript contains all actions, rewards, messages, and random seeds, so it captures everything a protocol can use. Only pulls of arm i contribute to the divergence, so $\mathrm{KL}(P_0, P_i) = \mathbb{E}_0 n_i(T)\ \mathrm{kl}(1/2, 1/2+\alpha) \le C\alpha^2 NT/K$, where $n_i(T)$ counts the pulls of arm i across all agents. Choosing α as a small constant times $\sqrt{K/(NT)}$ keeps this divergence small. Pinsker's inequality then forces average regret $\Omega(\sqrt{KT/N})$ across agents, and the largest expected regret is at least the average (Lemma D.2). The reference instance has no unique best arm, but it serves only for comparison, and the raised instance belongs to the model.

For the diameter term, take two vertices a and z at distance D, and set $r = \lceil D/2 \rceil$. The balls of radius $r-1$ around a and z are disjoint, so one of them contains at most $N/2$ vertices. Let v be its center, and let S be the set of vertices at distance at least r from v. Then $q = |S|/N \ge 1/2$, so the agents in S account for at least half of each global mean. Yet their observations cannot affect the actions of v during the first r rounds.

We build two Bernoulli instances indexed by $\sigma \in \{+1, -1\}$. At each agent in S, arms 1 and 2 have means $(1 + \sigma\lambda)/2$ and $(1 - \sigma\lambda)/2$, where $\lambda = 1/(2q) \le 1$. Outside S, both arms have mean $1/2$, and all other arms have mean zero everywhere. The global means of arms 1 and 2 are $3/4$ and $1/4$ when $\sigma = +1$, and the order reverses when $\sigma = -1$. Changing only the observations in S thus switches the unique best arm while keeping the gap at $1/2$. Hence, up to round r, the actions of v have the same distribution under both instances. For any common action distribution, the expected regrets of one round under the two instances sum to at least $1/2$. Summing over the first r rounds gives $\mathbb{E}_+ R_v(r) + \mathbb{E}_- R_v(r) \ge r/2$, so one of the two instances forces regret at least $r/4 \ge D/8$ at agent v. Neither unlimited message sizes nor messages in every round can remove this delay (Lemma D.4).

The two lower bounds use different instances but bound the same minimax regret. Their maximum is at least half their sum, which proves the lower bound in (5) (Theorem D.5).

5 Gap-Dependent Guarantees

The minimax rate describes the hardest instances. This section shows how the same AFE adapts to easier ones. We first refine its instance-dependent upper bound by counting the arms that compete at each gap scale. We then prove a lower bound that matches this refined bound on instances with similar gaps, up to the costs of waiting and localization. Finally, we use this result to compare AFE with existing guarantees.

5.1 INSTANCE-DEPENDENT UPPER BOUNDS

Arms with larger gaps can be eliminated with fewer samples. Summing the cost of each arm until its elimination gives the following logarithmic bound. The factor $1/N$ reflects that all agents sample in parallel.

Theorem 5.1 (Instance-dependent regret of AFE)**.** *For every connected G, $2 \le N, K \le T$, and every $\nu \in \mathcal{V}(G, K)$, AFE satisfies*

$$\mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le C \left(\Upsilon + \frac{\log(eNT)}{N} \sum_{i \neq b} \frac{1}{\Delta_i} \right), \tag{7}$$

where $d = D + 1$, $J = \log(eK)$, and $\Upsilon = d + \beta d J^{3/2}$, using AFE's localization precision β.

The sum in (7) includes the sampling cost of localization. The term Υ covers the rounds spent waiting for messages, including those during localization, and the extra rounds created by rounding sample sizes up to integers. The proof groups arms with similar gaps and charges each group's sampling until its arms are eliminated. It also bounds all regret after a wrong deletion, using the continuation bound from Section 4. Theorem C.3 and Appendix E give the full proof.

Theorem 5.1 applies the same factor $\log(eNT)$ to every arm. This factor is larger than necessary, because arms with large gaps leave early and fewer arms compete at finer precisions. To capture this effect, fix a suboptimal arm i. We count every arm whose gap is at most Δ_i fully, and we weight an arm with a larger gap by the squared ratio of the two gaps. This gives the effective count $k_i^{(2)}$ and the sampling scale $H_i^{(2)}$,

$$k_i^{(2)} = \sum_{j=1}^{K} \min\{1, (\Delta_i/\Delta_j)^2\}, \qquad H_i^{(2)} = \frac{k_i^{(2)}}{\Delta_i^2}, \tag{8}$$

where the term for $j = b$ equals one. Equivalently, $H_i^{(2)} = \sum_j \min\{\Delta_i^{-2}, \Delta_j^{-2}\}$. Up to constant and logarithmic factors, this is the number of pooled samples needed to bring every arm either to precision Δ_i or to its own elimination. With $\psi(x) = \max\{1, \log x\}$, the refined statistical term is

$$F_2 = \frac{1}{N} \sum_{i \neq b} \frac{\psi(NT/H_i^{(2)})}{\Delta_i}. \tag{9}$$

Each arm's logarithmic factor now depends on the ratio $NT/H_i^{(2)}$ between all available samples and the samples needed to resolve that arm, rather than on the horizon alone. Since $H_i^{(2)} \ge 2$, we have $\psi(NT/H_i^{(2)}) \le \log(eNT)$, so F_2 never exceeds the statistical term of Theorem 5.1. This square-weighted count also appears in the analysis of single-agent bandits (Lattimore, 2015).

For example, suppose m suboptimal arms have gap δ_1 and the other $K - m - 1$ have gap $\delta_2 > \delta_1$, where $1 \le m \le K - 2$. For an arm in the first group,

$$k_i^{(2)} = m + 1 + (K - m - 1)(\delta_1/\delta_2)^2, \qquad H_i^{(2)} = \frac{m+1}{\delta_1^2} + \frac{K - m - 1}{\delta_2^2}.$$

The best arm and the m arms with gap δ_1 count fully. The remaining arms contribute less, because they are eliminated at the coarser precision δ_2. When $(K - m - 1)(\delta_1/\delta_2)^2 \le m + 1$, the effective count lies between $m + 1$ and $2(m + 1)$, even if K is much larger. For an arm in the second group, the count is K. Different gap scales thus use different counts in the same bound. These counts appear only in the analysis, and AFE does not need them as inputs.

The refined bound must also account for localization. Localization uses one high confidence level for all arms, so only part of its sampling cost fits into F_2. The rest does not depend on T. With d, J, and β as above and $Q_0 = \lceil \log_2(1/\beta) \rceil$, this remaining localization cost is

$$\mathcal{B}_{\mathrm{loc}}^{(2)} = \begin{cases} 0, & \beta = 1, \\ \dfrac{K}{N\beta} \big[1 + \log K + \log(4dJ(Q_0 + 1)) \big], & \beta < 1. \end{cases} \tag{10}$$

Theorem 5.2 (Effective-competition regret of AFE)**.** *For every connected G, $2 \le N, K \le T$, and every $\nu \in \mathcal{V}(G, K)$, AFE satisfies*

$$\mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le C\big(F_2 + \Upsilon + \mathcal{B}_{\mathrm{loc}}^{(2)}\big). \tag{11}$$

This bound holds together with Theorems 4.1 and 5.1.

The gaps and effective counts in this theorem appear only in the analysis, and AFE learns its candidate sets from data. The proof tracks the random sizes of these sets. While the best arm remains, the survival probability of an arm decreases geometrically once its gap exceeds the current precision by a constant factor. Concavity controls the confidence factors at these random counts, and summing over gap scales yields F_2. The same accounting covers all rounds after the best arm is first deleted. Appendix E gives the full proof, including (10).

5.2 Tightness for similar gaps

To test the statistical term of Theorem 5.2, we consider instances in which all suboptimal arms have similar gaps. Their effective counts then have the same order, so F_2 reduces to a single statistical scale that a lower bound can match. Fix $0 < \Gamma \le 1/8$, and let $\mathcal{V}_\Gamma(G, K)$ contain the instances whose positive gaps all lie in $[\Gamma, 2\Gamma]$. We call this the *similar-gap class*. On this class, every suboptimal pull costs between Γ and 2Γ, including pulls made while waiting for messages, so $R_v(T) \le 2\Gamma T$ always. Define

$$\mathfrak{R}_\Gamma(G, K, T) = \inf_\pi \sup_{\nu \in \mathcal{V}_\Gamma(G,K)} \mathcal{R}_{\max_\nu}^{\pi}(T), \qquad S_\Gamma(T) = \frac{K}{N\Gamma} \log\left(1 + \frac{NT\Gamma^2}{K}\right). \tag{12}$$

The ratio $x = NT\Gamma^2/K$ compares the available samples with the samples needed to resolve K gaps of size Γ. For $x \le 1$, $S_\Gamma(T) = \Theta(\Gamma T)$. For $x \ge 1$, $S_\Gamma(T)$ has order $K\psi(x)/(N\Gamma)$. The class serves only the analysis, and AFE receives neither Γ nor the information that the instance belongs to the class.

Theorem 5.3 (Bounds for similar gaps)**.** *For every connected G, $2 \le N, K \le T$, and every $0 < \Gamma \le 1/8$, let $B_G = \Upsilon + \mathcal{B}_{\mathrm{loc}}^{(2)}$. Then*

$$\begin{aligned} c\big[S_\Gamma(T) + \Gamma D\big] &\le \mathfrak{R}_\Gamma(G, K, T) \\ &\le \sup_{\nu \in \mathcal{V}_\Gamma(G,K)} \mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \\ &\le C\big[S_\Gamma(T) + B_G\big]. \end{aligned} \tag{13}$$

The lower bound confirms that $S_\Gamma(T)$ is the right statistical scale. Whenever $B_G = O(S_\Gamma(T))$, AFE attains the class-optimal rate $\mathfrak{R}_\Gamma(G, K, T) = \Theta(S_\Gamma(T))$. The costs of waiting and localization are not matched in general, since the lower bound contains ΓD while the upper bound contains B_G.

For the statistical lower bound, take a Bernoulli instance in which arm 1 has global mean $1/2 + \Gamma$ and every other arm has global mean $1/2$, and add zero-sum offsets across agents. For each $i \ge 2$, a second instance raises the mean of arm i by 2Γ at every agent. All these instances belong to the similar-gap class. Under the first instance, every pull of a suboptimal arm costs at least Γ, so a protocol with small regret pulls some arm $i \ge 2$ rarely. The transcripts of the two instances then differ little, so the protocol also rarely pulls arm i under the second instance, where arm i is best. The Bretagnolle–Huber inequality turns this trade-off into the lower bound $S_\Gamma(T)/256$ on the largest average regret over the family, and the largest expected regret is at least this average. The argument uses the regret on the family itself to limit how much information a protocol can collect, so it needs no consistency assumption. A version of the construction in Section 4 with gaps scaled by Γ gives the term ΓD. For the upper bound, similar gaps imply $K/4 \le k_i^{(2)} \le K$, so F_2 reduces to $S_\Gamma(T)$ after the cap $2\Gamma T$. Theorem F.3 proves the full statement.

5.3 Comparison with existing bounds

We now use this criterion to compare AFE with existing bounds. On the following complete-graph sequence, the costs of waiting and localization are smaller than the statistical term, so the lower

bound certifies the rate of AFE. Let n tend to infinity through even integers and set

$$N = K = n, \qquad G = K_n, \qquad T = n^4, \qquad \ell = \log n, \qquad \Gamma = \frac{\sqrt{\ell}}{n^2}.$$

Then $x = \ell$, $S_\Gamma(T) = \Theta(n^2 \log \ell/\sqrt{\ell})$, and $B_G = O(\ell^{3/2}) = o(S_\Gamma(T))$. AFE therefore attains the optimal rate on the similar-gap class,

$$\mathfrak{R}_\Gamma(K_n, n, n^4) = \Theta\left(\frac{n^2 \log \ell}{\sqrt{\ell}}\right).$$

Table 1 compares the complete regret bounds of existing algorithms on the heterogeneous family of Proposition G.1. The natural cap is $T\Delta_{\max} = \Theta(n^2\sqrt{\ell})$. Even after taking this cap and the best available branch, every existing bound has order $n^2\sqrt{\ell}$. These bounds are thus larger than that of AFE by a factor of order $\ell/\log \ell$. Since $NT = n^5$, this factor is $\Theta(\log(NT)/\log\log(NT))$.

Table 1: Orders of the complete regret bounds on the sequence above. The last column takes the minimum over the available branches and the cap $T\Delta_{\max}$. The other branch is a gap-free or adversarial bound, depending on the method.

Method	Stochastic bound	Other branch	With cap
DRRB (Zhang et al., 2025a)	$\Theta(n^2\sqrt{\ell})$	$\Theta(n^2\sqrt{\ell})$	$\Theta(n^2\sqrt{\ell})$
EpoInc-SE (Zhang et al., 2025b)	$\Omega(n^2\sqrt{\ell})$	—	$\Theta(n^2\sqrt{\ell})$
FedFTRL (Hu et al., 2026)	$\Theta(n^3/\sqrt{\ell})$	$\Theta(n^2\ell)$	$\Theta(n^2\sqrt{\ell})$
Black-box (Qiu et al., 2026)	$\Theta(n^3\sqrt{\ell})$	$\Theta(n^2\ell)$	$\Theta(n^2\sqrt{\ell})$
AFE	$\Theta(n^2 \log \ell/\sqrt{\ell})$	$\Theta(n^2)$	$\Theta(n^2 \log \ell/\sqrt{\ell})$

The improvement comes from the confidence factor. On this family, $H_i^{(2)} = \Theta(K/\Gamma^2)$, so $\psi(NT/H_i^{(2)}) = \Theta(\log \ell)$. Replacing this factor by $\log(eNT) = \Theta(\ell)$ would give the statistical order $n^2\sqrt{\ell}$, the same as the existing bounds. Appendix G keeps every term of the existing bounds and converts their high-probability bounds to bounds in expectation. The comparison concerns proved bounds, not the actual regret of those algorithms, and several of these algorithms also offer anytime, adversarial, or communication guarantees that our objective does not measure.

6 Conclusion

We studied heterogeneous federated bandits on a communication graph, where every round counts toward regret. AFE attains the minimax rate $\Theta(D + \sqrt{KT/N})$ on every connected graph. It also adapts to the gaps, and its refined bound replaces the common factor $\log(NT)$ with a smaller factor for each arm. When all suboptimal arms have similar gaps, a lower bound shows that AFE is optimal up to the costs of waiting and localization.

Our analysis assumes independent bounded rewards, a static graph, known values of N, K, T, and D, and messages of unlimited size. For gap-dependent regret, our lower bound on instances with similar gaps contains the additive term ΓD, while AFE pays the larger costs of waiting and localization. Determining the optimal additive term remains open. It is also open whether the refined bound is optimal when the gaps differ widely. Extending these guarantees to limited message sizes or an unknown horizon is another direction for future work.

AI use statement

In this work, we used generative AI tools mainly for polishing the writing, searching the literature, and assisting with proofs. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All results in this paper are theoretical. Section 2 states the model and all assumptions. Appendix B specifies AFE completely, including the localization schedule, the MEDIAN subroutine, and all sampling lengths. Appendix A maps each result of the main text to its proof, and Appendices C to F prove all upper and lower bounds. Appendix G lists the exact expressions and parameters that we use to evaluate the existing bounds in Table 1.

References

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 39–75, 2017.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Idan Barnea, Tal Lancewicki, and Yishay Mansour. Individual regret in cooperative stochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Eshcar Hillel, Zohar S. Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Benjamin Howson, Sarah Lucie Filippi, and Ciara Pike-Burke. QuACK: A multipurpose queuing algorithm for cooperative k-armed bandits. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1927–1935, 2025.

Zicheng Hu, Zihao Wang, and Cheng Chen. A near-optimal best-of-both-worlds algorithm for federated bandits. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1238–1246, 2013.

Nikolai Karpov and Qin Zhang. Parallel best arm identification in heterogeneous environments. In *Proceedings of the 36th ACM Symposium on Parallelism in Algorithms and Architectures*, pp. 53–64, 2024a.

Nikolai Karpov and Qin Zhang. Communication-efficient collaborative regret minimization in multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13076–13084, 2024b.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248, 2016.

Tor Lattimore. Optimally confident UCB: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

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

Jingyuan Liu, Hao Qiu, Lin Yang, and Mengfan Xu. Distributed multi-agent bandits over Erdős–Rényi random networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

Hao Qiu, Mengxiao Zhang, and Nicolò Cesa-Bianchi. Near-optimal regret for distributed adversarial bandits: A black-box approach. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5465–5517, 2026.

Clémence Réda, Sattar Vakili, and Emilie Kaufmann. Near-optimal collaborative learning in bandits. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9603–9611, 2021.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pp. 126–146. IEEE, 2019.

Xuchuang Wang, Lin Yang, Yu-zhen Janice Chen, Xutong Liu, Mohammad Hajiesmaili, Don Towsley, and John C. S. Lui. Achieving near-optimal individual regret & low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020.

Mengfan Xu and Diego Klabjan. Multi-agent multi-armed bandit regret complexity and optimality. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1639–1647, 2025.

Haoran Zhang, Xuchuang Wang, Hao-Xu Chen, Hao Qiu, Lin Yang, and Yang Gao. Near-optimal regret bounds for federated multi-armed bandits with fully distributed communication. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4959–4981, 2025a.

Haoran Zhang, Yang Xu, Xuchuang Wang, Hao-Xu Chen, Hao Qiu, Lin Yang, and Yang Gao. Federated multi-armed bandits with efficient bit-level communications. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(1):1–29, 2021.

A GUIDE TO THE APPENDICES

The appendices keep the setting and notation of Section 2. Since $2 \le N, K \le T$, we have $D < T$ and $K < NT$, and several proofs use these two facts. Besides the largest expected regret $\mathcal{R}_{\max_\nu}^{\pi}(T)$ and the minimax regret $\mathfrak{R}(G, K, T)$, we use the average regret across agents,

$$\overline{\mathcal{R}}_\nu^{\pi}(T) = \frac{1}{N} \sum_{v=1}^{N} \mathbb{E}_\nu R_v^{\pi}(T).$$

The maximum is at least the average, so every lower bound on $\overline{\mathcal{R}}_\nu^{\pi}(T)$ also bounds $\mathcal{R}_{\max_\nu}^{\pi}(T)$. Appendix B gives the full specification of AFE in Algorithms 3 and 2. It also fixes the parameters $H = NT$, $d = D + 1$, $J = \log(eK)$, $\Lambda = \log(eH)$, and β, which the other appendices use without further comment.

The following table shows where each result of the main text is proved.

Main text	Appendix statement	Location
Theorem 4.1, upper bound	Theorem C.3, Lemma C.10	Appendix C
Theorem 4.1, lower bound	Theorem D.5	Appendix D
Theorem 5.1	Theorem C.3	Appendices C and E
Theorem 5.2	Theorem E.1	Appendix E
Theorem 5.3	Theorem F.3	Appendix F
Table 1	Proposition G.1	Appendix G

Corollary C.4 combines the two bounds of Theorem 4.1. Appendix H discusses further related work.

Two tools recur in the upper-bound proofs. Lemma C.6 is a pathwise identity. It separates the regret relative to the best remaining mean from the lasting loss caused by deletions. Lemma C.9 is a continuation bound. It controls AFE from the start of any elimination stage, so it covers both ordinary stages and all rounds after a wrong deletion. All lower bounds compare the distributions of the full transcript through Lemma D.1, so they hold for every protocol.

B FULL SPECIFICATION OF AFE

Algorithm 3 restates Algorithm 1 with the operations SAMPLE and EXPLOIT defined below, so that the proofs can refer to each step precisely. It adds the full localization schedule and the MEDIAN subroutine. AFE uses the parameters

$$H = NT, \qquad d = D + 1, \qquad J = \log(eK), \qquad \Lambda = \log(eH), \tag{14}$$

$$\beta = \min\{1, \sqrt{K/(NdJ)}\},$$

$$\Xi = d + \sqrt{Kd/N}\sqrt{J}\,\Lambda, \qquad \Upsilon = d + \beta d J^{3/2}. \tag{15}$$

We also write $\psi(x) = \max\{1, \log x\}$. All these quantities depend only on the known inputs, and AFE receives no mean or gap. The two quantities in (15) appear only in the analysis.

Every operation below stops at round T. An interrupted sampling block or message-passing step produces no estimate and removes no arm. EXPLOIT(i) means that every agent pulls arm i until round T. Ties between arms are broken in favor of the smaller label.

B.1 SAMPLING AND EXACT AGGREGATION

The operation SAMPLE$(S, u; A)$ estimates the arms in a set $S \subseteq A$, where A is the current candidate set. For u rounds, every agent pulls an arm drawn uniformly at random from S and records its reward sum for each arm. The agents then spend D more rounds passing these sums through the graph, and we call these rounds flooding. Each agent forwards every sum it receives, labeled by the operation number and by the agent that produced it, and drops duplicates. After these D rounds, every agent holds the same collection of sums from all N agents and computes

$$\widehat{\mu}_i = \frac{|S|}{Nu} \sum_{v=1}^{N} \sum_{s=1}^{u} \mathbf{1}\{A_v(s) = i\} Y_v(s), \quad i \in S. \tag{16}$$

The estimator divides by a known normalization instead of a random pull count. An agent that does not pull arm i contributes zero to its estimate. During flooding, every agent pulls the smallest-label arm in A, and these rewards are discarded. A message may carry $O(N|S|)$ real numbers and agent identifiers per edge and round. The model places no limit on message size, and we measure cost only in rounds.

Algorithm 2 Temporary median reference: MEDIAN(A, ε, p)

```
B ← A, q ← 0.
while |B| > 1 and rounds remain do
    θ_q ← (ε/4)(3/4)^q, ℓ_q ← log(12 · 2^q/p).
    u_q ← ⌈12|B|ℓ_q/(Nθ_q^2)⌉.
    Obtain estimates from SAMPLE(B, u_q; A).
    Retain the empirical top ⌈|B|/2⌉ arms in B; q ← q + 1.
end while
return the unique arm in B if the operation completes.
```

MEDIAN shrinks only the temporary set B and leaves A unchanged. It may discard the best arm of A from B, but that arm remains available to the elimination test. Its sampling blocks call SAMPLE$(B, u_q; A)$, so agents still pull arms from A while they wait for messages.

B.2 LOCALIZATION AND ELIMINATION

If $\beta < 1$, AFE first runs localization with a high confidence level. It starts with $A = [K]$ and sets

$$Q_0 = \lceil \log_2(1/\beta) \rceil, \qquad L_{\mathrm{loc}} = \log(4K(Q_0+1)T), \qquad e_j = 2^{-j}. \tag{17}$$

For $j = 0, \dots, Q_0$, localization stops if A is a singleton. Otherwise it obtains SAMPLE$(A, \lceil 48|A|L_{\mathrm{loc}}/(Ne_j^2) \rceil; A)$ and deletes the arms whose empirical gap to the empirical maximum exceeds $e_j/2$. The surviving set is A_0. If $\beta = 1$, AFE skips localization and sets $A_0 = [K]$. In both cases, AFE uses only the known inputs.

Algorithm 3 Adaptive Federated Elimination (AFE): full specification

Require: N, K, T, D, neighbor lists, agent IDs, common clock.

```
Compute (14); obtain A_0 by (17) if β < 1.
r ← 0; no previous reference is defined yet.
while rounds remain do
    a_r ← |A_r|, ε_r ← β2^{-r}.
    if a_r = 1 then
        EXPLOIT(the sole arm in A_r) and terminate.
    else if a_r ≥ NTε_r^2 then
        EXPLOIT(min A_0) if r = 0; otherwise EXPLOIT(J_{r-1}); terminate.
    else
        L_r ← log(128NTε_r^2/a_r), p_r ← e^{-L_r}.
        J_r ← MEDIAN(A_r, ε_r, p_r).
        With new data obtain SAMPLE(A_r, ⌈48a_r L_r/(Nε_r^2)⌉; A_r).
        A_{r+1} ← {i ∈ A_r : μ̂_{J_r} − μ̂_i ≤ ε_r}.
        r ← r + 1.
    end if
end while
```

Each candidate set contains the reference J_r and is therefore nonempty. In particular, the previous reference used at a stop still belongs to the current candidate set. All agents receive the same estimates, follow the same tie-breaking rule, and use the same operation lengths, so they stay synchronized without a central server. A stage runs only if $2 \le a_r < H\varepsilon_r^2$, which gives $L_r > \log 128$. A running stage therefore needs $\varepsilon_r > \sqrt{2/H}$, so AFE runs at most $O(\log(eH))$ elimination stages. The stopping rule uses only known parameters and observed quantities.

C UPPER BOUNDS FOR AFE

In conditional expectations, the history includes all past rewards and all randomness, although each agent observes only its own information and the messages it has received. Estimates of different arms need not be independent. The proofs first analyze each operation as if it ran to completion, extending it with fresh rewards when round T interrupts it. They never condition on an operation finishing before round T. AFE itself uses only the part of this ideal run that fits before round T, and Lemma C.6 shows that this part costs no more than the ideal run.

C.1 SAMPLING AND MEDIAN SELECTION

Lemma C.1 (Conditional concentration of the estimate)**.** *For an operation* SAMPLE$(S, u; A)$*, write* $a = |S|$*. Conditional on its start history, for each* $i \in S$*,*

$$\mathbb{E}\widehat{\mu}_i = \mu_i, \qquad \mathrm{Var}(\widehat{\mu}_i) \le \frac{a}{Nu}, \qquad \mathbb{P}\{|\widehat{\mu}_i - \mu_i| > z\} \le 2\exp\{-Nuz^2/(3a)\}, \quad 0 < z \le 1. \tag{18}$$

During a sampling round, the expected regret of each agent equals the average of the gaps Δ_i *over* $i \in S$*.*

Proof. Fix $i \in S$. The variables $X_{vs} = \mathbf{1}\{A_v(s) = i\}Y_v(s)$ are conditionally independent and lie in $[0, 1]$. Each has mean μ_{iv}/a and variance at most $1/a$. Summing their means and multiplying by $a/(Nu)$ gives $\mathbb{E}\widehat{\mu}_i = \mu_i$, and the same scaling gives the variance bound. Bernstein's inequality bounds the tail by $2\exp\{-Nuz^2/[a(2+2z/3)]\}$, which implies the displayed bound for $0 < z \le 1$. The last claim holds because each agent draws its arm uniformly from S, independently of the rewards in the current block. It conditions only on the history before the block, not on any later event. □

The estimate may fall outside $[0, 1]$, and AFE does not clip it. When a sampling length is $u = \lceil w \rceil$, all u rounds use uniform draws and enter the estimate. The proofs sometimes split their cost into a continuous part for w rounds and at most one extra round, but this split only simplifies the accounting.

Lemma C.2 (Median reference and its cost)**.** *Fix the start history of* MEDIAN(A, ε, p)*, with* $a = |A|$*,* $0 < \varepsilon \le 1$*, and* $0 < p \le 1/4$*. Let* $M = \max_{i \in A} \mu_i$ *and* $S = \sum_{i \in A}(M - \mu_i)$*. The output* J *satisfies*

$$\mathbb{P}(M - \mu_J > \varepsilon) \le p, \qquad \mathbb{E}(M - \mu_J) \le 4\varepsilon, \tag{19}$$

$$\mathbb{P}(M - \mu_J > x) \le p^{(x/\varepsilon)^2}, \qquad \varepsilon \le x \le 1. \tag{20}$$

If every arm in A *has gap at most* β*, the expected regret of each agent relative to* M *is at most*

$$\frac{C\log(12/p)}{N\varepsilon^2}(S + 2a\varepsilon) + \beta d\left\lceil \log_2 a \right\rceil. \tag{21}$$

The second term covers all rounds added by rounding and all flooding rounds.

Proof. The case $a = 1$ is immediate. At step q of MEDIAN, let B_q be the current temporary set, $a_q = |B_q|$, $M_q = \max_{i \in B_q} \mu_i$, and $Z_q = M_q - M_{q+1} \ge 0$. Suppose that $Z_q > x$ and that a best arm of B_q is not underestimated by more than $x/2$. Then at least $a_q/2$ of the retained arms are overestimated by more than $x/2$. Lemma C.1 and Markov's inequality, applied to the number of overestimated arms, give

$$\mathbb{P}(Z_q > x \mid \text{step history}) \le 6e^{-\ell_q x^2/\theta_q^2}, \qquad 0 < x \le 1. \tag{22}$$

This counting argument avoids a union bound over the a_q arms. Since $\sum_q \theta_q \le \varepsilon$ and $M - \mu_J = \sum_q Z_q$, a loss larger than x requires $Z_q > \theta_q x/\varepsilon$ for some q. Let $t = (x/\varepsilon)^2 \ge 1$. Events with a threshold above one are empty, because all gaps are at most one. A union bound over the steps gives

$$\mathbb{P}(M - \mu_J > x) \le \frac{6(p/12)^t}{1 - 2^{-t}} \le 12(p/12)^t \le p^t,$$

which proves (20). Its value at $x = \varepsilon$ gives the first bound in (19). Integrating the tail and using $p \le 1/4 < e^{-1}$ gives

$$\mathbb{E}(M - \mu_J) \le \varepsilon + \varepsilon \int_1^\infty e^{-t}\, dt < 2\varepsilon,$$

which proves the second bound in (19).

To bound the cost, fix the benchmark M at the initial maximum and let $G_q = a_q^{-1} \sum_{i \in B_q} (M - \mu_i)$. Let V be the empirical top $k = \lceil a_q/2 \rceil$ arms that MEDIAN retains, and let U be the true top k arms. Ordering gives

$$\begin{aligned} \frac{1}{k} \sum_{i \in V} (M - \mu_i) &\le \frac{1}{k} \sum_{i \in U} (M - \mu_i) + \frac{2}{k} \sum_{i \in B_q} |\widehat{\mu}_i - \mu_i| \\ &\le G_q + \frac{4}{a_q} \sum_{i \in B_q} |\widehat{\mu}_i - \mu_i|. \end{aligned}$$

The variance bound gives $\mathbb{E}|\widehat{\mu}_i - \mu_i| \le \theta_q / \sqrt{12 \ell_q}$, and consequently $\mathbb{E} G_q \le S/a + 2\varepsilon$. Given the initial size, $a_q = \lceil a/2^q \rceil$ is deterministic. The sampling lengths without rounding sum to at most

$$\sum_q \frac{12 a_q \ell_q}{N \theta_q^2} \le \frac{C a \log(12/p)}{N \varepsilon^2},$$

because $a_q \le 2a2^{-q}$ during execution and each step multiplies a_q/θ_q^2 by at most $\frac{1}{2}(\frac{4}{3})^2 = 8/9$, while ℓ_q grows only linearly in q. Multiplying these deterministic lengths by $\mathbb{E} G_q$ gives the first term of (21). Each operation adds at most one rounding round and D flooding rounds, that is, at most d rounds. In these rounds, agents pull arms from A, so each round costs at most β. There are at most $\lceil \log_2 a \rceil$ operations, which gives the second term. □

C.2 MAIN UPPER BOUNDS

Theorem C.3 (Simultaneous upper bounds for AFE)**.** *There is an absolute constant C such that Algorithm 3 satisfies, on every instance and for every agent v,*

$$\begin{aligned} \mathbb{E} R_v(T) &\le C\big(\sqrt{KT/N} + \Xi\big), \\ \mathbb{E} R_v(T) &\le C \left(\Upsilon + \frac{\Lambda}{N} \sum_{i \ne b} \frac{1}{\Delta_i} \right). \end{aligned} \tag{23}$$

The first bound gives the minimax rate, and the second is the instance-dependent bound of Theorem 5.1. Both include the costs of localization and of waiting for messages.

Corollary C.4 (Minimax rate on every graph)**.** *For every connected G and $2 \le N, K \le T$,*

$$\frac{1}{64} \left(D + \sqrt{KT/N} \right) \le \mathfrak{R}(G, K, T) \le C \left(D + \sqrt{KT/N} \right). \tag{24}$$

Appendix D proves the lower bound. Theorem C.3 and Lemma C.10 give the upper bound, and the condition $N, K \le T$ lets Lemma C.10 absorb the logarithmic remainder in Ξ. The same algorithm also satisfies the refined bound of Theorem E.1 in Appendix E.

C.3 LOCALIZATION

Lemma C.5 (Accuracy and cost of localization)**.** *Suppose $\beta < 1$. Couple the localization in* (17) *with its complete ideal run, using independent extra rewards whenever round T interrupts an operation. There is a good event of probability at least $1 - 1/(2T)$. On this event, the* completed output *A_0 contains b, and every arm in A_0 has gap at most β. We make no accuracy claim for an unfinished localization. On the good event, the cost of localization, or of any part of it that fits before round T, is at most*

$$\frac{384 K L_{\mathrm{loc}}}{N \beta} + 3d \tag{25}$$

on every trajectory. Alternatively, consider all localization steps that start before the first inaccurate estimate. Their expected sampling cost, without rounding, is at most

$$\frac{CL_{\mathrm{loc}}}{N}\sum_{i\neq b}\frac{\Delta_i}{\max\{\Delta_i,\beta\}^2}, \tag{26}$$

and rounding and flooding in these steps cost at most $3d$*. After the first inaccurate estimate, the remaining regret is at most* T*, so its expected contribution is at most* $1/2$*.*

Proof. At precision e_j, Lemma C.1 bounds the probability of each error larger than $e_j/4$ by $2e^{-L_{\mathrm{loc}}}$. A union bound over at most $K(Q_0+1)$ estimates bounds the failure probability by $1/(2T)$. On the good event, a best arm remains, and every arm that survives step j has gap at most e_j. At the start of a good step, all current gaps are therefore bounded by $h_0 = 1$ and $h_j = 2e_j$ for $j \geq 1$, and $\sum_j h_j \leq 3$. Multiplying the worst-case lengths $48KL_{\mathrm{loc}}/(Ne_j^2)+1+D$ by these h_j and summing gives (25), since $2^{Q_0} < 2/\beta$.

For (26), condition on the history at the start of a step, and restrict to the event, determined by that history, that all earlier estimates were accurate. A suboptimal arm i then survives at most until the first step with $e_j < \Delta_i$, or until step Q_0. Over these steps, its expected sampling regret without rounding is at most $48\Delta_i L_{\mathrm{loc}} N^{-1}\sum_j e_j^{-2}$. The last step in which it takes part has precision at least a constant times $\max\{\Delta_i,\beta\}$, so the geometric sum gives (26). The uniform draws in each step depend only on the history at its start. The first inaccurate step and all later rounds contribute at most T times the failure probability. The bound $\sum_j h_j \leq 3$ covers all rounding and flooding in the accurate steps. □

Since $K \geq \log(eK)$ and $d \leq N$, we have $\beta \geq 1/N$. Since $K \leq T < H$, we also have $L_{\mathrm{loc}} \leq C\Lambda$. AFE runs the elimination stages even after an inaccurate localization, and the good event serves only the analysis. All later samples are new.

C.4 Regret decomposition

The following identity holds on every trajectory. It separates the regret relative to the best remaining mean from the loss caused by drops in that mean, and it counts both over the original horizon of T rounds.

Lemma C.6 (Regret decomposition)**.** *Start from any history at the beginning of an elimination stage, with candidate set* A *and maximum* $M_0 = \max_{i\in A}\mu_i$*. Index the next* $h \leq T$ *rounds by* $\ell = 1,\dots,h$*. Let* $M(\ell)$ *be the maximum mean of the candidate set in effect when round* ℓ *is played. For every completed update* q *of the candidate set, let* τ_q *be the number of these rounds that have passed when the update takes effect, and let* $c_q \geq 0$ *be the resulting drop in the maximum. Then, on every trajectory,*

$$\sum_{\ell=1}^{h}(\mu_b - \mu_{A_v(\ell)}) = h(\mu_b - M_0) + \sum_{\ell=1}^{h}(M(\ell) - \mu_{A_v(\ell)}) + \sum_{q:\tau_q\leq h}(h-\tau_q)c_q. \tag{27}$$

Every summand on the right is nonnegative, and the first term vanishes when $b \in A$*. The last sum is at most* $T\sum_q c_q$*, and the middle sum covers sampling, rounding, flooding, and the final phase. If* $b \in A$ *and* b *is first deleted after round* $\sigma \leq h$*, and* M^+ *is the maximum after this deletion, then also*

$$\sum_{\ell=1}^{h}(\mu_b - \mu_{A_v(\ell)}) = \sum_{\ell=1}^{\sigma}(\mu_b - \mu_{A_v(\ell)}) + (h-\sigma)(\mu_b - M^+) + \sum_{\ell=\sigma+1}^{h}(M^+ - \mu_{A_v(\ell)}). \tag{28}$$

A continuation bound relative to M^+ *therefore already includes every later drop within that continuation.*

Proof. Every pull belongs to the current candidate set. This includes the pulls made while MEDIAN samples its temporary sets, and all pulls during flooding and in the final phase. The candidate sets are nested. Hence all deficits relative to the current maximum and all drops are nonnegative, and

$M_0 - M(\ell) = \sum_{q:\tau_q<\ell} c_q$. Substituting this telescoping relation into the gap of each round and exchanging the two finite sums gives (27). Equation (28) follows by adding and subtracting M^+ over the rounds after σ. An interrupted operation makes no update of the candidate set. Its partial cost relative to the current maximum is at most the cost of its complete ideal run, and extending the nested sets can only increase the accumulated drop. Hence the part of a run that fits before round T never costs more than the ideal run, and the estimation bounds need not condition on an operation finishing in time. □

C.5 Validation and candidate counts

Consider a running elimination stage with candidate set A, size a, maximum M, precision ε, and $L = \log(128H\varepsilon^2/a)$. Let A^+ be the candidate set after validation, and let $M^+ = \max_{i\in A^+} \mu_i$.

Lemma C.7 (Survival of suboptimal arms and loss of the best arm)**.** *For $g_i = M - \mu_i > 4\varepsilon$,*

$$\mathbb{P}(i \in A^+ \mid \textit{stage history}) \le 5e^{-Lg_i^2/(16\varepsilon^2)}. \tag{29}$$

Let b_A be a best arm in A and $\delta_J = M - \mu_J$. Then

$$\mathbb{P}(b_A \notin A^+ \mid \textit{history}, J) \le 4e^{-4L(1+\delta_J/\varepsilon)^2}, \tag{30}$$

$$\mathbb{E}(M - M^+ \mid \textit{history}) \le \varepsilon e^{-L}. \tag{31}$$

For $Z = \max_{i\in A^+}(M - \mu_i)$,

$$\mathbb{E}(Z^2 \mid \textit{history}) \le 16\varepsilon^2 \left[1 + \frac{\log(5a) + 1}{L}\right]. \tag{32}$$

Proof. Suppose that $\delta_J \le g_i/4$ and that the new estimation errors of J and i are both at most $g_i/4$. Then $\widehat{\mu}_J - \widehat{\mu}_i \ge g_i/4 > \varepsilon$, so arm i fails the test. Lemma C.2 bounds the probability that $\delta_J > g_i/4$ by $e^{-Lg_i^2/(16\varepsilon^2)}$, and Lemma C.1 bounds the probability of either large error by $4e^{-Lg_i^2/\varepsilon^2}$ in total. This proves (29).

The arm b_A is deleted only if the new error of J minus the new error of b_A exceeds $\delta_J + \varepsilon$. At least one of these errors then has magnitude larger than $(\delta_J + \varepsilon)/2 \le 1$, which gives (30). Because J always survives, $0 \le M - M^+ \le \delta_J \mathbf{1}\{b_A \notin A^+\}$. For $x \ge 0$ and $L \ge 1$, the function $4xe^{-4L(1+x)^2}$ is at most e^{-L}. Setting $x = \delta_J/\varepsilon$ proves (31).

Only the moment bound (32) uses a union bound over arms, and the sampling cost in the minimax analysis never does. For $x \ge 4\varepsilon$, (29) gives $\mathbb{P}(Z > x) \le \min\{1, 5ae^{-Lx^2/(16\varepsilon^2)}\}$. Integrating the tail of Z^2 and extending the integral gives

$$\mathbb{E}Z^2 \le 16\varepsilon^2 + \int_0^\infty \min\{1, 5ae^{-Lu/(16\varepsilon^2)}\}\, du,$$

which equals the stated bound. If $4\varepsilon > 1$, the same bound follows directly from $Z \le 1$. □

Lemma C.8 (Control of random candidate counts)**.** *Let $a^+ = |A^+|$, $\varepsilon^+ = \varepsilon/2$, and $S^+ = \sum_{i\in A^+}(M^+ - \mu_i)$. Define*

$$L^+(z) = \log\left[128 \max\{1, H(\varepsilon^+)^2/z\}\right], \quad z > 0,$$

even if the next stage stops. Conditional on the history at the start of the current running stage,

$$\mathbb{E}\left[L^+(a^+)(S^+ + 2a^+\varepsilon^+)\right] \le Ca\varepsilon L. \tag{33}$$

Proof. All expectations below condition on the history at the start of the stage. Write $g_i = M - \mu_i$ and $Q = \sum_{i\in A^+} g_i^2$. An arm with $g_i \le 4\varepsilon$ contributes at most $16\varepsilon^2$ to $\mathbb{E}Q$. For an arm with $g_i > 4\varepsilon$, put $u = g_i^2/(16\varepsilon^2) > 1$. Then (29) gives

$$g_i^2 \mathbb{P}(i \in A^+) \le 80\varepsilon^2 ue^{-Lu} \le 80\varepsilon^2 e^{-L} \le 16\varepsilon^2,$$

because $L > \log 128$ and ue^{-Lu} decreases for $u \ge 1$. Consequently $\mathbb{E}Q \le 16a\varepsilon^2$.

For $0 < z \le a$, we have $L^+(z) \le L + \log(a/z)$. Since $L > 2$, the function $\sqrt{z}[L + \log(a/z)]$ increases on this interval. Hence $\sqrt{z}L^+(z) \le \sqrt{a}L$ and $zL^+(z) \le aL$. Since $M^+ \le M$, the Cauchy–Schwarz inequality gives $S^+ \le \sum_{i\in A^+} g_i \le \sqrt{a^+Q}$, so

$$\mathbb{E}[L^+(a^+)S^+] \le \sqrt{a}L\,\mathbb{E}\sqrt{Q} \le \sqrt{a}L\sqrt{\mathbb{E}Q} \le 4a\varepsilon L.$$

Finally, $2\varepsilon^+ = \varepsilon$ gives $2a^+\varepsilon^+L^+(a^+) \le a\varepsilon L$ on every trajectory. Adding the two bounds proves (33). □

C.6 Continuation bound

The next lemma controls AFE from the start of any elimination stage, with its original parameters and the original deadline T. It serves both the minimax analysis below and, in Appendix E, the analysis after the first deletion of the best arm.

Lemma C.9 (Continuation from any elimination stage)**.** *Fix any history at the start of an elimination stage, with candidate set B, size $a \ge 2$, precision ε, and $a < H\varepsilon^2$. Let*

$$M = \max_{i\in B}\mu_i, \quad S = \sum_{i\in B}(M-\mu_i), \quad z = \max_{i\in B}(M-\mu_i), \quad L = \log(128H\varepsilon^2/a).$$

For every number of remaining rounds up to T, the conditional expected regret relative to M is at most

$$\frac{CL}{N\varepsilon^2}(S + 2a\varepsilon) + \frac{C\sqrt{aH}}{N} + \mathbb{E}\mathcal{O}. \tag{34}$$

Here $\mathcal{O}$ is the cost of rounding and flooding, measured at each stage relative to the maximum of that stage's candidate set, and

$$\mathbb{E}\mathcal{O} \le CdJ[z + \varepsilon\sqrt{J}]. \tag{35}$$

All future samples are new given the stated history. The bound assumes no separation between the means of the remaining arms.

Proof. Condition on the stated history, and extend the run with independent new rewards until AFE stops. Index the stages from the given one by $r \ge 0$, with $\varepsilon_r = \varepsilon 2^{-r}$, candidate count $a_r \le a$, maximum M_r, gap sum S_r, and confidence factor L_r while the stage runs. Lemma C.2 and the validation step give the first stage a sampling cost, without rounding, of at most $CL(S+2a\varepsilon)/(N\varepsilon^2)$ relative to M. For each later stage, Lemma C.8 gives

$$\mathbb{E}[\text{sampling cost of stage } r] \le \frac{C}{N}\mathbb{E}\left[\mathbf{1}\{r-1 \text{ runs}\}\frac{a_{r-1}L_{r-1}}{\varepsilon_{r-1}}\right]. \tag{36}$$

Each sampling cost is measured relative to the maximum of that stage's candidate set. On histories where the next stage stops, the right side contains extra nonnegative terms, which only weaken the bound. Put $\eta_a = \sqrt{a/H}$. On every history where stage r runs,

$$a_rL_r \le U_a(\varepsilon_r), \qquad U_a(u) := \begin{cases} a\log(128Hu^2/a), & u \ge \eta_a, \\ (\log 128)Hu^2, & u < \eta_a. \end{cases} \tag{37}$$

This holds because $w\log(128Hu^2/w)$ increases in w for $w < Hu^2$. Summing over the geometric grid of precisions on both sides of η_a gives

$$\sum_{\varepsilon_r \ge \eta_a}\frac{U_a(\varepsilon_r)}{\varepsilon_r} \le Ca/\eta_a, \qquad \sum_{\varepsilon_r < \eta_a}\frac{U_a(\varepsilon_r)}{\varepsilon_r} \le CH\eta_a. \tag{38}$$

In the first sum, the logarithm grows linearly while the factor $1/\varepsilon_r$ decreases geometrically. The second sum is geometric. All later sampling costs therefore sum to at most $C\sqrt{aH}/N$.

By (31), the expected drop in the maximum during a running stage is at most

$$\varepsilon_r e^{-L_r} = \frac{a_r}{128H\varepsilon_r} \le \frac{1}{128}\min\{a/(H\varepsilon_r), \varepsilon_r\}.$$

The same split at η_a bounds the total expected drop by $C\eta_a$. The regret decomposition (27), with initial maximum M, charges each drop over at most T rounds, so the drops cost at most $CT\eta_a = C\sqrt{aH}/N$.

Suppose AFE later stops because the candidate count reaches the threshold. The last running stage r then has $\varepsilon_r \le 2\eta_a$, and all agents pull its reference J_r. Dropping the indicator of stopping and using (19) gives

$$T \sum_{r:\varepsilon_r \le 2\eta_a} \mathbb{E}[\mathbf{1}\{r \text{ runs}\}(M_r - \mu_{J_r})] \le 4T \sum_{r:\varepsilon_r \le 2\eta_a} \varepsilon_r \le CT\eta_a.$$

A stop with a single arm left costs nothing relative to the current maximum. The given stage runs, so no stop occurs before it.

For rounding and flooding, the largest deficit in the first stage is z. For each later stage, drop the indicator that it runs and apply (32) given the preceding history. The largest deficit of the next candidate set, relative to its own maximum, then has expectation at most $C\varepsilon_r\sqrt{J}$, even if earlier maxima have been deleted. Each stage has at most CJ sampling operations and at most d extra rounds per operation. Since $\sum_r \varepsilon_r \le 2\varepsilon$, these costs give (35). The drops of the maximum during these rounds are already charged through the factor T above.

Finally, by Lemma C.6, the part of the run that fits before round T has no larger stage costs or accumulated drops than this ideal continuation. None of the bounds conditions on an operation finishing before round T. They therefore also cover interrupted operations, and the extra rewards used in the ideal run are never available to AFE itself. □

C.7 Proof of the minimax upper bound

Fix any history at the end of a successful localization, with candidate set A_0 of size a_0. Every arm in A_0 has gap at most β, and $b \in A_0$, so regret relative to the maximum of A_0 is the actual regret. If $a_0 = 1$, no further regret occurs. If AFE stops at once because $a_0 \ge H\beta^2$, every later pull costs at most β, and $T\beta \le \sqrt{a_0 H}/N$.

Otherwise $a_0 < H\beta^2$, and we apply Lemma C.9 with $S \le a_0\beta$ and $z \le \beta$. Put $t = \beta\sqrt{H/a_0}$, so that $t > 1$. The first sampling term is bounded by

$$\frac{Ca_0}{N\beta}\log(128H\beta^2/a_0) = \frac{C\sqrt{a_0 H}}{N}\frac{\log(128t^2)}{t} \le \frac{C\sqrt{a_0 H}}{N},$$

because $\log(128t^2)/t$ decreases for $t \ge 1$ and equals $\log 128$ at $t = 1$. In every case, the expected regret after localization is therefore at most $C\sqrt{KT/N} + C\beta d J^{3/2}$ over the remaining rounds.

If $\beta < 1$, Lemma C.5 bounds the cost of localization, or of any unfinished part of it, by $CK\Lambda/(N\beta) + 3d$, and it adds at most $1/2$ in expectation after an inaccurate localization. The sampling term equals $C\sqrt{Kd/N}\sqrt{J}\,\Lambda$. If $\beta = 1$, there is no localization. In both cases, $\beta \le \sqrt{K/(NdJ)}$ and $J \le \Lambda$ give

$$\beta d J^{3/2} \le \sqrt{Kd/N}\,J \le \sqrt{Kd/N}\sqrt{J}\,\Lambda.$$

Adding all localization costs proves the bound $C(\sqrt{KT/N} + \Xi)$ of Theorem C.3.

Lemma C.10 (Absorbing the logarithmic remainder)**.** *For $1 \le K \le T$ and $1 \le D \le N$,*

$$\frac{\sqrt{KD/N}}{D + \sqrt{KT/N}} \le T^{-1/6}.$$

Consequently, for every fixed $q > 0$ and $N, K \le T$,

$$\sqrt{KD/N}\,[\log(eNT)]^q \le C_q(D + \sqrt{KT/N}).$$

Proof. By the AM–GM inequality, the denominator is at least $2\sqrt{D}\,(KT/N)^{1/4}$, so the ratio is at most $\frac{1}{2}(K/(NT))^{1/4} \le \frac{1}{2}N^{-1/4}$. Keeping only the term $\sqrt{KT/N}$ in the denominator bounds the

ratio by $\sqrt{D/T} \le \sqrt{N/T}$. Multiplying the square of the first bound by the second bounds the cube of the ratio by $\frac{1}{4}T^{-1/2}$, which proves the first claim. For the second claim, use $NT \le T^2$ and $\sup_{x\ge 1} x^{-1/6}[\log(ex)]^q < \infty$. Taking $q = 3/2$ and using $d \le 2D$ and $J \le \Lambda$ gives the upper bound of Corollary C.4. □

C.8 Instance-dependent part

Fix any history at the end of a successful localization, so every arm in A_0 has gap at most β. In the ideal continuation, let E_r be the event that stage r runs and b is still in A_r. For $i \ne b$, set $q_{ir} = \mathbb{P}(E_r,\ i \in A_r)$. Both events are determined at the start of stage r. After AFE stops, all these probabilities are zero.

Lemma C.11 (Survival while the best arm remains)**.** *The probabilities satisfy $q_{i,r+1} \le q_{ir}$. Moreover, whenever $\Delta_i > 3\varepsilon_r$, they satisfy $q_{i,r+1} \le \rho q_{ir}$ with $\rho = 5/128 < 1/16$.*

Proof. The monotonicity holds because both events can only shrink from one stage to the next. Suppose E_r holds, so $b \in A_r$ and the maximum of A_r is μ_b. If the reference is within ε_r of μ_b and the two new estimation errors are at most $\varepsilon_r/4$, then an arm with $\Delta_i > 3\varepsilon_r$ fails the test, as shown in Section 3. The first condition fails with probability at most p_r, and the second with probability at most $4p_r$. Since $p_r < 1/128$, such an arm survives with probability at most $5/128$. □

Appendix E uses this lemma to control all later rounds, including every round after the first deletion of b. It proves the bound (60) after a successful localization. Appendix E.4 then derives the second bound of Theorem C.3 from this bound and Lemma C.5, so the two guarantees rest on the same regret accounting.

D Lower bounds

This appendix proves lower bounds that hold for every protocol. Protocols may use any shared or private randomness that does not depend on the instance, and they know the horizon. They do not know the means, the gaps, or the best arm. For each arm i and agent v, let

$$n_{iv}(T) = \sum_{t=1}^{T} \mathbf{1}\{A_v(t) = i\}, \qquad n_i(T) = \sum_{v=1}^{N} n_{iv}(T)$$

count the pulls of arm i by agent v and by all agents. The total expected regret equals $N\overline{\mathcal{R}}_\nu^\pi(T)$, and $\overline{\mathcal{R}}_\nu^\pi(T) \le \mathcal{R}_{\max\nu}^\pi(T)$ by Appendix A. We use the adaptive change-of-measure method (Kaufmann et al., 2016; Lattimore & Szepesvári, 2020) and keep the horizon T and the agent-level quantifiers explicit.

D.1 Transcript information

For $x, y \in (0,1)$, let

$$\mathrm{kl}(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$$

be the divergence between Bernoulli distributions. We extend kl to $[0,1]^2$ in the usual way, allowing the value $+\infty$.

Lemma D.1 (Transcript change of measure)**.** *Let $p, q \in (0,1)^{K\times N}$ be two matrices of Bernoulli means, and let $\mathbb{P}_p^T$ and $\mathbb{P}_q^T$ be the distributions of the full transcript of the same protocol through round T. Then*

$$\mathrm{KL}(\mathbb{P}_p^T, \mathbb{P}_q^T) = \sum_{i=1}^{K}\sum_{v=1}^{N} \mathbb{E}_p n_{iv}(T)\,\mathrm{kl}(p_{iv}, q_{iv}). \tag{39}$$

In particular, for every event E determined by the transcript,

$$\sum_{i,v} \mathbb{E}_p n_{iv}(T)\,\mathrm{kl}(p_{iv}, q_{iv}) \ge \mathrm{kl}\big(\mathbb{P}_p(E), \mathbb{P}_q(E)\big). \tag{40}$$

Proof. The transcript contains all random seeds, actions, observed rewards, and messages. Given the past, the protocol chooses actions and messages by the same kernels under both instances, so these choices contribute no divergence. The seeds have the same distribution under both instances, since this distribution does not depend on the instance. Within each round, order the rewards by agent ID. The actions of that round are already fixed, and the rewards are conditionally independent, so their conditional divergences add. The chain rule for KL divergence then gives (39). Applying the data-processing inequality to the map that returns $\mathbf{1}_E$ proves (40). The ordering within a round is used only in the analysis and gives the protocol no feedback within a round. □

To turn transcript information into regret lower bounds, we use two standard inequalities. For probability measures P and Q and an event E, the Bretagnolle–Huber and Pinsker inequalities give

$$P(E) + Q(E^c) \geq \tfrac{1}{2} \exp\{-\operatorname{KL}(P, Q)\}, \qquad \|P - Q\|_{\mathrm{TV}} \leq \sqrt{\operatorname{KL}(P, Q)/2}. \tag{41}$$

D.2 Minimax lower bound on every graph

We use the instance class $\mathcal{V}(G, K)$ and the minimax regret $\mathfrak{R}(G, K, T)$ of Section 2. The class contains all independent bounded rewards with arbitrary heterogeneity and a unique best arm, and the infimum in (2) ranges over all protocols on G.

Lemma D.2 (Statistical lower bound)**.** *For every connected graph and $2 \leq N, K \leq T$,*

$$\mathfrak{R}(G, K, T) \geq \frac{1}{32} \sqrt{\frac{KT}{N}}. \tag{42}$$

Proof. Fix any protocol and put $H = NT$. Under the reference instance in which every arm has Bernoulli mean $1/2$ at every agent, choose an arm i with $\mathbb{E}_0 n_i(T) \leq H/K$. The alternative instance i raises the mean of this arm at every agent to $1/2 + \alpha$, where

$$\alpha = \sqrt{K/(64H)} < 1/4.$$

Lemma D.1 applies directly to the protocol. Since $\operatorname{kl}(1/2, 1/2 + \alpha) \leq (16/3)\alpha^2$,

$$\operatorname{KL}(\mathbb{P}_0^T, \mathbb{P}_i^T) \leq \frac{16\alpha^2 H}{3K} \leq \frac{1}{12}.$$

The count $n_i(T)$ lies in $[0, H]$, so Pinsker's inequality gives

$$\mathbb{E}_i n_i(T) \leq \mathbb{E}_0 n_i(T) + H/\sqrt{24} \leq H/K + H/\sqrt{24}.$$

The alternative instance has a unique best arm, and

$$\mathcal{R}_{\max_i}^{\pi}(T) \geq \overline{\mathcal{R}}_i^{\pi}(T) = \frac{\alpha}{N}\big(H - \mathbb{E}_i n_i(T)\big) \geq \frac{\alpha H}{4N} = \frac{1}{32}\sqrt{KT/N},$$

where $K \geq 2$ gives $1 - 1/K - 1/\sqrt{24} \geq 1/4$. The reference instance has no unique best arm, but it serves only for the change of measure, and the alternative belongs to the class. Since the protocol was arbitrary, the minimax bound follows. □

Lemma D.3 (Causal indistinguishability)**.** *Suppose two instances coincide outside a vertex set S. Let r be an integer with $1 \leq r \leq D$. If $\operatorname{dist}_G(v, S) \geq r$, then $(A_v(1), \ldots, A_v(r))$ has the same distribution under the two instances.*

Proof. Couple all randomness that does not depend on the instance, and couple the rewards outside S so that they agree whenever the same arm is pulled at the same agent. The first difference between the two instances can appear only in a reward observed at an agent in S, at the end of round 1 or later. Its effect crosses at most one edge per round. Even if that reward is sent at once, its effect reaches an agent at distance r only at the end of round r, so it can first influence the action of that agent in round $r + 1$. Induction over rounds gives the claimed equality for the actions of v. The argument allows messages of any size and content. □

Lemma D.4 (Diameter lower bound)**.** *For every connected graph of diameter $D \geq 1$,*

$$\mathfrak{R}(G, K, T) \geq \frac{D}{8}. \tag{43}$$

Proof. Let a and z be two vertices at distance D, and set $r = \lceil D/2 \rceil$. The balls of radius $r-1$ around a and z are disjoint because $2(r-1) < D$, so one of them contains at most $N/2$ vertices. Let v be its center, and let $S = \{u : \mathrm{dist}_G(u, v) \geq r\}$. Then $q = |S|/N \geq 1/2$. Put $\lambda = 1/(2q) \leq 1$.

We build two Bernoulli instances indexed by $\sigma \in \{+1, -1\}$. At every vertex outside S, arms 1 and 2 both have mean $1/2$. At $u \in S$, their means are

$$p^{\sigma}_{1u} = (1 + \sigma\lambda)/2, \qquad p^{\sigma}_{2u} = (1 - \sigma\lambda)/2.$$

Every arm $i \geq 3$ gives reward zero at every vertex. Independent rewards with these means belong to the class. The global means of arms 1 and 2 are $3/4$ and $1/4$, and their order reverses with σ. Each instance therefore has a unique best arm, and every other arm has gap at least $1/2$.

Since $r \leq D < T$, Lemma D.3 shows that, for each $t \leq r$, the probabilities $p_i(t)$ that v pulls arm i in round t are the same under both signs. The two expected regrets of v in round t sum to

$$\tfrac{1}{2}[p_1(t) + p_2(t)] + \tfrac{3}{2}\sum_{i=3}^{K} p_i(t) \geq \tfrac{1}{2}.$$

Summing over the first r rounds shows that at least one sign gives $\mathbb{E}R_v(T) \geq r/4 \geq D/8$. Since $\mathcal{R}_{\max}$ is at least the expected regret of v, this proves (43). □

Theorem D.5 (Minimax lower bound on every graph)**.** *For every connected graph G of diameter D and $2 \leq N, K \leq T$,*

$$\inf_{\pi} \sup_{\nu \in \mathcal{V}(G,K)} \mathcal{R}_{\max}{}^{\pi}_{\nu}(T) \geq \frac{1}{64}\left(D + \sqrt{\frac{KT}{N}}\right). \tag{44}$$

Proof. Lemmas D.2 and D.4 bound the same minimax regret through the statistical cost and the cost of waiting. The larger of the two bounds is at least half their sum, which gives (44). □

Remark D.6 (Scope of the minimax statement). The two terms may come from different instances. Their sum is therefore a minimax statement, not a lower bound for each fixed instance. The diameter construction chooses one agent and one sign, and it does not force every agent to incur regret of order D on a common instance. No consistency assumption is used.

E Instance-dependent bounds for AFE

This appendix completes the instance-dependent bound of Theorem C.3 and proves Theorem E.1 below, which restates Theorem 5.2. Both proofs control how many suboptimal arms survive at each precision, and both use the same regret accounting, so we prove them together. The square-weighted count below comes from the analysis of single-agent bandits (Lattimore, 2015). Our analysis adds the cost of waiting for messages and the cost of localization.

For each suboptimal arm, define

$$k_i^{(2)} = \sum_{j=1}^{K} \min\{1, (\Delta_i/\Delta_j)^2\}, \qquad H_i^{(2)} = \frac{k_i^{(2)}}{\Delta_i^2}, \qquad F_2 = \frac{1}{N}\sum_{i \neq b} \frac{\psi(H/H_i^{(2)})}{\Delta_i}, \tag{45}$$

where the term for $j = b$ equals one since $\Delta_b = 0$. The quantities H, d, J, and ψ are those of Appendix B. We also write $\log_+(x) = \max\{0, \log x\}$ for $x > 0$ and define

$$\mathcal{B}^{(2)}_{\mathrm{loc}} = \begin{cases} 0, & \beta = 1, \\ \dfrac{K}{N\beta}\big[1 + \log K + \log(4dJ(Q_0+1))\big], & \beta < 1, \end{cases} \qquad Q_0 = \lceil \log_2(1/\beta) \rceil. \tag{46}$$

This localization cost depends on N, K, and G, but not on T.

Theorem E.1 (Effective-competition upper bound)**.** *There is an absolute constant C such that Algorithm 3 satisfies, for every instance and every agent v,*

$$\mathbb{E}R_v(T) \leq C\big(F_2 + \Upsilon + \mathcal{B}^{(2)}_{\mathrm{loc}}\big), \qquad \Upsilon = d + \beta d J^{3/2}. \tag{47}$$

The bounds of Theorem C.3 hold for the same run, so the minimax bound $C(D + \sqrt{KT/N})$ also holds.

Our comparisons with other bounds keep the localization cost $\mathcal{B}_{\mathrm{loc}}^{(2)}$ in (47). The theorem controls regret up to absolute constants. It does not give the exact KL-based constants of asymptotic lower bounds for heterogeneous agents, and it does not reach the sharper count with linear gap ratios.

E.1 Two deterministic comparisons

Sort the $k = K - 1$ suboptimal gaps in nondecreasing order, break ties arbitrarily, and index the arms by $i = 1, \dots, k$. Let $r_i = i + 1$ be the rank of arm i when the best arm is counted, and let $u_i = 1/\Delta_i$, so that $u_1 \geq \cdots \geq u_k > 0$.

Lemma E.2 (Rank comparison)**.** *We have*

$$\sum_{i=1}^{k} \frac{1}{\Delta_i} \log \frac{k_i^{(2)}}{r_i} \leq 4 \sum_{i=1}^{k} \frac{1}{\Delta_i}. \tag{48}$$

Consequently,

$$F_2 \leq F_{\mathrm{rank}} \coloneqq \frac{1}{N} \sum_{i=1}^{k} \frac{\psi(H\Delta_i^2/r_i)}{\Delta_i} \leq 5F_2. \tag{49}$$

Proof. After sorting, $k_i^{(2)} = r_i + u_i^{-2} \sum_{j>i} u_j^2$. Using $\log(1+x) \leq 2\sqrt{x}$ and $r_i \geq i$, we bound the left side of (48) by

$$2 \sum_{i=1}^{k} \frac{1}{\sqrt{i}} \left(\sum_{j>i} u_j^2 \right)^{1/2}.$$

Let $u_{k+1} = 0$ and $v_m = u_m - u_{m+1} \geq 0$, so that $u_j = \sum_{m \geq j} v_m$. Minkowski's inequality for the tail vector gives

$$\left(\sum_{j>i} u_j^2 \right)^{1/2} \leq \sum_{m>i} v_m \sqrt{m-i}.$$

We exchange the two finite nonnegative sums and use

$$\sum_{i<m} \frac{\sqrt{m-i}}{\sqrt{i}} \leq \sqrt{m} \sum_{i \leq m} i^{-1/2} \leq 2m.$$

This bounds the sum above by $4 \sum_m m v_m = 4 \sum_i u_i$, which proves (48). Finally, $r_i \leq k_i^{(2)}$, the inequality $\psi(xe^c) \leq \psi(x) + c$ for $c \geq 0$, and $\psi \geq 1$ give (49). □

For any fixed $\beta > 0$, partition all suboptimal arms into the dyadic bins

$$I_g = \{i : \delta_g/2 < \Delta_i \leq \delta_g\}, \qquad \delta_g = \beta 2^{-g}, \qquad n_g = |I_g|, \qquad g \in \mathbb{Z}.$$

Only finitely many bins are nonempty. Negative indices cover gaps larger than β in this deterministic comparison.

Lemma E.3 (Bin comparison)**.** *There is an absolute constant C such that*

$$\frac{1}{N} \sum_{g : n_g > 0} \frac{n_g}{\delta_g} \psi\left(\frac{H\delta_g^2}{n_g} \right) \leq CF_2. \tag{50}$$

Proof. For $i \in I_g$, we have $\Delta_i \leq \delta_g < 2\Delta_i$. Splitting a bin into its arms therefore bounds its contribution by

$$\sum_{i \in I_g} \frac{1}{\Delta_i} [\psi(H\Delta_i^2/n_g) + \log 4].$$

Let s_g be one plus the number of suboptimal arms in bins with strictly smaller gaps. Every rank in bin g is at most $R_g = s_g + n_g$. The inequality

$$\psi(H\Delta_i^2/n_g) \leq \psi(H\Delta_i^2/r_i) + \log_+(r_i/n_g)$$

shows that the extra logarithmic weight of the bin is at most

$$\frac{2n_g}{\delta_g}\log(1+s_g/n_g) \le \frac{2s_g}{\delta_g}.$$

Let $\delta_{\min}$ be the smallest δ_g among nonempty bins. Exchanging the finite sums over bins gives

$$\sum_g \frac{s_g}{\delta_g} = \sum_g \frac{1}{\delta_g} + \sum_h n_h \sum_{g:\delta_g>\delta_h} \frac{1}{\delta_g} \le \frac{2}{\delta_{\min}} + \sum_h \frac{n_h}{\delta_h} \le 3\sum_g \frac{n_g}{\delta_g} \le 3\sum_{i\neq b}\frac{1}{\Delta_i}.$$

Here the geometric sums run over nonempty dyadic bins, and the bin with $\delta_g = \delta_{\min}$ contains at least one arm. Lemma E.2 and $\psi \ge 1$ pay for this term and for the $\log 4$ terms, which proves (50). □

E.2 Sampling work while the best arm remains

Fix any history at the end of a successful localization. Then A_0 contains b, and every arm in A_0 has gap at most β, although A_0 need not contain every arm with a small gap. As in Appendix C.8, let E_r be the event that stage r runs and $b \in A_r$, and let $q_{ir} = \mathbb{P}(E_r, i \in A_r)$ for $i \neq b$. These probabilities are zero after AFE stops. Lemma C.11 gives

$$q_{i,r+1} \le q_{ir}, \qquad q_{i,r+1} \le \rho q_{ir} \quad \text{if } \Delta_i > 3\varepsilon_r, \qquad \rho = 5/128 < 1/16. \tag{51}$$

On E_r, let $S_r = \sum_{i\in A_r}\Delta_i$. The sampling work done while b remains is

$$\mathcal{W} = \frac{1}{N}\sum_{r\ge 0}\mathbb{E}\left[\mathbf{1}_{E_r}\frac{L_r}{\varepsilon_r^2}(S_r + 2a_r\varepsilon_r)\right], \tag{52}$$

where each product is zero on histories in which stage r does not run, since L_r is undefined there.

Lemma E.4 (Sampling work by gap bins)**.** *For every history at the end of a successful localization, $\mathcal{W} \le CF_2$ with an absolute constant C.*

Proof. For $g \ge 0$, every arm in I_g has $\Delta_i > \delta_g/2 > 3\varepsilon_r$ whenever $r \ge g+3$. Set $Z_{gr} = \mathbf{1}_{E_r}|A_r \cap I_g|$ and $t_{gr} = \rho^{(r-g-3)_+}$. Then (51) implies

$$\mathbb{E}Z_{gr} \le n_g t_{gr}. \tag{53}$$

The count n_g includes every arm of the original bin, even those that localization removed, so it remains a valid upper bound.

On E_r, we have $a_r \ge 2$ and $b \in A_r$, so $a_r \le 2\sum_{i\neq b}\mathbf{1}\{i \in A_r\}$ and

$$S_r + 2a_r\varepsilon_r \le \sum_{i\neq b,\, i\in A_r}(\Delta_i + 4\varepsilon_r).$$

With $h = H\varepsilon_r^2$, consider the increasing concave function

$$\phi_h(0) = 0, \qquad \phi_h(z) = z\log[128(1+h/z)], \quad z > 0.$$

When $Z_{gr} > 0$, we have $a_r \ge Z_{gr}$. Jensen's inequality and (53) therefore give

$$\mathbb{E}[Z_{gr}L_r] \le \mathbb{E}\phi_h(Z_{gr}) \le \phi_h(\mathbb{E}Z_{gr}) \le \phi_h(n_g t_{gr}). \tag{54}$$

Concavity thus handles the dependence between survival indicators without fixing the random candidate count. Combining these bounds gives

$$\mathcal{W} \le \frac{1}{N}\sum_{g\ge 0:n_g>0}\sum_{r\ge 0}\frac{\delta_g + 4\varepsilon_r}{\varepsilon_r^2}n_g t_{gr}\log\left[128\left(1+\frac{H\varepsilon_r^2}{n_g t_{gr}}\right)\right]. \tag{55}$$

All terms on the right are positive, so extending the series beyond the finite stopping time keeps the upper bound valid.

Fix a bin and write $\delta = \delta_g$, $n = n_g$, and $B = \log[128(1+H\delta^2/n)]$. For $r \le g+3$, put $v = g+3-r \ge 0$. Then $\varepsilon_r = \delta 2^v/8$ and $t_{gr} = 1$. The logarithm in (55) is at most $B + v\log 4$, and

the two precision weights are $64/(\delta 4^v)$ and $32/(\delta 2^v)$. Summing these geometric series, together with their versions weighted by v, gives at most CnB/δ.

For $r = g + 3 + t$ with $t \geq 1$, we have $\varepsilon_r = \delta/(8 \cdot 2^t)$ and $t_{gr} = \rho^t$. The logarithm is at most $B + t\log(1/(4\rho))$. After multiplication by the survival factor, the precision weights are $64(4\rho)^t/\delta$ and $32(2\rho)^t/\delta$. Both ratios 4ρ and 2ρ are smaller than one, so this part also sums to at most CnB/δ. Finally, $\log[128(1+x)] \leq C\psi(x)$ for every $x > 0$. The right side of (55) is therefore at most a constant times the left side of (50), and Lemma E.3 completes the proof. □

E.3 Regret after the first deletion

Fix a history at the start of stage r on which E_r holds, and write $a = a_r$, $\varepsilon = \varepsilon_r$, $L = L_r$, and $p = e^{-L}$. Let F be the event that stage r deletes b, and let $g_{J_r} = \mu_b - \mu_{J_r}$ be the gap of its reference. By (30), $\mathbb{P}(F \mid \text{history}) \leq 4p^4$ and

$$\mathbb{E}[g_{J_r}\mathbf{1}_F \mid \text{history}] \leq C\varepsilon p^4. \tag{56}$$

For the second bound, condition also on J_r and use $xe^{-8Lx} \leq 1/(8eL)$ with $x = g_{J_r}/\varepsilon$.

The reference always survives. If AFE stops after this stage, all agents pull J_r in the final phase. Otherwise, the new maximum M^+ satisfies $\mu_b - M^+ \leq g_{J_r}$. In both cases, T times (56) pays for the loss caused by the deletion, either in the final phase or as the drop from μ_b to M^+ in (28). If AFE continues, we add the bound of Lemma C.9, applied given the new history, with $\varepsilon^+ = \varepsilon/2$, $a^+ = |A^+|$, and $S^+ = \sum_{i \in A^+}(M^+ - \mu_i)$.

The first sampling term of that bound satisfies

$$\mathbb{E}\left[\mathbf{1}_{F,\text{continue}}\frac{L^+(a^+)(S^+ + 2a^+\varepsilon^+)}{N(\varepsilon^+)^2}\,\middle|\,\text{history}\right] \leq \frac{CaL}{N\varepsilon}. \tag{57}$$

To see this, drop the indicators and apply Lemma C.8. On histories where AFE continues, the extension L^+ equals the actual confidence factor. The remaining sampling term of Lemma C.9, multiplied by $\mathbb{P}(F)$, is at most $C\sqrt{aH}\,p^4/N$. Since $p = a/(128H\varepsilon^2)$,

$$T\varepsilon p^4 \leq \frac{Ca}{N\varepsilon}, \qquad \frac{\sqrt{aH}}{N}p^4 \leq \frac{Ca}{N\varepsilon}. \tag{58}$$

For rounding and flooding, let $Z = \max_{i \in A^+}(\mu_b - \mu_i)$. The largest deficit of the new set relative to M^+ is at most Z. Equation (32) and the Cauchy–Schwarz inequality give $\mathbb{E}[Z\mathbf{1}_F \mid \text{history}] \leq C\varepsilon p^2\sqrt{J}$. The other term in (35) is multiplied by $\mathbb{P}(F) \leq 4p^4$. All regret after a possible first deletion in stage r is therefore at most

$$\frac{CaL}{N\varepsilon} + CdJ\varepsilon(p^2 + p^4)\sqrt{J}. \tag{59}$$

Summing over the stages while b remains gives at most $C\mathcal{W} + C\beta dJ^{3/2}$. The first term is part of (52), because $aL/(N\varepsilon) \leq L(S_r + 2a\varepsilon)/(N\varepsilon^2)$. The second follows from $p < 1/128$ and $\sum_r \varepsilon_r \leq 2\beta$.

It remains to collect the other parts of the regret while b remains. By Lemma C.2 and the validation step, the sampling cost without rounding is at most $C\mathcal{W}$. Rounding and flooding cost at most $C\beta dJ^{3/2}$. Indeed, the largest gap in A_0 is at most β, and after we drop the indicator that the next stage runs, (32) bounds the expected largest gap in stage r by $C\varepsilon_{r-1}\sqrt{J}$.

Suppose AFE stops at a later stage because the candidate count reaches the threshold while b remains. The previous stage then had $a \geq H\varepsilon^2/4$, so $T \leq 4a/(N\varepsilon^2)$. Dropping the indicator of stopping and using the expected deficit bound of Lemma C.2 bounds the final phase by $Ca/(N\varepsilon)$, which is again part of $C\mathcal{W}$. A stop with a single arm left costs nothing while b remains. If AFE stops at once after localization because $a_0 \geq H\beta^2$, then

$$T\beta \leq \frac{a_0}{N\beta} \leq 2\sum_{i \in A_0 \setminus \{b\}}\frac{1}{N\Delta_i} \leq 2F_2.$$

Lemma E.4 therefore shows that, given any history at the end of a successful localization, all later regret is at most

$$CF_2 + C\beta dJ^{3/2}. \tag{60}$$

Since F_2 depends only on the original instance, this bound holds whichever set A_0 localization returns.

E.4 Proof of the logarithmic bound

Bound (60) also gives the second bound of Theorem C.3. Since $0 < \Delta_i \le 1$ and $k_i^{(2)} \ge 1$,

$$\psi(H\Delta_i^2/k_i^{(2)}) \le \Lambda, \qquad F_2 \le \frac{\Lambda}{N}\sum_{i\neq b}\frac{1}{\Delta_i}.$$

By Lemma C.5, $L_{\mathrm{loc}} \le C\Lambda$, and $\Delta_i/\max\{\Delta_i,\beta\}^2 \le 1/\Delta_i$, the sampling cost of localization is at most the same sum. Its other costs are at most $3d + 1/2$, including all rounds after an inaccurate localization. An unfinished localization costs only the part that has run. Averaging (60) over the histories at the end of localization therefore gives

$$\mathbb{E}R_v(T) \le C\left[\Upsilon + \frac{\Lambda}{N}\sum_{i\neq b}\frac{1}{\Delta_i}\right].$$

When $\beta = 1$, there is no localization and no localization cost. Together with Appendix C.7, this completes the proof of Theorem C.3.

E.5 Localization cost and proof of the refined bound

There is no localization cost when $\beta = 1$. Otherwise, let $c_\beta = \log[4dJ(Q_0+1)] > 0$. Since $H = NT$ and $\beta^2 = K/(NdJ)$, the definition in (17) gives $L_{\mathrm{loc}} = \log(H\beta^2) + c_\beta$. For any $1 \le k \le K$ and $\Delta > 0$,

$$\frac{\Delta L_{\mathrm{loc}}}{\max\{\Delta,\beta\}^2} \le \frac{\psi(H\Delta^2/k)}{\Delta} + \frac{c_\beta + 1 + \log K}{\beta}. \tag{61}$$

For $\Delta \ge \beta$, use $\log(H\beta^2) \le \psi(H\Delta^2/k) + \log K$ and $c_\beta/\Delta \le c_\beta/\beta$. For $x = \Delta/\beta < 1$, expand

$$\log(H\beta^2) = \log(H\Delta^2/k) + \log k + 2\log(1/x)$$

and multiply by x/β. The first term is at most $\psi(H\Delta^2/k)/(\beta x)$. The others sum to at most $(\log K + c_\beta + 2/e)/\beta$, since $x\log(1/x) \le 1/e$. Because $2/e < 1$, (61) follows.

Applying this inequality with $k = k_i^{(2)}$ to the localization cost (26) on the good event bounds its sampling cost by $CF_2 + C\mathcal{B}_{\mathrm{loc}}^{(2)}$. The other localization costs are $3d$ and at most $1/2$ for failure. Combining these bounds with (60) proves Theorem E.1. As in Appendix C, all statements compare the actual run with its ideal continuation, and no estimate is used before its D rounds of flooding end.

F Instances with similar gaps

This appendix proves Theorem 5.3. We first bound the statistical cost and the cost of waiting from below on the similar-gap class, and then compare their sum with the upper bound of AFE. The statistical construction adapts Lattimore (2015, Appendix F, Theorem 19) to heterogeneous bounded rewards and to every horizon. All bounds here concern the worst case over the class, not optimality for each gap vector.

For $0 < \Gamma \le 1/8$, the similar-gap class $\mathcal{V}_\Gamma$ contains the instances in $\mathcal{V}(G,K)$ whose gaps satisfy

$$\Gamma \le \Delta_i \le 2\Gamma \qquad (i \neq b).$$

We keep the graph G and the horizon T, and define

$$x = \frac{NT\Gamma^2}{K}, \qquad S_\Gamma = \frac{K}{N\Gamma}\log(1+x), \qquad B_G = \Upsilon + \mathcal{B}_{\mathrm{loc}}^{(2)}. \tag{62}$$

The parameter Γ only defines the class, and AFE does not receive it. The lower bounds remain valid even if a protocol knows Γ and the finite family used below.

Proposition F.1 (Lower bound from a heterogeneous family)**.** *Fix any matrix (z_{iv}) with*

$$|z_{iv}| \le 1/8, \qquad \sum_{v=1}^{N} z_{iv} = 0 \quad (i \in [K]).$$

Define independent Bernoulli instances $\nu^0, \nu^2, \dots, \nu^K$ by

$$p^0_{1v} = 1/2 + \Gamma + z_{1v}, \qquad p^0_{iv} = 1/2 + z_{iv} \quad (i \ge 2),$$

where ν^i changes only the entries of arm i, to $p^i_{iv} = p^0_{iv} + 2\Gamma$. Write $\mathcal{F}_\Gamma(z) = \{\nu^0, \nu^2, \dots, \nu^K\}$. All members of this family belong to $\mathcal{V}_\Gamma$. For every protocol on every connected graph,

$$\max_{\nu \in \mathcal{F}_\Gamma(z)} \overline{\mathcal{R}}^{\pi}_{\nu}(T) \ge \frac{1}{256} S_\Gamma. \tag{63}$$

The bound requires no consistency assumption.

Proof. The offsets sum to zero over agents, so they do not change the global means. Every member of the family therefore has positive gaps in $[\Gamma, 2\Gamma]$, and all its means lie in $[3/8, 7/8]$. For a changed entry, the inequality $\log u \le u - 1$ gives

$$\mathrm{kl}(p, p + 2\Gamma) \le \frac{4\Gamma^2}{(p + 2\Gamma)(1 - p - 2\Gamma)} \le 64\Gamma^2 =: h.$$

Let R be the maximum average regret in (63), and put $A = (K-1)/N$. Under ν^0, every pull of an arm $i \ge 2$ costs Γ, so $\Gamma \sum_{i=2}^{K} \mathbb{E}_0 n_i(T) \le NR$. Some fixed $i \ge 2$ therefore satisfies $\mathbb{E}_0 n_i(T) \le R/(\Gamma A)$, and Lemma D.1 gives

$$\mathrm{KL}(\mathbb{P}^T_0, \mathbb{P}^T_i) \le h \mathbb{E}_0 n_i(T) \le hR/(\Gamma A).$$

Now append to the transcript an agent and a round drawn uniformly at random, independently of everything else. This leaves the divergence unchanged. Let E be the event that the chosen agent pulls arm i in the chosen round. The regret bounds imply

$$\mathbb{P}_0(E) = \frac{\mathbb{E}_0 n_i(T)}{NT} \le \frac{R}{\Gamma(K-1)T}, \qquad \mathbb{P}_i(E^c) \le \frac{R}{\Gamma T}.$$

The Bretagnolle–Huber inequality in (41) then yields $R \exp(hR/(\Gamma A)) \ge \Gamma T/4$. Put $u = hR/(\Gamma A)$. Since $ue^u \le e^{2u} - 1$,

$$R \ge \frac{\Gamma A}{2h} \log\left(1 + \frac{hT}{4A}\right) = \frac{A}{128\Gamma} \log\left(1 + \frac{16T\Gamma^2}{A}\right) \ge \frac{1}{256} S_\Gamma,$$

where the last step uses $K/(2N) \le A \le K/N$. The transcript contains all rewards, actions, and messages of all agents, so the bound imposes no restriction on sampling or communication beyond the model. □

Lemma F.2 (Cost of waiting for similar gaps)**.** *For every connected graph,*

$$\inf_{\pi} \sup_{\nu \in \mathcal{V}_\Gamma} \mathcal{R}_{\max \nu}^{\pi}(T) \ge \frac{\Gamma D}{4}. \tag{64}$$

Proof. Take the two mean matrices p^σ, the agent v, the set S, and the radius $r = \lceil D/2 \rceil$ from the proof of Lemma D.4. Replace each mean by

$$\widetilde{p}^{\sigma}_{iu} = \tfrac{1}{2} - \Gamma + 2\Gamma p^{\sigma}_{iu}, \qquad \sigma \in \{+1, -1\},$$

and keep independent Bernoulli rewards. All means lie in $[1/2 - \Gamma, 1/2 + \Gamma] \subset (0, 1)$. Every global gap is multiplied by 2Γ, so the positive gaps are Γ, and also $3\Gamma/2$ if $K \ge 3$. Both instances therefore belong to $\mathcal{V}_\Gamma$.

The two instances still coincide outside S, so Lemma D.3 shows that the actions of v have the same distribution under both instances through round $r \le D < T$. The gap calculation in Lemma D.4, scaled by 2Γ, shows that for any common action distribution the two expected regrets in one round sum to at least $2\Gamma \cdot (1/2) = \Gamma$. One of the two instances thus gives v expected regret at least $\Gamma r/2 \ge \Gamma D/4$. □

Theorem F.3 (Minimax bounds for similar gaps)**.** *For all integers $2 \le N, K \le T$, every connected graph, and $0 < \Gamma \le 1/8$, Algorithm 3 satisfies*

$$\begin{aligned}\frac{1}{512}\big[S_\Gamma + \Gamma D\big] &\le \inf_{\pi} \sup_{\nu \in \mathcal{V}_\Gamma} \mathcal{R}_{\max_\nu}^{\pi}(T) \\ &\le \sup_{\nu \in \mathcal{V}_\Gamma} \mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le C\big(S_\Gamma + B_G\big).\end{aligned} \tag{65}$$

The upper bound includes the costs of waiting and localization $B_G = \Upsilon + \mathcal{B}_{\mathrm{loc}}^{(2)}$.

Proof. Proposition F.1, the inequality $\mathcal{R}_{\max} \ge \overline{\mathcal{R}}$, and $\mathcal{F}_\Gamma(z) \subseteq \mathcal{V}_\Gamma$ give the lower bound $S_\Gamma/256$. Lemma F.2 gives the lower bound $\Gamma D/4$, possibly on a different instance. The larger of the two bounds is at least half their sum, which proves the lower side of (65).

For the upper side, every arm has gap at most 2Γ, so every round costs at most 2Γ. The bound $R_v(T) \le 2\Gamma T$ therefore holds on every trajectory and covers sampling, rounding, flooding, interrupted operations, and the final phase. In addition, every instance in the class satisfies

$$\frac{K}{4} \le k_i^{(2)} \le K, \qquad x \le \frac{H\Delta_i^2}{k_i^{(2)}} \le 16x.$$

Theorem E.1 then gives

$$\mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le C\left[\frac{K}{N\Gamma}\psi(16x) + B_G\right]. \tag{66}$$

If $x \le 1$, then $\log(1+x) \ge x/2$, so $\Gamma T \le 2S_\Gamma$, and the bound $R_v(T) \le 2\Gamma T$ gives $\mathcal{R}_{\max_\nu}^{\pi_{\mathrm{AFE}}}(T) \le 4S_\Gamma$. If $x \ge 1$, then $\psi(16x) \le C\log(1+x)$, and (66) gives the stated upper bound. All these upper bounds hold for the same algorithm, which never uses Γ. □

G Comparison with existing bounds

This appendix compares the complete regret bounds of existing algorithms with the bound of AFE. All bounds concern bounded independent heterogeneous rewards on a static undirected graph, and we keep every term of each bound. Every agent satisfies $R_v(T) \le T\Delta_{\max}$, where $\Delta_{\max} = \max_i \Delta_i$, so we take the minimum of each bound with this cap. When a method also proves a gap-free or adversarial bound, we take the smaller of its two branches.

Proposition G.1 (Separation from existing bounds)**.** *Let n tend to infinity through even integers and set*

$$N = K = n, \quad G = K_n, \quad D = 1, \quad T = n^4, \quad \ell = \log n, \quad \Gamma = \sqrt{\ell}/n^2.$$

In Proposition F.1, take $z_{iv} = s_v a_i$, where half the signs s_v equal $+1$, the other half equal -1, and $(a_1, a_2, a_3, \dots, a_n) = (-1/8, 1/8, 1/16, \dots, 1/16)$. On every member of this family, each existing bound in Table 1, after taking the minimum with $T\Delta_{\max}$, has order $n^2\sqrt{\ell}$. In contrast, AFE attains the optimal order $n^2 \log \ell/\sqrt{\ell}$ uniformly over the whole class $\mathcal{V}_\Gamma$. For absolute constants $c, C > 0$,

$$c\frac{n^2 \log \ell}{\sqrt{\ell}} \le \inf_{\pi} \max_{\nu \in \mathcal{F}_\Gamma(z)} \mathcal{R}_{\max_\nu}^{\pi}(n^4) \le \inf_{\pi} \sup_{\nu \in \mathcal{V}_\Gamma} \mathcal{R}_{\max_\nu}^{\pi}(n^4) \le C\frac{n^2 \log \ell}{\sqrt{\ell}}. \tag{67}$$

The existing bounds are therefore larger by a factor of order $\ell/\log \ell$.

Proof. Since $|z_{iv}| \le 1/8$ and half the signs s_v are positive, the offsets satisfy the conditions of Proposition F.1. Every member of the family therefore belongs to $\mathcal{V}_\Gamma$, and all its positive gaps lie in $[\Gamma, 2\Gamma]$. Write

$$Z = \sum_{i \ne b} \frac{1}{\Delta_i} = \Theta\left(\frac{n^3}{\sqrt{\ell}}\right), \qquad T\Delta_{\max} = \Theta(n^2\sqrt{\ell}).$$

All comparisons use the averaging matrix $W = \mathbf{1}\mathbf{1}^\top/n$, whose second largest singular value is $\sigma_2(W) = 0$. For DRRB and EpoInc-SE, we set the consensus prefactors $Q = C_G = \sqrt{n}$, and for EpoInc-SE we set the initialization length $a = 1$. DRRB and EpoInc-SE state their bounds on

an event of probability at least $1 - 2NK/T$. Since $R_v(T) \leq T$, adding $2NK$ turns each of these bounds into a bound in expectation, for all N and K. On our sequence, $2NK = 2n^2$ is smaller than the cap.

For DRRB (Zhang et al., 2025a, Thm. 1, Cor. 2), the leading terms of the two branches, $(16 \log T/N)Z$ and $8\sqrt{KT \log T/N}$, both have order $n^2\sqrt{\ell}$. The remaining terms, including $(D+1)\sum_i \Delta_i$ or $K(D+1)$, the consensus cost $O(QK/(1-\sigma_2(W)))$, and the charge $2NK$, are $O(n^2)$.

For EpoInc-SE (Zhang et al., 2025b, App. D.2), we keep the complete bound over epochs. With $\sigma_2(W) = 0$, $a = 1$, and $D = 1$, it reads

$$U_{\mathrm{E}} = 2n^2 + \sum_{i \neq b} \frac{\Delta_i}{2}(\bar{r}_i + 1)(\bar{r}_i + 2),$$
$$\bar{r}_i = \left\lceil \frac{4\sqrt{2 \log T}}{\sqrt{n}\Delta_i} - 2 + \frac{8(2\sqrt{n}+1)}{\Delta_i} \right\rceil + 2.$$

Both factors $\bar{r}_i + 1$ and $\bar{r}_i + 2$ exceed $y_i = 4\sqrt{2 \log T}/(\sqrt{n}\Delta_i)$. Hence

$$U_{\mathrm{E}} \geq \sum_{i \neq b} \frac{\Delta_i y_i^2}{2} = \frac{16 \log T}{n} Z = \Theta(n^2\sqrt{\ell}).$$

The cap has the same order, so the minimum of U_{E} and the cap is $\Theta(n^2\sqrt{\ell})$, whatever the size of the remaining terms.

For FedFTRL (Hu et al., 2026, Thm. 1), the graph parameter is $A_P = \min\{\log(NT), \sqrt{N}\}/(1 - \sigma_2(W)) + 2 + D = \Theta(\ell)$. The dominant term of the stochastic bound is $(90A_P/\log K)Z = \Theta(n^3/\sqrt{\ell})$, and the dominant term of the adversarial bound is $13\sqrt{A_P T \log K} = \Theta(n^2\ell)$. For the black-box approach, equation (4) of Qiu et al. (2026) gives the block length $B = \Theta(\log(KNT)) = \Theta(\ell)$. The dominant terms of their Theorem 6 are $(B^2/\log K)Z = \Theta(n^3\sqrt{\ell})$ in the stochastic bound and $\sqrt{BT \log K} = \Theta(n^2\ell)$ in the adversarial bound. Here $T/B \geq 2$, and rounding the last block adds only $O(B\Delta_{\max})$. Both adversarial bounds also contain a term of order $\sqrt{KT/N} = n^2$, which is smaller. All other terms of these bounds have no larger order, which gives the entries of Table 1. For both methods, the cap is smaller than both branches.

Finally, $x = NT\Gamma^2/K = \ell$ and $S_\Gamma = \Theta(n^2 \log \ell/\sqrt{\ell})$. The localization precision is $\beta = (2\log(en))^{-1/2}$, so the costs of waiting and localization satisfy $\Upsilon = O(\ell)$ and $\mathcal{B}^{(2)}_{\mathrm{loc}} = O(\ell^{3/2}) = o(S_\Gamma)$. Theorem F.3 therefore gives the upper bound in (67) for AFE, uniformly over $\mathcal{V}_\Gamma$. Proposition F.1 and the inequality $\mathcal{R}_{\max} \geq \overline{\mathcal{R}}$ give the lower bound on the family. □

The four methods work under different communication models, and they offer guarantees that our objective does not measure. DRRB and EpoInc-SE place no limit on the messages that agents broadcast to their neighbors, and EpoInc-SE also bounds the number of bits sent. FedFTRL is anytime and bounds the expected communication per agent and round. FedFTRL and the gossip-based black-box approach also have guarantees for adversarial rewards. Our comparison uses only their regret bounds on a common stochastic family. It concerns these proved bounds, not the actual regret of the algorithms.

H Further related work

Federated bandit algorithms optimize the global mean of heterogeneous local rewards and communicate through a server or by gossip (Shi & Shen, 2021; Zhu et al., 2021). For weighted global objectives, Réda et al. (2022) develop pure-exploration algorithms and prove asymptotic regret lower bounds for Gaussian rewards. On communication graphs, DRRB combines consensus with elimination and comes with finite-time regret bounds and asymptotic lower bounds (Zhang et al., 2025a). Xu & Klabjan (2025) establish statistical regret lower bounds on networks, and Liu et al. (2025) obtain logarithmic regret when edges are available at random. Recent work also studies bit-efficient communication (Zhang et al., 2025b), anytime stochastic and adversarial guarantees for FedFTRL

(Hu et al., 2026), and a reduction from gossip to delayed bandits (Qiu et al., 2026). For the static heterogeneous stochastic model, we prove matching upper and lower bounds on every connected graph, and the same algorithm also satisfies instance-dependent bounds. Appendix G compares complete regret bounds.

Collaborative pure exploration relates the speedup from parallel agents to the number of communication rounds (Hillel et al., 2013; Tao et al., 2019). Batched bandits measure the regret cost of observing feedback only between batches (Perchet et al., 2016; Gao et al., 2019). Karpov & Zhang (2024b) bound the total regret of homogeneous collaborative bandits with a limited number of global exchanges, and they allow adaptive local sampling between exchanges. Karpov & Zhang (2024a) characterize the identification time under heterogeneous rewards and limited global exchanges, and Agarwal et al. (2017) characterize the query complexity of identification as a function of the number of batches. In our model, every round, including rounds spent waiting for messages, counts toward the regret of each agent, and agents may communicate in every round.

For homogeneous rewards, Wang et al. (2020) achieve near-optimal total regret with efficient communication. On networks, running-consensus UCB (Landgren et al., 2016) and accelerated consensus (Martínez-Rubio et al., 2019) show how the mixing of the network affects regret. UCB-TCOM provides communication-efficient guarantees for each agent and extensions to communication graphs with delays (Wang et al., 2023). QuACK transfers single-agent guarantees to group regret when each arm has a common reward distribution across agents (Howson et al., 2025, Assumption 1, Theorem 1), and Coop-SE provides regret bounds for each agent (Barnea et al., 2025). In contrast, AFE estimates the unweighted global means when the reward distributions differ arbitrarily across agents.

AFE combines two classical tools, median selection and validation with new samples (Even-Dar et al., 2006; Karnin et al., 2013), with the estimation of global means and a regret accounting that covers every round. Logarithmic regret bounds go back to asymptotic allocation theory (Lai & Robbins, 1985) and to finite-time upper confidence bounds (Auer et al., 2002). Our bounds also build on minimax and effective-competition analyses of single-agent bandits (Audibert & Bubeck, 2009; Lattimore, 2015; 2018).