BEST-OF-BOTH-WORLDS LEARNING WITH DELAYED GRAPH FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study best-of-both-worlds (BoBW) online learning with delayed graph feedback, where an action reveals the losses of related actions only after a delay. Existing BoBW algorithms handle either feedback graphs or delays, but not both. We present the first BoBW algorithm for this setting. It requires no knowledge of the horizon, the suboptimality gaps, the delays, or the independence number. For undirected graphs, its regret is $O(\sqrt{\alpha T \log(eKT)} + \Psi_T + 1)$ against an oblivious adversary and $O(H_G(\Delta)\log(eKT)\log(eT) + \Delta_{\max}\sigma_{\max} + 1)$ in the stochastic regime, where α is the independence number, $\Psi_T \leq \sqrt{D \log K}$ is a delay term in which excessively delayed rounds may be skipped, D is the total delay, $H_G(\Delta)$ is the largest sum of inverse gaps over independent sets of suboptimal arms, and $\sigma_{\max}$ is the maximum number of rounds whose feedback is pending at the same time. The delay thus enters only additively, and in the stochastic regime its coefficient is a gap rather than an inverse gap. Under bandit and full-information feedback this coefficient improves to the average gap Δ_{avg}.

1 INTRODUCTION

In online learning with graph feedback, playing an action reveals the losses of its neighbors in a feedback graph, which interpolates between bandit and full-information feedback (Mannor & Shamir, 2011). A standard example is promotion on a social network: offering a promotion to one user also reveals how the user's friends respond (Caron et al., 2012; Buccapatnam et al., 2014), and the adversarial regret scales with the independence number of the friendship graph rather than with the number of users (Alon et al., 2017). In practice these responses, such as purchases, arrive only after a delay (Chapelle, 2014; Vernade et al., 2017), so the learner keeps choosing users while earlier responses are pending. The users who reveal the most, such as highly connected ones, need not be the best users to promote, and delays can make their cost accumulate over many rounds. The learner should exploit stationary responses and retain worst-case guarantees when they change, for instance under competing campaigns, without knowing which case it faces.

Best-of-both-worlds (BoBW) algorithms provide such guarantees (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Zimmert & Seldin, 2021). Delayed bandits are well studied under stochastic or adversarial losses (Joulani et al., 2013; Thune et al., 2019; Zimmert & Seldin, 2020), and BoBW algorithms have been extended to them (Masoudian et al., 2022; 2024). These FTRL-based algorithms have stochastic delay terms with inverse gaps, and the elimination-based Delayed SAPO of Schlisselberg et al. (2025) reduces this term to the average gap times the maximum number $\sigma_{\max}$ of pending rounds. For immediate graph feedback, BoBW algorithms exist for several graph classes (Erez & Koren, 2021; Kong et al., 2022; Ito et al., 2022; Rouyer et al., 2022; Dann et al., 2023). The stochastic bounds of Ito et al. (2022) and Dann et al. (2023) depend on the minimum gap, whereas the algorithm of Rouyer et al. (2022) retains individual gaps, with stochastic regret $O(H_G(\Delta)\log^2 T)$ up to T-independent terms, where $H_G(\Delta)$ is the largest sum of inverse gaps over independent sets of suboptimal arms. The closest models that combine delays with graph structure, communication networks between agents (Cesa-Bianchi et al., 2016) and temporal feedback graphs (Gatmiry & Schneider, 2024), consider adversarial losses. Existing BoBW algorithms handle either delays or feedback graphs, but not both.

This leads to our question: *can one algorithm, under delayed graph feedback, retain the graph-dependent BoBW guarantees of Rouyer et al. (2022) while paying for delays only additively?* Running

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.

Table 1: Pseudo-regret bounds of the closest prior results and of this paper, up to constant factors and additive constants. For bandit feedback, $H_G(\Delta) = \sum_{i \neq i^*} 1/\Delta_i$.

	Adversarial	Stochastic: learning	Stochastic: delay
Graph feedback, no delays			
Rouyer et al. (2022)[a]	$\sqrt{\alpha T \log K} + K$	$H_G(\Delta) \log^2 T$	—
Bandit feedback with delays			
Masoudian et al. (2024)[b]	$\sqrt{KT} + \bar{\Psi}_T$ $+S^* + K\hat{\sigma}_{\max}$	$\sum_{i \neq i^*} \log T/\Delta_i$	$\sum_{i \neq i^*} \frac{\hat{\sigma}_{\max}}{\Delta_i \log K}$ $+K\hat{\sigma}_{\max} + S^*$
Schlisselberg et al. (2025)[c]	$\sqrt{KT \log T} + \sqrt{D}$	$\sum_{i \neq i^*} \log T/\Delta_i$	$\Delta_{\text{avg}}\sigma_{\max}$
Lower bounds[d]	$\Omega(\sqrt{KT} + \bar{\Psi}_T)$	$\Omega(\sum_{i \neq i^*} \log T/\Delta_i)$	$\Omega(\Delta_{\text{avg}}\sigma_{\max})$
Graph feedback with delays			
This paper[e]	$\sqrt{\alpha T \log(eKT)} + \Psi_T$	$H_G(\Delta) \log(eKT) \log(eT)$	$\Delta_G \sigma_{\max}$

[a] Leading term of their Corollary 1 for undirected graphs, which adds $O(\alpha \log T)$ and T-independent terms.
[b] $\bar{\Psi}_T = \min_S\{|S| + \sqrt{D_{\bar{S}} \log K}\}$, where $D_{\bar{S}}$ is the total delay outside a skipped set S. $\hat{\sigma}_{\max} \leq \sigma_{\max}$ counts pending observations after skipping, and S^* is the number of rounds skipped by their algorithm (bounded in their Appendix H).
[c] Requires T. The adversarial bound holds up to logarithmic factors, and the delay term is that of their Appendix F variant (their main theorem has an extra $\log K$).
[d] Worst case over families of instances. Adversarial: Masoudian et al. (2022), for nonincreasing delays. Stochastic learning: Lai & Robbins (1985), asymptotically for consistent algorithms. Stochastic delay: Lancewicki et al. (2021), for constant delays and the worst assignment of gaps to arms.
[e] $\Psi_T \leq \bar{\Psi}_T$. The delay term is at most $2\Delta_{\text{avg}}\sigma_{\max}$ under bandit and full-information feedback (Proposition 3).

$d + 1$ copies of the algorithm of Rouyer et al. (2022) on interleaved rounds handles a known delay bound d, but multiplies the stochastic term by $d + 1$ and the adversarial term by $\sqrt{d + 1}$. Elimination, as in Delayed SAPO, avoids this multiplicative dependence, but graph feedback changes how delays cost regret: arms are observed without being played, so a pending round costs the gap of the action chosen to observe them, and the per-arm counting argument for delayed bandits does not transfer directly. Since this cost decreases only once informative feedback arrives, the difficulty is to show that the rounds played in the meantime add only an additive term. The coefficient of this term also depends on which arms remain active and on how ties among them are broken: a fixed tie-breaking rule may keep selecting a costly informative action, incurring regret up to $\Delta_{\max}$ per pending round on a complete graph.

Approach. Three ideas address these difficulties. First, the analysis splits the cost of each round into its decrease while the round's feedback is pending and its value when the feedback arrives. Since at most $\sigma_{\max}$ rounds are pending at the same time, the decreases add up to an additive delay term. The other two ideas refine its coefficient. Second, arms are eliminated roughly in decreasing order of their gaps, as in Schlisselberg et al. (2025), so that play does not concentrate on the worst remaining arms. Third, ties among the remaining arms are broken at random, so that the algorithm does not repeatedly select the active arm with the largest gap. These ideas are implemented within the structure of Delayed SAPO (Schlisselberg et al., 2025): the algorithm eliminates arms with confidence intervals, keeps observing eliminated arms at lower rates, and switches to exponential weights with skipping (Zimmert & Seldin, 2020; Gyorgy & Joulani, 2021) if a test on the received losses fails. The test rarely fails under stochastic losses, and while it passes, it controls the regret for arbitrary losses.

Contributions. We give the first BoBW algorithm for delayed graph feedback and answer the question affirmatively for undirected graphs, up to logarithmic factors. Without knowledge of the horizon, its pseudo-regret is $O(H_G(\Delta) \log(eKT) \log(eT) + \Delta_G \sigma_{\max} + 1)$ in the stochastic regime and $O(\sqrt{\alpha T \log(eKT)} + \Psi_T + 1)$ against an oblivious adversary (Theorem 1), where Δ_G is the largest average gap over maximal independent sets of the graph (Δ_{avg} under bandit feedback and $\Delta_{\max}$ under full information), and $\Psi_T \leq \sqrt{D \log K}$ is a skipping delay term (Zimmert & Seldin, 2020). Table 1 compares these bounds with prior results. The delay thus enters additively, with a gap rather than an inverse gap as its coefficient. Our main technical contribution is a delayed-summation lemma (Lemma 2) that accounts for the delay cost per round rather than per arm, since under graph

feedback arms are observed without being played. Unlike Delayed SAPO, the algorithm does not need the horizon: it runs in epochs and recomputes the confidence intervals of each epoch from the observations of the previous one, so that the delay cost is not paid again in every epoch. Random tie-breaking further improves the coefficient to $O(\Delta_{\mathrm{avg}})$ under bandit and full-information feedback, whereas on general graphs the average gap no longer suffices: no algorithm attains an $O(\Delta_{\mathrm{avg}})$ coefficient uniformly over graphs with a comparable learning term (Propositions 3 and 4). For directed graphs, analogous bounds hold with α replaced by the size of a maximum acyclic induced subgraph, and a high-probability variant holds in the stochastic regime (Appendix B).

2 Setting and Main Result

There are $K \geq 2$ actions, also called arms, and a fixed, known undirected graph G on $[K] = \{1, \dots, K\}$ with a self-loop at every vertex; $N_G(i)$ denotes the closed neighborhood of i. In round s, the learner draws I_s from a distribution p_s, incurs ℓ_{s,I_s}, and at the end of round $s + d_s$ receives the index s together with the losses $(\ell_{s,j})_{j \in N_G(I_s)}$. The delays d_s are nonnegative integers fixed before interaction, so feedback with $d_s = 0$ arrives at the end of round s.

Loss vectors belong to $[0,1]^K$. In the adversarial regime, an oblivious adversary fixes them before interaction. In the stochastic regime they are i.i.d. across rounds and independent of the delays, and coordinates within a vector may be dependent. The means μ_i have a unique minimizer i^*, and we write $\Delta_i = \mu_i - \mu_{i^*}$ for the suboptimality gaps, $\Delta = (\Delta_i)_{i \in [K]}$, $\Delta_{\min} = \min_{i \neq i^*} \Delta_i$, $\Delta_{\max} = \max_i \Delta_i$, and $\Delta_{\mathrm{avg}} = \frac{1}{K} \sum_{i=1}^K \Delta_i$. We use the pseudo-regret

$$R_T = \mathbb{E} \sum_{s=1}^{T} \ell_{s,I_s} - \min_i \mathbb{E} \sum_{s=1}^{T} \ell_{s,i}.$$

For fixed adversarial losses the minimizing arm is the best arm in hindsight, and for stochastic losses $R_T = \mathbb{E} \sum_{s=1}^{T} \langle p_s, \Delta \rangle$.

Graph quantities. Let $\alpha = \alpha(G)$ be the independence number; independence ignores self-loops. The first term of the stochastic bound, which we call the learning term, is governed by

$$H_G(\Delta) = \max \Big\{ \sum_{i \in S} \frac{1}{\Delta_i} : \; S \subseteq [K] \setminus \{i^*\} \text{ independent in } G \Big\}. \tag{1}$$

It satisfies $H_G(\Delta) \leq \alpha / \Delta_{\min}$ and has appeared in prior work on undirected feedback graphs (Lykouris et al., 2020; Rouyer et al., 2022). The delay term is governed by the graph-average gap

$$\Delta_G = \max_{C \in \mathcal{M}(G)} \frac{1}{|C|} \sum_{i \in C} \Delta_i, \tag{2}$$

where $\mathcal{M}(G)$ is the family of maximal independent sets of G (maximal by inclusion). It is the largest average gap of a set of pairwise non-adjacent actions that observes all arms. It equals Δ_{avg} under bandit feedback and $\Delta_{\max}$ under full information, and $\Delta_G \leq \Delta_{\max}$ always.

Delay quantities. We call round s pending at round t if $s \leq t < s + d_s$, that is, if its feedback has not yet arrived. In the stochastic regime, the delay cost comes from rounds played before earlier feedback can rule out suboptimal arms, so it depends on how many rounds are pending at the same time. In the adversarial regime, we use the skipping bound of Zimmert & Seldin (2020), in which each round is either skipped at unit cost or contributes its delay, truncated at the remaining horizon. Let

$$\begin{aligned} \sigma_{\max} &= \max_{1 \leq t \leq T} |\{s \leq t : s + d_s > t\}|, \\ \Psi_T &= \min_{U \subseteq [T]} \left\{ |U| + \sqrt{\log K \sum_{s \in [T] \setminus U} \min\{d_s, T - s\}} \right\}. \end{aligned} \tag{3}$$

Thus $\sigma_{\max}$ is the maximum number of rounds that are pending at the same time, as in Masoudian et al. (2024) and Schlisselberg et al. (2025); under graph feedback it counts pending rounds, each of which

may carry several observed losses. If all delays are at most d, then $\sigma_{\max} \le d$, and a single round with a very long delay contributes at most one to $\sigma_{\max}$. In Ψ_T, the set U collects rounds whose feedback is skipped, and the minimum is taken in hindsight; the choice $U = \varnothing$ gives $\Psi_T \le \sqrt{D \log K}$ for the total delay $D = \sum_{s \le T} d_s$.

Theorem 1 (Anytime best-of-both-worlds guarantee)**.** *Under the preceding assumptions, Algorithm 1, run with the anytime schedule of Appendix B, satisfies the following bounds for every fixed $T \ge 2$. In the adversarial regime,*

$$R_T = O\big(\sqrt{\alpha T \log(eKT)} + \Psi_T + 1\big). \tag{4}$$

In the stochastic regime,

$$R_T = O\big(H_G(\Delta) \log(eKT) \log(eT) + \Delta_G \sigma_{\max} + 1\big). \tag{5}$$

The algorithm takes only G as input, and the implied constants are universal. Without delays, the bounds match those of Rouyer et al. (2022) up to logarithmic factors, and under bandit feedback the delay term matches that of Schlisselberg et al. (2025) without requiring the horizon, while the learning term is larger by a logarithmic factor (Table 1). The adversarial bound matches the $\Omega(\sqrt{\alpha T})$ lower bound of Alon et al. (2017) up to logarithmic factors.

3 Algorithm

The algorithm runs in epochs whose lengths grow as $N_{m+1} = N_m^2$. Epochs let the confidence parameter depend on a known length although the horizon is unknown, and squaring keeps the total length of all but the last two epochs at $O(\sqrt{T})$ (Section 4.3). Each epoch starts with a statistical phase, which eliminates arms using confidence intervals, plays an exploration set that observes all arms through the graph, and tests whether the received losses remain consistent with the regret that the algorithm predicted. If the test fails, the algorithm runs a delay-adaptive exponential-weights algorithm until the epoch ends. Wherever a guarantee would need α, the algorithm uses an observed quantity that is at most α: the size of the exploration sets in the test and the ratios $\sum_i p_{t,i}/Q_i(p_t)$ in the fallback. Algorithm 1 summarizes one epoch of length N.

Within an epoch, t and s denote local rounds, d_s is the delay of the corresponding global round, and all sets and counts below involve only rounds of the current epoch. An arm stays active until its gap is estimated within a factor two and no active arm may have a gap more than twice as large (Section 3.1); it is then eliminated. Eliminated arms are still observed, and may be played, at lower rates, because the test compares the learner's loss with that of every arm, including eliminated ones.

For p in the probability simplex Δ^{K-1} and an arm i, let $Q_i(p) = \sum_{j : i \in N_G(j)} p_j$ be the probability of observing i when an action is drawn from p. The algorithm uses the importance-weighted estimates $\widehat{\ell}_{s,i} = \ell_{s,i} \mathbf{1}\{i \in N_G(I_s)\}/Q_i(p_s)$, whose denominator is the observation probability of round s. Statistics computed after the arrivals of round t carry the index t, so the decision of round t uses the gap estimates $g_{t-1,i}$ of Section 3.1 and the active set $\mathcal{A}_t = \{i : g_{t-1,i} = 0\}$.

3.1 Elimination with a common threshold

Let $\lambda = \log(16KN^5)$ be the confidence parameter. Let $n_{t,i}$ and $W_{t,i}$ be the number and sum of the observations of arm i received in the current epoch. Each arm keeps a confidence interval, obtained by intersecting its previous interval with $[W_{t,i}/n_{t,i} \pm \sqrt{\lambda/(2n_{t,i})}] \cap [0,1]$; an arm without observations in the current epoch keeps its interval. Write the endpoints as $\underline{\mu}_{t,i}, \overline{\mu}_{t,i}$. An empty intersection counts as a failure of the consistency test (Section 3.3). Otherwise compute the lower and upper confidence bounds on the gap of arm i:

$$\underline{\Delta}_{t,i} = [\underline{\mu}_{t,i} - \min_j \overline{\mu}_{t,j}]_+, \qquad \overline{\Delta}_{t,i} = \overline{\mu}_{t,i} - \min_j \underline{\mu}_{t,j}. \tag{6}$$

An arm whose gap is known up to a factor two can be observed at a lower rate. If such an arm left the active set while arms with larger gaps remained active, play would concentrate on the larger-gap arms, which Schlisselberg et al. (2025) avoid in bandits by eliminating arms roughly in decreasing order of their gaps. We obtain the same effect with a common threshold. Specifically, an arm j leaves the active set only if its positive gap is estimated within a factor two, $0 < \underline{\Delta}_{t,j}$ and $\overline{\Delta}_{t,j} \le 2\underline{\Delta}_{t,j}$,

Algorithm 1 One epoch of learning with delayed graph feedback

Input: Graph G, epoch length N, confidence parameter λ, initial intervals $[\underline{\mu}_i, \overline{\mu}_i]$
Initialize: $n_i, W_i, Z_i \leftarrow 0$ for all i; $v, \hat{\alpha} \leftarrow 0$
Initialize gap bounds, g_i, and $\mathcal{A}$ by (6)–(7)
for $t = 1, \dots, N$ **do**
  Sample a uniform permutation π_t; $\mathcal{C}_t \leftarrow$ exploration set built greedily in increasing $(g_i, \pi_t(i))$ order; $\hat{\alpha} \leftarrow \max\{\hat{\alpha}, |\mathcal{C}_t|\}$
  $f_t \leftarrow$ exploration rates from (8)
  $p_t \leftarrow f_t + (1 - \|f_t\|_1)\,\mathrm{Unif}(\mathcal{C}_t \cap \mathcal{A})$ ▷ *action distribution*
  $a_t \leftarrow \langle p_t, \overline{\Delta} \rangle$; store a_t
  Sample $I_t \sim p_t$ and play I_t; store I_t and $Q_i(p_t)$ for $i \in N_G(I_t)$
  for rounds s in this epoch with $s + d_s = t$ **do**
    Receive $\ell_{s,j}$ for $j \in N_G(I_s)$; $n_j \leftarrow n_j + 1$, $W_j \leftarrow W_j + \ell_{s,j}$ for these j
    $Z_i \leftarrow Z_i + \ell_{s,I_s} - \widehat{\ell}_{s,i} - a_s$ for all i; $v \leftarrow v + 1$ ▷ *received excess loss*
  end for
  $[\underline{\mu}_i, \overline{\mu}_i] \leftarrow [\underline{\mu}_i, \overline{\mu}_i] \cap [W_i/n_i \pm \sqrt{\lambda/(2n_i)}] \cap [0, 1]$ for each i with $n_i > 0$
  if some interval is empty **then** Run Algorithm 2 for $N - t$ rounds; **return** $\varnothing$
  $(\underline{\Delta}, \overline{\Delta}, g, \mathcal{A}) \leftarrow$ rules in (6)–(7) ▷ *elimination*
  if $Z_i + v g_i > h_t(v) + \frac{1}{4} t g_i$ for some i **then** Run Algorithm 2 for $N - t$ rounds; **return** $\varnothing$
end for
return $(W_i, n_i)_{i \in [K]}$

and no active arm i has $\overline{\Delta}_{t,i} > 2\underline{\Delta}_{t,j}$. The smallest active set satisfying these conditions is obtained in one pass through the common threshold ξ_t:

$$\xi_t = \max_{i:\overline{\Delta}_{t,i} > 2\underline{\Delta}_{t,i}} \overline{\Delta}_{t,i}, \quad g_{t,i} = \underline{\Delta}_{t,i} \mathbf{1}\{2\underline{\Delta}_{t,i} \geq \xi_t\}, \quad \mathcal{A}_{t+1} = \{i : g_{t,i} = 0\}, \tag{7}$$

where $\max \varnothing = 0$. We call $g_{t,i}$ the gap estimate (zero for active arms). Lemma 7 shows that the active set is nonempty and minimal, that $g_{t,i}$ is nondecreasing and ξ_t is nonincreasing along nested intervals, and that, when the intervals contain the means, active arms satisfy $\Delta_i \leq \xi_t$ and eliminated arms satisfy $g_{t,i} \leq \Delta_i \leq 2 g_{t,i}$.

3.2 Graph exploration with random tie-breaking

For a positive gap estimate $g_{t-1,i}$, the test of Section 3.3 tolerates a cumulative estimation error of order $t g_{t-1,i}$. The variance of importance-weighted estimates is controlled by $Q_i(p_s)^{-1}$, so an observation probability of order $\lambda/(t g_{t-1,i}^2)$ keeps the deviation of the cumulative estimate of arm i of order $t g_{t-1,i}$. Active arms, whose gaps are not yet estimated precisely, are instead observed in every round with probability of order at least $1/\alpha$.

The algorithm supplies these observations through a small set of actions. In every round of the statistical phase, it draws a fresh uniform permutation π_t, orders the arms by increasing $(g_{t-1,i}, \pi_t(i))$, and repeatedly selects the first uncovered arm and covers its closed neighborhood. The resulting exploration set $\mathcal{C}_t$ is an independent dominating set of G. Active arms have gap estimate zero and come first, so the active part $J_t = \mathcal{C}_t \cap \mathcal{A}_t$ is the maximal independent set of the induced subgraph $G[\mathcal{A}_t]$ built greedily in the order π_t, and $c_t = |J_t| \geq 1$. Distributing the active-arm mass over the independent set J_t, rather than over all active arms, keeps $c_t \leq \alpha$, so every active arm is observed with probability at least $1/(4\alpha)$ by (9) below. The order matters. On a complete graph J_t is a single arm, so a fixed tie-breaking order could select the same active arm with a large gap in every round and incur regret up to $\Delta_{\max}$ per pending round. With a fresh random order, J_t is a uniformly random active arm, whose expected gap is the average gap of the active arms. With $\kappa = 80$, define the rates $f_{t,i}$ of eliminated arms and the action distribution p_t by

$$f_{t,i} = \begin{cases} \min\{(4|\mathcal{C}_t|)^{-1}, \kappa\lambda/(t g_{t-1,i}^2)\}, & i \in \mathcal{C}_t \setminus \mathcal{A}_t, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$
$$p_t = f_t + (1 - \|f_t\|_1)\,\mathrm{Unif}(J_t),$$

where $\mathrm{Unif}(J_t)$ is the uniform distribution on J_t. The rates of eliminated arms sum to at most $1/4$, so p_t is a distribution. Every arm is observed through an arm of $\mathcal{C}_t$ with no larger gap estimate, and gap estimates only increase within an epoch, which gives

$$Q_i(p_s) \geq (4c_s)^{-1} \quad (i \in \mathcal{A}_s), \qquad \max_{1\leq s\leq t} Q_i(p_s)^{-1} \leq 4\hat{\alpha}_t + \frac{tg_{t,i}^2}{\kappa\lambda}, \qquad \hat{\alpha}_t = \max_{1\leq s\leq t} |\mathcal{C}_s|. \tag{9}$$

3.3 A CONSISTENCY TEST ON RECEIVED LOSSES

Let $O_t = \{s \leq t : s + d_s \leq t\}$ be the rounds of the current epoch whose feedback has arrived by the end of round t, let $v_t = |O_t|$, and let $\sigma_t = t - v_t$ be the number of pending rounds. After drawing π_s and before drawing I_s, the algorithm saves the regret prediction $a_s = \langle p_s, \overline{\Delta}_{s-1}\rangle$ of round s. It is an upper-confidence estimate of the regret of p_s and is at least $\langle p_s, \Delta\rangle$ when the intervals contain the means. For the analysis, let $\mathcal{F}^+_{s-1}$ contain everything drawn before I_s, and $\mathcal{F}_{s-1}$ everything drawn before π_s. Both include the loss vectors of earlier rounds whose feedback has not arrived and the observations used to initialize the epoch, whereas the algorithm uses only the feedback received so far. The corresponding conditional expectation $\mathbb{E}^+_{s-1}$ fixes p_s and is used for the test, and $\mathbb{E}_{s-1}$ averages over the random order and is used for the delay cost. For stochastic losses, unbiasedness and the inequality $a_s \geq \langle p_s, \Delta\rangle$ on the event that the intervals contain the means give

$$\mathbb{E}^+_{s-1}[\ell_{s,I_s} - \widehat{\ell}_{s,i} - a_s] = \langle p_s, \Delta\rangle - \Delta_i - a_s \leq -\Delta_i. \tag{10}$$

When feedback arrives, the algorithm adds the received terms to the test statistic

$$Z_{t,i} = \sum_{s\in O_t} (\ell_{s,I_s} - \widehat{\ell}_{s,i} - a_s), \qquad h_t(v) = 8\sqrt{\hat{\alpha}_t v\lambda} + 16\hat{\alpha}_t\lambda, \tag{11}$$

and continues while all confidence intervals are nonempty and

$$Z_{t,i} + v_t g_{t,i} \leq h_t(v_t) + \tfrac{1}{4} t g_{t,i} \qquad \text{for every } i. \tag{12}$$

Here h_t is the concentration width of $Z_{t,i}$ given the variance bound (9). It uses the observed size $\hat{\alpha}_t \leq \alpha$ of the exploration sets, each of which is an independent set of G, and the term $\frac{1}{4} t g_{t,i}$ absorbs the gap-dependent part of the variance. The term $v_t g_{t,i}$ makes the test stricter for eliminated arms: in the regret bound (14) below, it cancels the gap-dependent tolerance up to a term proportional to the number of pending rounds, while under stochastic losses it is offset by the drift $-\Delta_i \leq -g_{t,i}$ in (10). Section 4 explains why this single test serves both regimes.

3.4 AFTER A TEST FAILURE AND BETWEEN EPOCHS

When the test first fails at round τ, the algorithm runs Algorithm 2 (Appendix G) for the rest of the epoch. It is exponential weights with skipping (Gyorgy & Joulani, 2021) on importance-weighted graph estimates, and its learning rate and skipping threshold are set from the observed ratios $\sum_i p_{t,i}/Q_i(p_t) \leq \alpha$ (Rouyer et al., 2022; Alon et al., 2017). A new epoch recomputes its intervals from the raw sums and counts of the previous epoch (Section 4), or starts from $[0, 1]$ after a test failure. Late feedback from a finished epoch is discarded (Appendix B).

4 ANALYSIS

Appendices C–I give complete proofs.

4.1 ONE TEST FOR BOTH REGIMES

Let

$$M_t(i) = \sum_{s=1}^{t} \mathbb{E}^+_{s-1}[\ell_{s,I_s} - \ell_{s,i}]$$

be the sum of the conditional mean loss differences against arm i over the first t rounds of the epoch; its expectation is the expected regret against arm i. Concentration with (9) gives, except with

probability at most N^{-3} and simultaneously until the test fails,

$$\left| Z_{t,i} - \sum_{s \in O_t} \left(\mathbb{E}^+_{s-1}[\ell_{s,I_s} - \ell_{s,i}] - a_s \right) \right| \le h_t(v_t) + \tfrac{1}{4} t g_{t,i}. \tag{13}$$

The good event is the intersection of this event, the event that $n_{t,i} \ge (1 - e^{-1}) \sum_{s \in O_t} Q_i(p_s) - \lambda$ for every arm i, and, for stochastic losses, the event that all confidence intervals contain the means. On the good event, stochastic losses satisfy $g_{t,i} \le \Delta_i$ and (10), so the upper deviation gives

$$Z_{t,i} + v_t g_{t,i} \le h_t(v_t) + \tfrac{1}{4} t g_{t,i} + v_t(g_{t,i} - \Delta_i) \le h_t(v_t) + \tfrac{1}{4} t g_{t,i},$$

and the intervals remain nonempty, so the test does not fail. For arbitrary losses, the lower deviation and a passed test give

$$M_t(i) \le \sum_{s \in O_t} a_s + 2h_t(v_t) + \sigma_t + (t/2 - v_t) g_{t,i} \le \sum_{s=1}^{t} a_s + 2h_t(v_t) + \tfrac{3}{2}\sigma_t, \tag{14}$$

where the pending rounds contribute at most σ_t, and $t/2 - v_t \le \sigma_t/2$ and $g_{t,i} \le 1$ give the second inequality. Thus, as long as the test passes, the regret is controlled by the regret predictions a_s, the tolerance, and the number of pending rounds.

4.2 Stochastic regime: an additive delay cost

For stochastic losses, let $\bar{\gamma}_s = \mathbb{E}_{s-1}\langle p_s, \Delta \rangle$ be the conditional regret of round s, averaged over the permutation π_s. The expected regret of the first n rounds of the epoch is $\mathbb{E} \sum_{s \le n} \bar{\gamma}_s$, and $\bar{\gamma}_s$ splits into the active part $\bar{\gamma}^A_s = \mathbb{E}_{s-1}[(1 - \|f_s\|_1)|J_s|^{-1} \sum_{i \in J_s} \Delta_i]$ and the exploration of eliminated arms, which is at most $4\kappa H_G(\Delta)\lambda/s$ by (8) on the good event, since the eliminated arms of $\mathcal{C}_s$ form an independent set of suboptimal arms with $\Delta_i \le 2g_{s-1,i}$.

Let $X_t = \sum_{s \in O_t} (4c_s)^{-1}$ be the sum of the lower bounds $(4c_s)^{-1}$ on the probability of observing an active arm, over the rounds whose feedback has arrived by the end of round t, and let $\sigma_{\max}(n) = \max_{t \le n} \sigma_t$. Every arm of $\mathcal{A}_{t+1} \subseteq \mathcal{A}_s$ is observed with probability at least $(4c_s)^{-1}$ by (9), so on the good event it has at least $(1 - e^{-1})X_t - \lambda$ observations. Since $\xi_t = \max_{i \in \mathcal{A}_{t+1}} \overline{\Delta}_{t,i}$ (Lemma 7), this gives, even for arbitrary losses before a failure,

$$\xi_t \le \min\{1, \sqrt{64\lambda / X_t}\}. \tag{15}$$

The coefficient of the delay cost depends on the random order through the following quantity. Call $A \subseteq [K]$ admissible if $i^* \in A$ and $\Delta_i \le 2\Delta_j$ for all $i \in A$ and $j \notin A$; on the good event, every active set is admissible, since active arms have $\Delta_i \le \xi_t$ and eliminated arms have $\xi_t \le 2g_{t,j} \le 2\Delta_j$. For a uniform random permutation π of $[K]$, let $J_\pi(A)$ be the maximal independent set of $G[A]$ built greedily in the order π, so that $J_s = J_{\pi_s}(\mathcal{A}_s)$, and define

$$\Delta^\pi_G = \max\left\{ \frac{\alpha}{16 H_G(\Delta)}, \max_{A \text{ admissible}} \mathbb{E}_\pi \left[\frac{\sum_{i \in J_\pi(A)} \Delta_i}{|J_\pi(A)|} \right] \right\}. \tag{16}$$

The second term is the expected average gap of the active part of the exploration set, in the worst case over admissible active sets. The first term is at most $\Delta_{\max}/4$ and is needed for the second condition in (17) below. Extending $J_\pi(A)$ greedily to a maximal independent set of G and using admissibility shows $\Delta^\pi_G \le 2\Delta_G$ (Appendix D.2). While at least two arms are active, call $b_t = \min\{\Delta^\pi_G, \xi_t\}$ the level at round t; it is nonincreasing. Lemma 16 gives

$$\bar{\gamma}^A_s \le b_{s-1}, \qquad c_s \le 16 H_G(\Delta) b_{s-1}. \tag{17}$$

The first bound averages over π_s; the second holds for every permutation, since J_s is an independent set of active arms with gaps at most ξ_{s-1} and the first term of Δ^π_G covers the case $b_{s-1} = \Delta^\pi_G$. Thus the level b_{s-1} bounds the per-round cost of the active arms. By (15), it decreases as X_t grows, and since X_t includes only rounds whose feedback has arrived, the decrease comes only after the delays. The following lemma, proved as Lemma 14 in Appendix D, turns this into an additive delay cost.

Lemma 2 (Delayed summation). *Let $H > 0$, $\lambda \geq 1$, and $1 \geq b_0 \geq b_1 \geq \cdots \geq b_n \geq 0$. For arbitrary delays, with O_t and $\sigma_{\max}(n)$ as above, let $X_t = \sum_{s \in O_t} \omega_s$ with weights $\omega_s \in (0, 1/4]$. If $b_t^2 \leq C_1 \lambda / X_t$ whenever $X_t > 0$, and $C_2 H \omega_s b_{s-1} \geq 1$ whenever $b_{s-1} > 0$, for constants $C_1 \geq 1/4$ and $C_2 > 0$, then*

$$\sum_{s=1}^{n} b_{s-1} \leq C_1 C_2 \, H\lambda \log(en) + 2b_0(1 + \sigma_{\max}(n)).$$

Here (15) and (17) give $C_1 = 64$ and $C_2 = 64$. The proof splits each b_{s-1} into the decrease of the level between round s and the arrival of its feedback, and the level at arrival. The decreases add up to at most $(1 + \sigma_{\max}(n))b_0$: the level only decreases from b_0, and each decrease is shared by the current round and the at most $\sigma_{\max}(n)$ rounds that are pending. Since each b_{s-1} is at most twice its decrease plus a term in the level at arrival, the decreases contribute $2b_0(1 + \sigma_{\max}(n))$. The levels at arrival are bounded through the observations that arrive, and the two conditions make their total of order $H\lambda \log n$. Since $b_0 \leq \Delta_G^\pi$, the delay-dependent part is at most $2\Delta_G^\pi(1 + \sigma_{\max}(n))$.

After the last round with at least two active arms, the remaining active arm is optimal on the good event, and only eliminated arms contribute. Hence, on the good event, every prefix of the epoch satisfies, for a universal constant C,

$$\sum_{s=1}^{n} \bar{\gamma}_s \leq C H_G(\Delta) \lambda \log(en) + 2b_0(1 + \sigma_{\max}(n)). \tag{18}$$

Across epochs, analyzing each epoch from the intervals $[0, 1]$ would add a separate delay term for every epoch. Instead, each epoch keeps the raw sums and counts of the previous one and recomputes the intervals at its own confidence parameter. Once the previous epoch has at least $4(1 + \sigma_{\max})$ rounds, at least three quarters of its feedback has arrived, so the initial level of the new epoch is of order at most $\sqrt{\alpha\lambda / N_{m-1}}$. On consecutive good epochs, these upper bounds on the initial levels decrease geometrically, so the delay cost is not multiplied by the number of epochs. The total delay cost of good epochs is therefore $O(\Delta_G^\pi(1 + \sigma_{\max}) + H_G(\Delta) \log(eKT))$ (Lemmas 18 and 19). Bad epochs contribute $O(1)$ expected regret, since their failure probabilities weighted by the epoch lengths are summable (Appendix D.4). Since $\Delta_G^\pi \leq 2\Delta_G$, this gives the stochastic bound of Theorem 1 and the following bound, which is sharper, for example under full information.

Proposition 3 (Refined stochastic bound). *In the stochastic regime, the algorithm of Theorem 1 satisfies $R_T = O(H_G(\Delta) \log(eKT) \log(eT) + \Delta_G^\pi \sigma_{\max} + 1)$. For the graph with only self-loops (bandit feedback) and the complete graph (full information), $\Delta_G^\pi \leq 2\Delta_{\text{avg}}$, so the delay term is $O(\Delta_{\text{avg}} \sigma_{\max})$.*

4.3 Adversarial regime

Before a failure. For arbitrary losses, $c_s \leq \alpha$ gives $X_{s-1} \geq (s - 1 - \sigma_{\max}(n))_+ / (4\alpha)$, and summing (15) over the rounds, together with the rates of eliminated arms, gives, for a universal constant C,

$$\sum_{s=1}^{n} a_s \leq C\sqrt{\alpha n \lambda} + \sigma_{\max}(n) + 1. \tag{19}$$

Stopping at the first failure τ, the last passed test bounds the regret before τ through (14) and (19), and Appendix F relates $\sigma_{\max}(n)$ to the skipping term.

After a failure. From round $\tau + 1$, conditional on its starting history, Algorithm 2 faces fixed remaining losses and delays with independent randomness. Proposition 22 in Appendix G bounds its regret over n rounds by $O\big(\sqrt{\alpha n \log K} + |U| + 1 + \sqrt{\log K \sum_{s \in [n] \setminus U} \min\{d_s, n - s\}}\big)$ for every fixed $U \subseteq [n]$, following the skipping analysis of Gyorgy & Joulani (2021) with this data-dependent tuning.

Across epochs. Proposition 21 combines the bounds before and after a failure within an epoch, and Appendix F sums them over epochs with a common comparator and skipped set.

5 LIMITS OF THE AVERAGE GAP

Under bandit and full-information feedback, Proposition 3 matches, for constant delay d, a delay term of order $\Delta_{\mathrm{avg}} \min\{T, d+1\}$, which is unavoidable in the worst case over assignments of the gaps to the arms (Proposition 31; cf. Lancewicki et al., 2021). On general graphs, the average gap cannot replace Δ_G^{π}.

Proposition 4. *There is no algorithm that, for some constant C, satisfies*

$$R_T \le C\big(H_G(\Delta)\log(eKT)\log(eT) + \Delta_{\mathrm{avg}}\sigma_{\max} + 1\big)$$

for every connected undirected graph, bounded i.i.d. loss instance with a unique optimal arm, horizon T, and constant delay, even if the horizon and the delay are known. Along a sequence of connected graphs with $\alpha(G) = 2$ and diameter two, with $K = d$ and $T = \Theta(d \log\log d)$ as the constant delay d grows, every algorithm has an instance on which R_T exceeds $H_G(\Delta)\log(eKT)\log(eT) + \Delta_{\mathrm{avg}}\sigma_{\max} + 1$ by a factor $\Omega(\log\log T)$.

In the construction, a leaf v is joined to one clique arm b with a large gap, so the loss of the leaf is revealed only by playing v or b, and a constant delay prevents the learner from learning in time which instance it faces (Appendix H). Unlike in the complete-graph example of Section 3.2, on this construction Δ_G^{π} is not $O(\Delta_{\mathrm{avg}})$, even with a random order: when v is suboptimal with gap δ, the set $A = [K] \setminus \{b\}$ is admissible and v is isolated in $G[A]$, so $\Delta_G^{\pi} \ge \delta/2$, whereas $\Delta_{\mathrm{avg}}/\delta \to 0$ as d grows in the construction. Whether Δ_G^{π} is necessary remains open.

6 RELATED WORK

Delayed feedback. For stochastic bandits, Joulani et al. (2013) give black-box reductions that make the delay cost additive, and Lancewicki et al. (2021) allow unrestricted delay distributions. For adversarial bandits, Cesa-Bianchi et al. (2016) study constant delays, Thune et al. (2019) and Zimmert & Seldin (2020) handle arbitrary delays by skipping excessively delayed rounds, and Gyorgy & Joulani (2021) adapt the learning rate to the observed delays. Among BoBW algorithms, Masoudian et al. (2022) require the maximal delay for arbitrary delays, and Masoudian et al. (2024) tolerate arbitrarily large delays through implicit exploration and adaptive skipping.

Feedback graphs. Alon et al. (2015; 2017) relate the minimax adversarial regret to the independence number and related graph parameters, Kocák et al. (2014) use implicit exploration, and Eldowa et al. (2023) study the minimax regret in more detail. For stochastic losses, Lykouris et al. (2020) bound Thompson sampling and UCB through independent sets, and Marinov et al. (2022) study finite-time and asymptotic optimality. Esposito et al. (2022) study adversarial losses with stochastic feedback graphs. BoBW algorithms for feedback graphs cover different graph classes (Erez & Koren, 2021; Kong et al., 2022; Ito et al., 2022; Rouyer et al., 2022). Dann et al. (2023) give a general reduction to BoBW, which requires a base algorithm with a local self-bounding guarantee; they obtain such algorithms from ones that are stable under importance weighting. To our knowledge, no algorithm for delayed graph feedback is known to provide such a guarantee.

Graphs and delays. Cesa-Bianchi et al. (2016) study agents that exchange delayed messages over a communication network, and Gatmiry & Schneider (2024) study temporal feedback graphs, which specify the past rounds observed at each round. Both consider adversarial losses. Appendix A discusses the reduction of Dann et al. (2023) and further related work in more detail.

7 CONCLUSION

We gave a BoBW algorithm for delayed graph feedback that pays for delays only additively in both regimes, with a gap rather than an inverse gap as the coefficient in the stochastic regime. The delays are fixed before interaction, the stochastic guarantees assume i.i.d. losses with a unique optimal arm, and the adversarial guarantees assume an oblivious adversary. Under bandit feedback, our learning term is larger than that of Delayed SAPO by a logarithmic factor. Our analysis does not cover stochastic losses with adversarial corruptions, which self-bounding analyses handle (Zimmert & Seldin, 2021; Masoudian et al., 2022).

AI USE STATEMENT

In this work, we used generative AI tools to check the correctness of the proofs and the consistency of the paper. Tasks involving data sets, experiments, or code are not applicable to this theoretical work. Additionally, we used generative AI tools to polish and refine the writing and to search for related papers. We have reviewed all AI-assisted work. The authors checked every issue that the tools reported in the proofs and every suggested correction before adopting it, and read the papers found through search before citing them. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This work is theoretical and contains no experiments. The model and all assumptions are stated in Section 2, and Algorithm 1, the anytime schedule (Appendix B), and Algorithm 2 (Appendix G) specify every parameter, including all constants. Complete proofs of all results are given in the appendix: the notation and the dependencies between the proofs are summarized at the beginning of Appendix C, the stochastic and adversarial bounds are proved in Appendices D and F, the high-probability bound in Appendix E, and the lower bounds in Appendix H. Appendix I describes the computation and storage required to implement the algorithm.

REFERENCES

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017. doi: 10.1137/140989455.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 42.1–42.23, 2012. URL https://proceedings.mlr.press/v23/bubeck12b.html.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. In *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2014.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 142–151, 2012.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, 2016.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5503–5570. PMLR, 2023. URL https://proceedings.mlr.press/v195/dann23a.html.

Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2023.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2021.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *International Conference on Machine Learning*, 2023.

Khashayar Gatmiry and Jon Schneider. Adversarial online learning with temporal feedback graphs. In *Conference on Learning Theory*, 2024.

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, 2021.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, 2014.

Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *International Conference on Machine Learning*, 2022.

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

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5969–5978. PMLR, 2021. URL https://proceedings.mlr.press/v139/lancewicki21a.html. Full version with appendices: arXiv:2106.02436.

Thodoris Lykouris, Eva Tardos, and Drishti Wali. Feedback graph regret bounds for Thompson Sampling and UCB. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 592–614, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1e32e235eee1f970470a3a6658dfdd5-Paper.pdf.

Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9e79aefb538d02a7c0610fa43bdb0d0f-Abstract-Conference.html.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2022.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. In *Advances in Neural Information Processing Systems*, 2024.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Ofir Schlisselberg, Tal Lancewicki, Peter Auer, and Yishay Mansour. Improved best-of-both-worlds regret for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2025.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1287–1295, 2014. URL https://proceedings.mlr.press/v32/seldinb14.html.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, 2019.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Dirk van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and MDPs. In *Conference on Learning Theory*, 2023.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL https://jmlr.org/papers/v22/19-753.html.

A Additional related work

Delayed feedback. Joulani et al. (2013) give black-box reductions: in the adversarial setting they run several copies of a non-delayed algorithm, each of which receives its feedback before it acts again, and in the stochastic setting a single base algorithm is fed from queues of received feedback, which makes the delay cost additive. Lancewicki et al. (2021) study stochastic bandits with unrestricted delay distributions. For adversarial bandits, Cesa-Bianchi et al. (2016) study constant delays, and later algorithms handle arbitrary delays with an additive term of order $\sqrt{D \log K}$, where D is the total delay, by tuning the learning rate to the observed delays or by skipping feedback that arrives too late (Thune et al., 2019; Zimmert & Seldin, 2020; Gyorgy & Joulani, 2021). van der Hoeven et al. (2023) extend nonstochastic delayed analyses to combinatorial semi-bandits, linear bandits, and MDPs, and Esposito et al. (2023) study delayed bandits with intermediate observations.

Best of both worlds. BoBW algorithms either test which regime they are in and switch (Bubeck & Slivkins, 2012), or use a single algorithm that adapts to both regimes (Seldin & Slivkins, 2014; Zimmert & Seldin, 2021). For delayed bandits, the extension of Masoudian et al. (2022) to arbitrary delays requires knowledge of the maximal delay, Masoudian et al. (2024) tolerate arbitrarily large delays through implicit exploration and adaptive skipping, and Schlisselberg et al. (2025) match the known lower bounds in both regimes up to logarithmic factors. Dann et al. (2023) give a general reduction that turns a family of FTRL and OMD algorithms into BoBW algorithms. A natural alternative is to apply this reduction to a delayed graph algorithm. The reduction requires a base algorithm with a local self-bounding guarantee that holds at stopping times, and Dann et al. (2023) construct such base algorithms from algorithms that are stable under importance weighting. To our knowledge, no base algorithm with this guarantee is known under delayed graph feedback. Moreover, the stochastic bounds of their instantiations for graphs depend on the minimum gap rather than on $H_G(\Delta)$, and existing self-bounding analyses of delayed FTRL-based methods produce delay terms with inverse gaps (Masoudian et al., 2022; 2024).

Feedback graphs. Mannor & Shamir (2011) introduce learning with side observations, and Alon et al. (2015; 2017) relate the minimax adversarial regret to the independence number and related graph parameters; Kocák et al. (2014) use implicit exploration in this setting, and Eldowa et al. (2023) study the minimax regret in more detail. For BoBW guarantees with feedback graphs, Erez & Koren (2021), Kong et al. (2022), Ito et al. (2022), Rouyer et al. (2022), and Dann et al. (2023) give algorithms under different graph classes; the stochastic bounds of Ito et al. (2022) and Dann et al. (2023) depend on the minimum gap, while Rouyer et al. (2022) retain individual gaps through $H_G(\Delta)$. For purely stochastic losses, Lykouris et al. (2020) bound Thompson sampling and UCB through independent sets, and Marinov et al. (2022) study finite-time and asymptotic optimality. Esposito et al. (2022) study adversarial losses with stochastic feedback graphs.

Graphs and delays. The two models closest to ours that combine graph structure and delays consider adversarial losses. Cesa-Bianchi et al. (2016) study cooperating agents that exchange messages over a communication network and drop messages that take more than d hops to arrive. Gatmiry & Schneider (2024) study temporal feedback graphs, which specify the past rounds whose losses are observed at each round.

B Algorithm details, directed graphs, and high-probability bounds

Anytime schedule. Use $N_0 = 2K$ and $N_{m+1} = N_m^2$. In epoch m, use $\lambda = \log(16KN_m^5)$, or the high-probability variant below, fixed throughout that epoch. If the preceding epoch returned $(W_i^{\mathrm{old}}, n_i^{\mathrm{old}})_{i\in[K]}$, initialize $[\underline{\mu}_{0,i}, \overline{\mu}_{0,i}] = [W_i^{\mathrm{old}}/n_i^{\mathrm{old}} \pm \sqrt{\lambda/(2n_i^{\mathrm{old}})}] \cap [0,1]$ when $n_i^{\mathrm{old}} > 0$ and $[0,1]$ when $n_i^{\mathrm{old}} = 0$. Initially or after a failure, use $[0,1]^K$. Counters for the current epoch start at zero. Pending records and late arrivals from completed epochs are cleared at epoch boundaries. Reusing the observations of the previous epoch keeps its gap information, so a new epoch does not restart from uninformed intervals and pay the initial delay cost again.

The graph-average gap. For the graph-average gap Δ_G of (2), the first term of Δ_G^π is at most $\Delta_G/4$, and $\Delta_G^\pi \le 2\Delta_G \le 2\Delta_{\max}$; the self-loop-only graph has $\Delta_G = \Delta_{\text{avg}}$, and a complete graph has $\Delta_G = \Delta_{\max}$, which is the cost of a fixed exploration order.

Directed graphs. For a directed graph with every self-loop, use $N^{\text{out}}(j)$ for the losses revealed by action j and $Q_i(p) = \sum_{j:(j,i)\in E} p_j$. The exploration set is out-dominating: each selected arm covers its out-neighborhood. Every edge between two distinct selected vertices points from the later one to the earlier one, since otherwise the later vertex would already have been covered. Thus the exploration set induces an acyclic graph. Define

$$\rho = \operatorname{mas}(G), \qquad H_G^{\text{ac}}(\Delta) = \max_{I:\, G[I] \text{ acyclic},\ i^* \notin I} \sum_{i\in I} \frac{1}{\Delta_i}, \tag{20}$$

ignoring self-loops but not reciprocal two-cycles. For directed graphs, let $\mathcal{M}(G)$ be the family of induced acyclic out-dominating sets and take the maximum defining Δ_G in (2) over this family. For an undirected graph represented by reciprocal edges, the directed parameters reduce to α and $H_G(\Delta)$.

Theorem 5 (Directed feedback)**.** *Use out-neighborhoods in Algorithm 1. In the directed formulas for observation and the consistency test, replace $\hat{\alpha}_t$ by $\hat{\rho}_t = \max_{1\le s\le t} |\mathcal{C}_s| \le \rho$. In the adversarial regime,*

$$R_T = O(\sqrt{\rho T \log(eKT)} + \Psi_T + 1).$$

In the stochastic regime,

$$R_T = O(H_G^{\text{ac}}(\Delta) \log(eKT) \log(eT) + \Delta_G \sigma_{\max} + 1).$$

The proof uses the order of selected vertices to establish out-domination and bounds for induced acyclic sets. The algorithm does not require $\operatorname{mas}(G)$; the adversarial bound depends on its actual value. The stochastic delay coefficient remains Δ_G; extending the refinement to Δ_G^π to directed graphs is left open.

High-probability bounds. With a larger confidence parameter, the algorithm also controls the realized regret against the best arm in hindsight.

Corollary 6 (High-probability stochastic regret)**.** *For $\delta \in (0,1)$, use $\lambda_\delta = \log(16KN^5) + \log(64/\delta)$ in all confidence intervals, sampling probabilities, consistency tests, and interval recomputation steps. With probability at least $1-\delta$, simultaneously for all $T \ge 2$ under the stochastic assumptions, for an undirected graph,*

$$\sum_{s=1}^{T} \ell_{s,I_s} - \min_i \sum_{s=1}^{T} \ell_{s,i} = O(H_G(\Delta) \log(eKT/\delta) \log(eT) + \Delta_G^\pi \sigma_{\max} + 1). \tag{21}$$

For a directed graph, the same left-hand side is $O(H_G^{\text{ac}}(\Delta) \log(eKT/\delta) \log(eT) + \Delta_G \sigma_{\max} + 1)$. The same parameter setting has expected adversarial regret, with $\rho = \alpha$ for undirected graphs,

$$O\big(\sqrt{\rho T \log(eKT)} + \sqrt{\rho T \log(64/\delta)} + \Psi_T + 1\big). \tag{22}$$

The realized excess loss against i^* vanishes whenever $I_s = i^*$, so its conditional second moment is at most $\bar{\gamma}_s/\Delta_{\min}$, and a supermartingale argument converts the bound without an additional $\sqrt{T}$ term (Appendix E).

C Statistical estimates and the consistency test

We use a common analysis for both graph classes, with the directed parameters $\rho = \operatorname{mas}(G)$ and $H_G^{\text{ac}}(\Delta)$ of (20); ρ is an analysis parameter, not an input of the algorithm. For undirected graphs, $\rho = \alpha$ and $H_G^{\text{ac}}(\Delta) = H_G(\Delta)$. The statistical test uses $\widehat{\rho}_t = \max_{s\le t} |\mathcal{C}_s| \le \rho$ (denoted $\widehat{\alpha}_t$ for undirected graphs) and $h_t(v)$ from Section 3; no graph parameter is supplied to that test.

We call round s the source of the feedback it generates. Martingale increments are indexed by source rounds: since the delays are fixed, the set O_t of received sources is deterministic for each t, so for each fixed t a received statistic is a deterministically masked sum of source-order martingale

increments, even though its terms arrive out of order. Within the current epoch's statistical phase, the received statistics at local time t include exactly the sources $s \le t$ with $s + d_s \le t$. The algorithm updates its statistics when feedback arrives.

At every round in the statistical phase, a fresh uniform permutation π_t breaks ties when the algorithm sorts by $(g_{t-1,i}, \pi_t(i))$, for either graph class. The history $\mathcal{F}_{t-1}$ precedes this permutation, and $\mathcal{F}^+_{t-1}$ includes it but precedes the action and current loss. On rounds run by Algorithm 2, no permutation is drawn and we set $\mathcal{F}^+_{t-1} = \mathcal{F}_{t-1}$. Thus $\mathcal{C}_t$, p_t, and $Q_i(p_t)$ are fixed conditional on $\mathcal{F}^+_{t-1}$ in the statistical phase, while stochastic averages before the permutation use $\mathbb{E}_{t-1}$. The two conditionings serve different purposes. The test and its concentration are analyzed given $\mathcal{F}^+_{t-1}$, where p_t and the observation probabilities in the variance bounds are fixed, whereas the active regret, which drives the delay cost, is bounded given $\mathcal{F}_{t-1}$, because its bound by Δ^π_G holds only on average over π_t (Lemma 16).

Write $\gamma^A_t = \sum_{i \in \mathcal{A}_t} p_{t,i}\Delta_i$, $\bar{\gamma}^A_t = \mathbb{E}_{t-1}[\gamma^A_t]$, $\bar{\gamma}_t = \mathbb{E}_{t-1}[\langle p_t, \Delta\rangle]$, and $\bar{G}_t = \sum_{s=1}^t \bar{\gamma}_s$. For the delayed-sum arguments, the graph coefficient β equals Δ^π_G for undirected graphs and $2\Delta_G$ for directed graphs. The quantities Δ^π_G and $J_\pi(A)$ are defined in Section 4, and the family $\mathfrak{A}_\Delta$ of admissible sets in Appendix D.2.

Notation	Role
t, N	current local round and epoch length
O_t, v_t, σ_t	received sources, their number, and pending count within an epoch
$\sigma_{\max}$	maximum number of rounds pending at the same time, through T
$\sigma_{\max}(n) = \max_{t \le n} \sigma_t$	peak through local round n
$\sigma^{(m)}_{\max}$	peak within global epoch m
$\mathcal{A}_t, \mathcal{C}_t, J_t, c_t$	active set, exploration set, active exploration set, and its size
$\pi_t, \mathcal{F}_{t-1}, \mathcal{F}^+_{t-1}$	fresh uniform permutation and histories before and after it
$\mathbb{E}_{t-1}, \mathbb{E}^+_{t-1}$	conditional means before and after π_t is drawn
$\gamma^A_t, \bar{\gamma}^A_t$	realized-permutation active gap contribution and its pre-permutation mean
$\bar{\gamma}_t, \bar{G}_t$	pre-permutation total expected gap and its cumulative sum
β	Δ^π_G for undirected graphs, $2\Delta_G$ for directed graphs
$\widehat{\rho}_t$	largest exploration set through local round t
$\xi_t, g_{t,i}, \overline{\Delta}_{t,i}$	common elimination threshold, gap estimate, and upper gap bound
$a_s, Z_{t,i}$	regret prediction saved when round s is played, and received test statistic
$M_t(i)$	sum of post-permutation conditional mean loss differences
X_t, ε_m	sum of observation-probability lower bounds over received rounds, and initial levels
$\mathcal{P}^-_t, \mathcal{P}_t$	subroutine sources before and after the age test
$\Sigma_t, V_t, B_t, \theta_t$	cumulative wait and ratio, scale, current ratio

Reading order. The proofs follow three dependency chains. The shared chain, in the current appendix, establishes the threshold geometry, the observation rates, one concentration event, the consistency test (Proposition 10), the stopped-regret identity (Lemma 11), and the upper-confidence sum (Proposition 13). The stochastic chain verifies the hypotheses of the delayed-summation Lemma 14 within an epoch (Corollary 17) and sums the local bounds across epochs through interval recomputation and the weighted pending-feedback sum (Appendix D). Appendix E proves Corollary 6. The adversarial chain combines the test and the restart at the first failure (Lemma 20) with the subroutine bound of Proposition 22 (Appendices F and G). The effective history of Algorithm 2 is denoted $\mathcal{G}_t$ to distinguish it from the source-order history $\mathcal{F}_t$. Appendix H proves the lower bounds, and Appendix I records implementation costs.

C.1 THE ACTIVE SET AND THE OBSERVATION PROBABILITIES

The algorithm and its constants are defined in Section 3. We prove the two deterministic properties used below. Interval endpoints and gap estimates in this appendix are evaluated after arrivals; decision t uses their values at $t-1$. At each round in the statistical phase, a fresh uniform permutation breaks ties between equal gap estimates before the arm is drawn. All graph and observation assertions below hold for every realized permutation.

Lemma 7 (The common threshold and its minimality)**.** *For a collection of nonempty intervals, abbreviate $x_i = \underline{\Delta}_i$, $y_i = \overline{\Delta}_i$, and put $\mathcal{U} = \{i : x_i = 0 \text{ or } y_i > 2x_i\}$. Then $\mathcal{U} \neq \varnothing$ and*

$$\xi = \max_{i\in\mathcal{U}} y_i = \max_{i\in\mathcal{A}} y_i, \quad \text{and} \quad y_j \le 2g_j, \quad g_j = x_j \ge \xi/2 \quad (j \notin \mathcal{A}). \tag{23}$$

The set $\mathcal{A}$ is the unique smallest set, under inclusion, satisfying both

$$x_j > 0,\ y_j \le 2x_j \quad (j \notin \mathcal{A}), \quad \text{and} \quad y_i \le 2x_j \quad (i \in \mathcal{A},\ j \notin \mathcal{A}).$$

Along nested boxes, ξ_t is nonincreasing, $g_{t,i}$ is nondecreasing, and $\mathcal{A}_t$ decreases. On mean coverage,

$$i^* \in \mathcal{A}, \quad 0 \le g_i \le \Delta_i \le y_i, \quad \text{and} \quad \Delta_j \le 2g_j \quad (j \notin \mathcal{A}).$$

For $w_i = \overline{\mu}_i - \underline{\mu}_i$ and $w_* = \min_i \overline{\mu}_i - \min_i \underline{\mu}_i$, every nonempty interval box satisfies

$$y_i \le 2(w_i + w_*) \quad (i \in \mathcal{U}) \quad \text{and} \quad \xi \le 4 \max_{i\in\mathcal{A}} w_i. \tag{24}$$

An arm attaining the minimum lower endpoint belongs to $\mathcal{U}$ and has width at least w_*.

Proof. Zero lower gap bounds omitted from the maximum in (7) have $y_i = 0$, so both definitions of ξ agree. Every member of $\mathcal{U}$ remains active: if $x_i > 0$ and $y_i > 2x_i$, then $2x_i < \xi$. For an additional active arm, $0 < 2x_i < \xi$ and $y_i \le 2x_i$, so $y_i < \xi$. This proves the second equality in (23) and all inequalities for eliminated arms.

Every set satisfying the displayed two conditions must include $\mathcal{U}$. Its maximum upper gap bound is therefore at least ξ, so it must also include all positive-x_i arms with $2x_i < \xi$. These are exactly the remaining members of $\mathcal{A}$. Since $\mathcal{A}$ itself satisfies the conditions, it is the unique smallest such set.

Nested intervals make x_i increase and y_i decrease. Thus $\mathcal{U}$ decreases, ξ decreases, and a positive g_i remains positive and increases. Mean coverage gives $x_i \le \Delta_i \le y_i$ and $x_{i^*} = 0$. Finally, if $x_i > 0$ then $y_i - x_i = w_i + w_*$. For an arm failing the factor-two condition, $y_i < 2(w_i + w_*)$. If $x_i = 0$, the stronger bound $y_i \le w_i + w_*$ applies. An arm attaining $\min_j \underline{\mu}_j$ belongs to $\mathcal{U}$ and has $w_i \ge w_*$ because $\overline{\mu}_i \ge \min_j \overline{\mu}_j$. It also belongs to $\mathcal{A}$, so $w_* \le \max_{i\in\mathcal{A}} w_i$. Maximizing the first bound in (24) over $\mathcal{U}$ and using (23) proves the second. □

The first condition gives each eliminated arm a reliable positive gap estimate, which sets its exploration rate. The second enforces top-down elimination: on coverage, $\Delta_i \le y_i \le 2x_j \le 2\Delta_j$ for every active i and eliminated j, so no arm leaves the active set while an arm with more than twice its gap is still active. Hence, for undirected graphs, every active set is admissible on coverage (Appendix D.2), so its averaged active regret is at most Δ_G^π (Lemma 16); admissibility also gives $\Delta_G^\pi \le 2\Delta_G$. The threshold is the smallest value compatible with keeping all insufficiently estimated gaps active.

Lemma 8 (Ordered domination and observation bounds)**.** *The exploration set $\mathcal{C}_t$ is out-dominating and induced acyclic, $1 \le c_t \le |\mathcal{C}_t| \le \rho$, and each vertex i is observed by some selected vertex j with $g_{t-1,j} \le g_{t-1,i}$. Moreover the active exploration set J_t out-dominates $\mathcal{A}_t$. The distribution (8) is predictable and, before the consistency test fails,*

$$Q_i(p_s) \ge (4c_s)^{-1} \qquad (i \in \mathcal{A}_s), \tag{25}$$

$$Q_{t,i}^{-1,\max} := \max_{s\le t} Q_i(p_s)^{-1} \le 4\widehat{\rho}_t + \frac{t g_{t,i}^2}{\kappa\lambda}, \quad \text{where} \quad \widehat{\rho}_t = \max_{1\le s\le t} |\mathcal{C}_s| \le \rho. \tag{26}$$

Proof. Write selected vertices in their selection order, sorting by $(g_{t-1,i}, \pi_t(i))$. No arc goes from an earlier selected vertex to a later one, since the later vertex would have been covered. Thus every nonself arc goes backwards in this order, proving acyclicity. Out-domination is by a vertex selected no later in the gap-estimate order. Zero gap estimates are processed first, so J_t covers $\mathcal{A}_t$ and is

nonempty. Since $\|f_t\|_1 \le 1/4$, the direct mass $(1 - \|f_t\|_1)/c_t$ on each member of J_t makes p_t a distribution and gives $Q_i(p_t) \ge 3/(4c_t)$ for active i. Retaining the weaker bound (25), use the appropriate exploration rate of the covering vertex and gap-estimate monotonicity to obtain

$$Q_i(p_s) \ge \min\{(4|\mathcal{C}_s|)^{-1}, \kappa\lambda/(sg_{t,i}^2)\}.$$

Taking reciprocals and maximizing over $s \le t$ yields (26). The second term in this minimum is interpreted as infinity at $g_{t,i} = 0$. The exploration set is fixed after the permutation and before drawing at each round, so $\widehat{\rho}_t$ is predictable with respect to the post-permutation history. □

This proof covers directed graphs directly. Symmetry is needed only to identify acyclic sets with independent sets; the shared concentration and delayed-summation steps do not use it.

C.2 Concentration and the consistency test

Index increments by source rounds. Let $\mathcal{F}_{s-1}$ contain all past action randomization, past latent loss vectors, and the epoch's initial historical state, including the recomputed intervals. Observable history is a sub-sigma-field. On rounds in the statistical phase let $\mathcal{F}^+_{s-1} = \mathcal{F}_{s-1} \vee \sigma(\pi_s)$; on rounds run by Algorithm 2 set $\mathcal{F}^+_{s-1} = \mathcal{F}_{s-1}$, since no permutation is drawn. The distribution p_s, observation probabilities, regret prediction a_s, and exploration size c_s are $\mathcal{F}^+_{s-1}$-measurable. Write $\mathbb{E}^+_{s-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}^+_{s-1}]$. Define

$$r_{s,i} = \mathbb{E}[\ell_{s,I_s} - \ell_{s,i} \mid \mathcal{F}^+_{s-1}], \quad D_{s,i} = \ell_{s,I_s} - \widehat{\ell}_{s,i},$$
$$M_t(i) = \sum_{s=1}^{t} r_{s,i}.$$

Current action randomization is independent of the current loss vector conditionally on $\mathcal{F}^+_{s-1}$. Thus $\mathbb{E}^+_{s-1} D_{s,i} = r_{s,i}$ in both regimes. In particular

$$(D_{s,i} - a_s) - (r_{s,i} - a_s) = D_{s,i} - r_{s,i}.$$

The regret prediction a_s is $\mathcal{F}^+_{s-1}$-measurable, so subtracting it adds no random error.

Lemma 9 (Counts, empirical means, and received test deviations)**.** *Let τ be the first test failure after an arrival step. Except on an event of probability at most $(6KN^2 + KN)e^{-\lambda} \le e^{-\lambda_{\text{off}}} N^{-3}$, where $\lambda_{\text{off}} = \lambda - \log(16KN^5)$, the following hold simultaneously for all arms and all applicable $t \le \tau \wedge N$:*

$$(1 - e^{-1}) \sum_{s \in O_t} Q_i(p_s) - \lambda \le n_{t,i}, \tag{27}$$

$$\Big|Z_{t,i} - \sum_{s \in O_t} (r_{s,i} - a_s)\Big| \le h_t(v_t) + \tfrac{1}{4} t g_{t,i}. \tag{28}$$

We call the event in (27) the lower-count event; together with (28) and, for stochastic losses, interval coverage, it forms the good event of Section 4. Equation (28) is asserted when the updated intervals are nonempty. Under i.i.d. losses, the same event also gives, whenever $n_{t,i} > 0$,

$$\big|W_{t,i}/n_{t,i} - \mu_i\big| \le \sqrt{\lambda/(2n_{t,i})}. \tag{29}$$

Coverage of the incoming intervals is handled separately in Lemma 18.

Proof. For fixed deterministic t, the mask $w_s(t) = \mathbf{1}\{s + d_s \le t\}$ is deterministic. Also $\chi_s = \mathbf{1}\{s \le \tau\}$ is $\mathcal{F}_{s-1}$-measurable, hence also $\mathcal{F}^+_{s-1}$-measurable. We apply the following arguments to increments multiplied by $\chi_s w_s(t)$; this directly analyzes the actual stopped policy. On $t \le \tau$, all sources through t were played in the statistical phase.

Observation counts. Write $q_s = Q_i(p_s)$ and $Y_s = \mathbf{1}\{i \in N^{\text{out}}(I_s)\}$. Conditional means satisfy $\mathbb{E}^+_{s-1} Y_s = q_s$. The exponential process with increment

$$-Y_s + (1 - e^{-1}) q_s$$

has expectation at most one after exponentiation, since $\mathbb{E}^+_{s-1}e^{-Y_s} = 1-(1-e^{-1})q_s \le e^{-(1-e^{-1})q_s}$. Markov's inequality gives (27) for each arm and deterministic t except with probability $e^{-\lambda}$.

Empirical means. Under i.i.d. losses, conditional Hoeffding, first conditioning on the current action, gives for fixed z

$$\mathbb{E}\exp\left\{z\sum_{s=1}^{t}\chi_s w_s(t)Y_s(\ell_{s,i}-\mu_i) - \frac{z^2}{8}\sum_{s=1}^{t}\chi_s w_s(t)Y_s\right\} \le 1.$$

Fix (i,t,n) with deterministic $n \ge 1$. On $\{n_{t,i}=n\}$ the second sum equals n, so the exponent equals $z(W_{t,i}-n\mu_i) - z^2 n/8$. Markov's inequality with $z = 4x/n$ and $x = \sqrt{\lambda n/2}$ bounds the probability of $\{n_{t,i}=n\}\cap\{W_{t,i}-n\mu_i \ge x\}$ by $e^{-2x^2/n} = e^{-\lambda}$, and the lower tail is handled in the same way. A union over $n \le N$ gives (29) for the random count $n_{t,i}$ at cost $2Ne^{-\lambda}$ per arm and evaluation time, and therefore ordinary interval coverage.

Received test deviations. For the source-round loss difference, both losses are nonnegative, so $D_{s,i} \in [-1/q_s, 1]$, $r_{s,i} \in [-1,1]$, and

$$D_{s,i}^2 \le \ell_{s,I_s}^2 + \widehat{\ell}_{s,i}^2, \quad \mathbb{E}^+_{s-1}(D_{s,i}-r_{s,i})^2 \le 1+1/q_s, \quad |D_{s,i}-r_{s,i}| \le 1+1/q_s.$$

No independence between played and estimated loss is used. For the martingale argument use the interleaved filtration $\mathcal{F}_0 \subseteq \mathcal{F}_0^+ \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_1^+ \subseteq \cdots$. The step from $\mathcal{F}_{s-1}$ to $\mathcal{F}^+_{s-1}$, where π_s is drawn, carries a zero increment, and the step from $\mathcal{F}^+_{s-1}$ to $\mathcal{F}_s$ carries the centered increment $D_{s,i}-r_{s,i}$. The latter has conditional mean zero and the displayed variance and range bounds given $\mathcal{F}^+_{s-1}$. Fix the arm i, a deterministic evaluation time t and a deterministic cap b. Because the delays are fixed, the mask $w_s(t) = \mathbf{1}\{s+d_s \le t\}$ is deterministic, so the masked increments defined next are martingale differences for this filtration. For the deterministic dyadic cap b, multiply the increments by the predictable indicator that no source with $\chi_s/q_s > b$ has yet occurred, stopping before that source. After test failure, set these increments to zero; no sampling probability after the test failure is needed. The masked increments have range at most $1+b$ and predictable variance at most $(1+b)v_t$. Scalar Freedman (Tropp, 2011, Theorem 1.1), with $x = \sqrt{2V\lambda} + \frac{2}{3}R\lambda$ satisfying $x^2 \ge 2\lambda(V + Rx/3)$, applied at the fixed horizon t to both signs, bounds the deviation by

$$\sqrt{2(1+b)v_t\lambda} + \tfrac{2}{3}(1+b)\lambda$$

except with probability $2e^{-\lambda}$ per arm, evaluation time and cap. A union bound is then taken over arms, evaluation times, caps and signs. Choose the cap $Q_{t,i}^{-1,\max} \le b < 2Q_{t,i}^{-1,\max}$ only after this uniform event is obtained. Then $1+b \le 9\widehat{\rho}_t + 2tg_{t,i}^2/(\kappa\lambda)$ by Lemma 8, and $v_t \le t$, $g_{t,i}^2 \le g_{t,i} \le 1$ bound the display by

$$\sqrt{18\widehat{\rho}_t v_t\lambda} + 6\widehat{\rho}_t\lambda + \big(2/\sqrt{\kappa} + 4/(3\kappa)\big)tg_{t,i} \le h_t(v_t) + \tfrac{1}{4}tg_{t,i},$$

since $2/\sqrt{80} + 4/240 < 0.2403 < 1/4$. Since $N \ge 2K \ge 4$ and $\kappa\lambda \ge 1$, $Q_{t,i}^{-1,\max} \le 4K+N \le 3N$, so the $L_N = 1+\lceil\log_2(3N)\rceil \le 2N$ dyadic caps $1,2,\dots,2^{\lceil\log_2(3N)\rceil}$ suffice. Their union event is formed before substituting the random bound $Q_{t,i}^{-1,\max} \le 4\widehat{\rho}_t + tg_{t,i}^2/(\kappa\lambda)$; consequently the same event supports the data-dependent $\widehat{\rho}_t$. Counts cost $KNe^{-\lambda}$, empirical means $2KN^2e^{-\lambda}$ over arms, evaluation times and counts, and test deviations $2KNL_Ne^{-\lambda} \le 4KN^2e^{-\lambda}$ over arms, evaluation times, signs and caps. This proves the stated bound, since $16KN^2e^{-\lambda} = e^{-\lambda_{\mathrm{off}}}N^{-3}$. At $v_t = 0$ the centered sum is exactly zero. These arguments use source order and condition first on the realized permutation, without a union over permutations; they do not use conditional unbiasedness in arrival order. □

Proposition 10 (The consistency test in the two environments)**.** *Assume* (28)*. On stochastic mean coverage,* (12) *does not fail. At every time when the test has not failed,*

$$M_t(i) \le \sum_{s\in O_t} a_s + 2h_t(v_t) + \sigma_t + (t/2-v_t)g_{t,i} \le \sum_{s=1}^{t} a_s + 2h_t(v_t) + \tfrac{3}{2}\sigma_t. \tag{30}$$

Proof. Stochastic validity. Under coverage, $r_{s,i} - a_s = \langle p_s, \Delta \rangle - \Delta_i - a_s \le -\Delta_i$. Therefore

$$Z_{t,i} + v_t g_{t,i} \le h_t(v_t) + \tfrac{1}{4} t g_{t,i} + v_t(g_{t,i} - \Delta_i) \le h_t(v_t) + \tfrac{1}{4} t g_{t,i}.$$

Suppose the test first fails at some $\tau \le N$. Lemma 9 still applies at $t = \tau$, since it holds for all $t \le \tau \wedge N$. Coverage keeps all intervals nonempty, so (28) is available at τ, and the display shows that (12) holds at τ. This contradicts the failure at τ, so the test does not fail.

Regret while the test passes. Without stochastic assumptions, omitted sources contribute at most σ_t to $M_t(i)$. The opposite deviation inequality and the satisfied consistency test give

$$M_t(i) \le \sum_{s \in O_t} a_s + Z_{t,i} + h_t(v_t) + \tfrac{1}{4} t g_{t,i} + \sigma_t \le \sum_{s \in O_t} a_s + 2h_t(v_t) + \sigma_t + (t/2 - v_t) g_{t,i}.$$

Now $t/2 - v_t = (\sigma_t - v_t)/2 \le \sigma_t/2$, and $0 \le g_{t,i} \le 1$. □

The same two-sided estimation bound proves both assertions.

Lemma 11 (Expected regret at the test failure time)**.** *Take $\tau = \infty$ if the test never fails. For every deterministic $n \le N$ and fixed comparator i,*

$$\mathbb{E} \sum_{s=1}^{\tau \wedge n} (\ell_{s,I_s} - \ell_{s,i}) = \mathbb{E} M_{\tau \wedge n}(i). \tag{31}$$

Proof. The indicator $\mathbf{1}\{s \le \tau \wedge n\}$ asks whether the test failed before decision s and is $\mathcal{F}_{s-1}$-measurable. Multiply this indicator by $(\ell_{s,I_s} - \ell_{s,i}) - r_{s,i}$ and sum over $s \le n$. Each summand has conditional mean zero given $\mathcal{F}^+_{s-1}$, and the tower property over the fresh permutation makes its expectation zero. □

Remark. The identity is applied at $\tau \wedge n$ rather than at $\tau - 1$. The event $\{s \le \tau - 1\} = \{\tau \ge s + 1\}$ involves the failure decision of round s, so it does not in general have the $\mathcal{F}_{s-1}$-measurability needed for the martingale argument. Proposition 21 therefore bounds $M_{\tau \wedge n}(i)$ through the last passed test and adds at most one for the failing round.

C.3 A COMMON CONFIDENCE BOUND

For analysis only, define

$$\omega_s = (4c_s)^{-1} \quad \text{and} \quad X_t = \sum_{s \in O_t} \omega_s. \tag{32}$$

Lemma 12 (Contraction of the common gap threshold)**.** *On the lower-count event, whenever the updated intervals are nonempty,*

$$\overline{\Delta}_{t,i} \le \xi_t \le \min\{1, \sqrt{64\lambda / X_t}\} \qquad (i \in \mathcal{A}_{t+1}). \tag{33}$$

The minimum is one when $X_t = 0$.

Proof. Every currently active arm was active at each earlier round of this epoch, so the lower-count event gives it at least $(1 - e^{-1})X_t - \lambda$ observations. If $0 < X_t \le 64\lambda$, use $\xi_t \le 1 \le \sqrt{64\lambda/X_t}$; at $X_t = 0$ the square-root term is infinite. If $X_t > 64\lambda$, each active arm has $n_{t,i} \ge (1 - e^{-1} - 1/64)X_t \ge X_t/2$, and its interval has width at most $\sqrt{2\lambda/n_{t,i}} \le 2\sqrt{\lambda/X_t}$, since intersecting and clipping cannot increase width. The aggregate bound in (24) gives $\xi_t^2 \le 64\lambda/X_t$. Apply (23) to each active arm. Mean coverage is not required. □

Exploration of eliminated arms. Write $\mathcal{C}^+_s = \mathcal{C}_s \setminus \mathcal{A}_s$. The factor-two elimination rule and the specified exploration rates give

$$\sum_{i \notin \mathcal{A}_s} p_{s,i} \overline{\Delta}_{s-1,i} \le 2 \sum_{i \in \mathcal{C}^+_s} \min\left\{ \frac{g_{s-1,i}}{4|\mathcal{C}_s|}, \frac{\kappa\lambda}{s g_{s-1,i}} \right\}. \tag{34}$$

Using $\min\{x, y\} \le \sqrt{xy}$ bounds this by $\sqrt{\kappa\rho\lambda/s}$ without mean coverage. On stochastic coverage, instead use $g_{s-1,i} \ge \Delta_i/2$ and the inverse-gap weight of the acyclic set $\mathcal{C}_s^+$: the same display is at most $4\kappa H_G^{\mathrm{ac}}(\Delta)\lambda/s = 320H_G^{\mathrm{ac}}(\Delta)\lambda/s$. Since $\Delta_i \le \overline{\Delta}_{s-1,i}$ on coverage, this also bounds the exploration regret of eliminated arms. In particular, on stochastic coverage the eliminated-arm regret obeys

$$\sum_{i \notin \mathcal{A}_s} p_{s,i}\Delta_i \le 4\kappa H_G^{\mathrm{ac}}(\Delta)\lambda/s. \tag{35}$$

C.4 The nonstochastic upper-confidence bound

Proposition 13 (Upper-confidence sum)**.** *On the lower-count event, before a test failure,*

$$\sum_{s=1}^{n} a_s \le 50\sqrt{\rho n\lambda} + \sigma_{\max}(n) + 1. \tag{36}$$

Proof. For active arms, the observation bound gives

$$X_{s-1} \ge \frac{(s - 1 - \sigma_{\max}(n))_+}{4\rho}.$$

The first $\sigma_{\max}(n)+1$ contributions are at most one. Lemma 12 bounds the remaining contributions by $\sqrt{256\rho\lambda/(s-1-\sigma_{\max}(n))}$; summing $j^{-1/2}$ gives at most $2\sqrt{256\rho n\lambda}$. The bound on exploration of eliminated arms (34) contributes at most $2\sqrt{\kappa\rho n\lambda}$. Since $2\sqrt{256} + 2\sqrt{80} < 50$, the result follows. □

No stochastic mean was used.

D Stochastic regret

The proof combines a deterministic delay bound with graph-dependent bounds on active regret and exploration size. These give a local regret bound; interval recomputation then controls the sum over epochs.

D.1 A deterministic bound for delayed observations

Lemma 14 sums a nonincreasing level over the prefix during which at least two arms remain active. Round s pays b_{s-1}, but its feedback lowers the level only when it arrives, and the proof splits b_{s-1} into a staleness part $b_{s-1} - y_s$, where y_s is the level at arrival, and an information part. In this decomposition only the bound on the staleness part depends on the delays, and it is additive: each decrease of the level is counted once for the current source and once for each pending source, which gives at most $(1 + \sigma_{\max}(n))b_0$. The information part is weighted by ω_s through the weight condition $64H\omega_s b_{s-1} \ge 1$ and is bounded by the integral of $\min\{b_0^2, 64\lambda/x\}$ over the received weight.

Lemma 14 (Delayed summation)**.** *Let $H > 0$, $1 \ge b_0 \ge b_1 \ge \cdots \ge b_n \ge 0$ and $\lambda \ge 1$. Let $X_0 = 0$ and let X_t accumulate positive weights ω_s when source $s \in [n]$ arrives at $s + d_s$, where d_s is a nonnegative integer, and let $\sigma_t = |\{s \le t : s + d_s > t\}|$ and $\sigma_{\max}(n) = \max_{t \le n} \sigma_t$. Assume $b_t^2 \le 64\lambda/X_t$ for every $0 \le t \le n$ with $X_t > 0$, $X_n \le n/4$, and $64H\omega_s b_{s-1} \ge 1$ whenever $b_{s-1} > 0$. Then*

$$\sum_{s=1}^{n} b_{s-1} \le 4096H\lambda\log(en) + 2b_0(1 + \sigma_{\max}(n)). \tag{37}$$

Proof. Set $y_s = b_{s+d_s}$ if source s arrives by n, and $y_s = 0$ otherwise; thus feedback arriving after this prefix does not enter X_n or y_s. Counting each decrease of the threshold once for the current source and once for each source still pending gives

$$\sum_{s=1}^{n}(b_{s-1} - y_s) = \sum_{t=1}^{n}(1 + \sigma_{t-1})(b_{t-1} - b_t) + \sigma_n b_n \le (1 + \sigma_{\max}(n))b_0.$$

The terminal term accounts for feedback still pending at n. For $b_{s-1} > 0$, the inequality $(b_{s-1} - y_s)^2 \geq 0$ and the weight assumption give $b_{s-1} \leq 2(b_{s-1}-y_s)+y_s^2/b_{s-1} \leq 2(b_{s-1}-y_s)+64H\omega_s y_s^2$; zero terms contribute nothing. Grouping sources by their arrival time, including simultaneous arrivals, yields

$$\sum_{\substack{s=1 \\ s+d_s \leq n}}^{n} \omega_s y_s^2 = \sum_{t=1}^{n} (X_t - X_{t-1}) b_t^2 \leq \int_0^{X_n} \min\{b_0^2, 64\lambda/x\}\, dx \leq 64\lambda \log(en).$$

If $b_0 = 0$, the integral vanishes. Otherwise, split at $64\lambda/b_0^2$; the bound follows from $\lambda \geq 1$, $n \geq 1$, and $X_n \leq n/4$. Combining the two displays proves the claim. □

The proof uses the two constants 64 only through the two conditions of the lemma. For general constants $C_1 \geq 1/4$ and $C_2 > 0$ in these conditions, the same argument gives $C_1 C_2 H\lambda \log(en) + 2b_0(1 + \sigma_{\max}(n))$, which is the form stated in Lemma 2.

D.2 The graph-average gap and its randomized refinement

Let $\mathcal{M}(G)$ contain the induced acyclic sets that out-dominate the graph (Appendix B). The actual greedy set belongs to this family. In an undirected graph the family consists of maximal independent sets. Define

$$\Delta_G = \max_{C \in \mathcal{M}(G)} \frac{1}{|C|} \sum_{i \in C} \Delta_i.$$

This analysis parameter bounds the average gap of every exploration set that the algorithm may select.

For undirected G, every independent set extends to a maximal independent set. Hence the same family also represents the learning quantity:

$$H_G(\Delta) = \max_{C \in \mathcal{M}(G)} \sum_{i \in C \setminus \{i^*\}} \frac{1}{\Delta_i}.$$

Lemma 15 (The graph coefficient in the restart bound)**.** *For the graph coefficient,*

$$\rho/(2\Delta_G) \leq 2H_G^{\mathrm{ac}}(\Delta).$$

Proof. A maximum-cardinality induced acyclic set C is out-dominating: a vertex receiving no edge from C could be added without creating a cycle. For $\rho \geq 2$, remove the optimal vertex if present. Cauchy–Schwarz gives

$$H_G^{\mathrm{ac}}(\Delta) \sum_{i \in C} \Delta_i \geq (\rho - 1)^2, \quad \Delta_G \geq \frac{(\rho-1)^2}{\rho H_G^{\mathrm{ac}}(\Delta)} \geq \frac{\rho}{4H_G^{\mathrm{ac}}(\Delta)}.$$

If $\rho = 1$, every pair of distinct vertices is reciprocal, so $\Delta_G = \Delta_{\max}$ and $H_G^{\mathrm{ac}} = 1/\Delta_{\min}$; the claim follows directly. □

For undirected graphs, call a nonempty set $A \subseteq [K]$ admissible if $i^* \in A$ and $\Delta_i \leq 2\Delta_j$ for every $i \in A$ and $j \notin A$; write $\mathfrak{A}_\Delta$ for this family. The full set is admissible. Coverage and Lemma 7 make every actual active set admissible. For a uniform permutation π, let $J_\pi(A)$ be the greedy maximal independent set of the subgraph induced by A, in permutation order. Define the undirected analysis parameter

$$\Delta_G^\pi = \max\left\{ \frac{\alpha(G)}{16H_G(\Delta)}, \max_{A \in \mathfrak{A}_\Delta} \mathbb{E}_\pi \left[\frac{1}{|J_\pi(A)|} \sum_{i \in J_\pi(A)} \Delta_i \right] \right\}. \tag{38}$$

In the remainder of this appendix let $\beta = \Delta_G^\pi$ for undirected graphs and $\beta = 2\Delta_G$ for directed graphs. The floor $\alpha(G)/(16H_G(\Delta))$ keeps the exploration size comparable to the graph coefficient: it gives $c_s \leq \rho \leq 16H_G^{\mathrm{ac}}(\Delta)\beta$, so the weight condition $64H\omega_s b_{s-1} \geq 1$ of Lemma 14 holds also when the level equals Δ_G^π; for directed graphs the same bound $\rho/\beta \leq 16H_G^{\mathrm{ac}}(\Delta)$ follows from Lemma 15. The bound $\rho/\beta \leq 16H_G^{\mathrm{ac}}(\Delta)$ also lets the factor ρ/β in the restart bound (43) be absorbed into $H_G^{\mathrm{ac}}(\Delta)$. The active-regret and exploration-size bounds needed for delayed summation are proved in Lemma 16.

Comparison with the graph-average gap. Moreover $\beta \le 2\Delta_G$ in an undirected graph: extend each $J_\pi(A)$ by continuing the greedy order through $[K] \setminus A$, obtaining a maximal independent set C. Admissibility implies every added vertex has gap at least half the average gap of $J_\pi(A)$, so the average over C is at least half that over $J_\pi(A)$. The floor is at most $2\Delta_G$ by Lemma 15.

Bandit and full-information feedback. For the self-loop-only graph, the preceding comparison and $\Delta_G = \Delta_{\mathrm{avg}}$ give $\Delta_G^\pi \le 2\Delta_{\mathrm{avg}}$. For a complete graph, $J_\pi(A)$ is a uniform singleton of A. If its mean gap is d_A, admissibility gives $\Delta_j \ge d_A/2$ for every $j \notin A$, so $d_A \le 2\Delta_{\mathrm{avg}}$. The floor is $\Delta_{\min}/16 \le \Delta_{\mathrm{avg}}$, hence $\Delta_G^\pi \le 2\Delta_{\mathrm{avg}}$ there as well.

Lemma 16 (Averaged active regret and realized exploration size)**.** *Assume mean coverage, starting from covered initial intervals. Put $b_{s-1} = \min\{\beta, \xi_{s-1}\}$. In both graph classes let $\bar\gamma_s^A = \mathbb{E}_{\pi_s}[\sum_{i\in\mathcal{A}_s} p_{s,i}^{\pi_s}\Delta_i \mid \mathcal{F}_{s-1}]$. Then $\bar\gamma_s^A \le b_{s-1}$. For every realized permutation, if at least two active arms remain,*

$$c_s \le 16 H_G^{\mathrm{ac}}(\Delta) b_{s-1}. \tag{39}$$

These claims hold for arbitrary covered initial intervals, including intervals recomputed at epoch boundaries.

Proof. Suppress time indices and write $c = |J|$. For each realized permutation the active contribution is exactly $(1 - \|f\|_1)c^{-1}\sum_{i\in J}\Delta_i$. In an undirected graph, all active gap estimates vanish, so the fresh permutation induces $J = J_\pi(\mathcal{A})$. Since $0 \le 1 - \|f\|_1 \le 1$, keeping this factor inside the permutation expectation and using admissibility proves $\bar\gamma^A \le \Delta_G^\pi$. Coverage gives $\Delta_i \le \overline{\Delta}_i \le \xi$ for active i, proving $\bar\gamma^A \le \xi$. The bound by Δ_G^π holds after averaging over the permutation, and the bound by ξ holds for every permutation.

For directed graphs, the elimination rule implies $\Delta_i \le 2\Delta_j$ whenever i is active and j is eliminated. Every eliminated arm in the exploration set therefore has gap at least $\frac{1}{2c}\sum_{i\in J}\Delta_i$, whence $\sum_{i\in\mathcal{C}_s}\Delta_i \ge \frac{|\mathcal{C}_s|}{2c}\sum_{i\in J}\Delta_i$. Thus $c^{-1}\sum_{i\in J}\Delta_i \le 2\Delta_G$, and coverage gives the additional bound by ξ. These bounds hold for each permutation, so averaging the active contribution gives $\bar\gamma^A \le b$.

For either graph class and each realized set J, at least $c-1$ members are suboptimal. Coverage gives $\Delta_i \le \xi$ for each of them. The set $J \setminus \{i^*\}$ is induced acyclic, so it is admissible in the definition of $H_G^{\mathrm{ac}}(\Delta)$ and $H_G^{\mathrm{ac}}(\Delta) \ge \sum_{i\in J\setminus\{i^*\}}\Delta_i^{-1} \ge (c-1)/\xi$. If at least two active arms remain, an active suboptimal singleton also shows $H_G^{\mathrm{ac}}(\Delta)\xi \ge 1$, even when $J = \{i^*\}$. Consequently $c \le 1 + H_G^{\mathrm{ac}}(\Delta)\xi \le 2H_G^{\mathrm{ac}}(\Delta)\xi$. The realized set is induced acyclic, hence $c \le \rho \le 16 H_G^{\mathrm{ac}}(\Delta)\beta$, where the directed inequality uses Lemma 15 and the undirected inequality follows from the floor in (38). Combining these bounds gives $c \le 16 H_G^{\mathrm{ac}}(\Delta)\min\{\beta,\xi\}$. □

D.3 Regret within an epoch

Fix a covered history before the fresh permutation. Lemma 16 bounds the active regret after averaging over that permutation, while its exploration-size bound holds for every realized permutation. Thus the realized observation weights satisfy the hypothesis of the deterministic delayed-summation lemma on each complete good path. The conditional permutation mean is taken before this pathwise application; it is not conditioned on the future concentration event.

Corollary 17 (Local stochastic regret)**.** *For any covered initial intervals, put $b_0 = \min\{\beta, \xi_0\}$ and use $\bar G_n$ as defined in Appendix C. On the complete concentration event of Lemma 9, together with incoming and within-epoch mean coverage, Proposition 10 rules out a test failure. On this event, for every evaluated $n \le N$,*

$$\bar G_n \le C H_G^{\mathrm{ac}}(\Delta)\lambda\log(en) + 2b_0(1 + \sigma_{\max}(n)). \tag{40}$$

Proof. If $b_0 = 0$, covered intervals imply that the active contribution vanishes; (35) bounds exploration of eliminated arms. Otherwise, let $m \le n$ be the last decision with at least two active arms, if one exists. Lemma 16 bounds each active term, averaged over π_s, by b_{s-1} and the realized c_s by $16H_G^{\mathrm{ac}}(\Delta)b_{s-1}$ through m, where $b_t = \min\{\beta, \xi_t\}$. Nested intervals make b_t nonincreasing, and Lemma 12 gives $b_t^2 \le 64\lambda/X_t$ when $X_t > 0$. With the realized $\omega_s = (4c_s)^{-1}$, the size bound implies $64H_G^{\mathrm{ac}}(\Delta)\omega_s b_{s-1} \ge 1$. Apply Lemma 14 through m and add the harmonic sum of the

realized bound (35) on the exploration of eliminated arms through n; its average over π_s obeys the same bound. After m, active regret vanishes because the remaining active arm is optimal on coverage. If no such m exists, only the exploration of eliminated arms contributes. □

D.4 Interval recomputation and regret across epochs

Use $N_0 = 2K$, $N_{m+1} = N_m^2$, and $t_m = 1 + \sum_{j=0}^{m-1} N_j$. For an evaluation time T, let $m_T = \max\{m : t_m \le T\}$ be the last epoch started by T. For epoch m, let $\xi_{m,0}$ and $\mathcal{A}_{m,1}$ be the threshold and the active set computed from its incoming intervals. Set $\lambda_m = \log(16KN_m^5) + \lambda_{\text{off}}$, with $\lambda_{\text{off}} = 0$ for the default rule and $\lambda_{\text{off}} = \log(64/\delta)$ for the high-probability version with parameter $\delta \in (0,1)$; the offset λ_{off} is constant across epochs. For a fixed evaluation time T, let $n_m = \min\{N_m, T - t_m + 1\}$ for started epochs. The peaks of pending sources generated within each epoch, measured over its evaluated prefix, are $\sigma_{\max}^{(m)} \le \min\{N_m, \sigma_{\max}\}$, and $\lambda_* = \max_{t_m \le T} \lambda_m = O(\log(eKT) + \lambda_{\text{off}})$.

Coverage and accuracy at a new epoch. Use the interval-recomputation rule of Appendix B. Restarting each epoch from $[0,1]^K$ would leave $b_0 = \min\{\beta, 1\}$ in Corollary 17 and hence a delay term $2\min\{\beta,1\}(1 + \sigma_{\max}^{(m)})$ in every epoch; reusing the raw sums instead makes the initial level small. Coverage at the new confidence level holds for the retained samples even if the old gap estimates were inaccurate. A quantitative bound on the new threshold additionally uses coverage in the preceding epoch: an arm that is active at the new threshold $u = \xi_{m,0}$ was observed at each old source time s at rate at least $\min\{(4\rho)^{-1}, \kappa\lambda_{m-1}/(sg_i^2)\} \ge \min\{(4\rho)^{-1}, \kappa\lambda_{m-1}/(su^2)\}$, since $g_i \le \Delta_i \le u$. Summing the second term harmonically over s then gives the arm enough observations to force $u = O(\sqrt{\rho\lambda_m/N_{m-1}})$ (Lemma 18).

Lemma 18 (Accuracy of recomputed intervals)**.** *Under i.i.d. losses, with probability at least* $1 - 2KN_{m-1}e^{-\lambda_m}$*, every interval recomputed at the start of epoch* m *contains* μ_i*, whether or not the old gap estimates were correct. Suppose the preceding epoch of length* $M = N_{m-1}$ *had no test failure, at most* $\bar\sigma$ *of its sources were pending at its end,* $M \ge 4(1+\bar\sigma)$*, and the preceding lower-count and coverage events and the new interval coverage event hold. Assume also* $\lambda_{m-1} \le \lambda_m \le 2\lambda_{m-1}$*, as for the stated schedule, where* $\lambda_m = \log(16K) + 10\log N_{m-1} + \lambda_{\text{off}}$*. Then*

$$\xi_{m,0} \le c_{\text{init}}\sqrt{\frac{\rho\lambda_m}{M}}, \quad \textit{where} \quad c_{\text{init}}^2 = 1280. \tag{41}$$

If $M < 4(1+\bar\sigma)$*, use only* $0 \le \xi_{m,0} \le 1$.

Proof. Coverage at the new confidence level. The retained (W_i, n_i) are the statistics $(W_{M,i}, n_{M,i})$ of Lemma 9 for the previous epoch at its endpoint $t = M$. The recomputed radius is $\sqrt{\lambda_m/(2n_i)}$, so the event that some recomputed interval misses μ_i is contained in the union over arms and deterministic counts $n \le M$ of the Hoeffding events of Lemma 9 at level λ_m. Each has probability at most $2e^{-\lambda_m}$, so the union costs $2KMe^{-\lambda_m} \le 2KN_m e^{-\lambda_m}$. Together with Lemma 9, the total is at most $(6KN_m^2 + 3KN_m)e^{-\lambda_m} \le 16KN_m^2 e^{-\lambda_m} = e^{-\lambda_{\text{off}}}N_m^{-3}$. This argument does not assume accurate old gap estimates.

Contraction of the initial threshold. For accuracy let $\lambda = \lambda_{m-1}$, $\lambda' = \lambda_m$, so $\lambda \le \lambda' \le 2\lambda$, and write $u = \xi_{m,0}$. If $u \le \sqrt{1280\rho\lambda'/M}$ there is nothing to prove. Otherwise $u \le 1$ gives $M > 1280\rho\lambda' \ge 1280$. For every newly active arm i, old mean coverage gives $g_i \le \Delta_i$ for the previous final gap estimate g_i, and new mean coverage and (23) give $\Delta_i \le \overline{\Delta}_i' \le u$. Monotonicity of the gap estimates gives, at each old source time s,

$$Q_i(p_s) \ge \min\{(4\rho)^{-1}, \kappa\lambda/(su^2)\}.$$

This lower bound is nonincreasing in s, so removing at most $\bar\sigma$ pending sources is worst when the earliest $\bar\sigma$ are removed. Put $z = 320\rho\lambda/u^2$ and $k = \max\{\bar\sigma + 1, \lceil z \rceil\}$. Since $z < M/4$ and $\bar\sigma + 1 \le M/4$, we have $k \le M/4 + 1$. For $s \ge k$ the second term attains the minimum, so, with $\kappa = 80$,

$$\sum_{s \in O_M} Q_i(p_s) \ge \frac{80\lambda}{u^2}\sum_{s=k}^{M}\frac{1}{s} \ge \frac{80\lambda}{u^2}\log\frac{M+1}{k}.$$

Since $M > 1280$, $(M+1)/k \geq 4(M+1)/(M+4) > 19/5 > e^{4/3}$. The lower-count event (27) of the preceding epoch, at its old level $\lambda = \lambda_{m-1}$ and $t = M$, and $u \leq 1$ therefore give

$$n_i u^2 \geq \tfrac{320}{3}(1-e^{-1})\lambda - \lambda u^2 \geq \big(\tfrac{320}{3}(1-e^{-1}) - 1\big)\lambda > 64\lambda,$$

where the last step uses $1 - e^{-1} > 5/8$. In particular $n_i > 0$, and the new interval of i has width $w_i' \leq \sqrt{2\lambda'/n_i}$. Apply the deterministic width bound in (24) to the new nonempty box and take an active arm of maximal width: $u^2 \leq 16 w_i'^2 \leq 32\lambda'/n_i$, so $n_i u^2 \leq 32\lambda' \leq 64\lambda$. This contradiction proves (41). The same conclusion holds if $u = 0$. □

This proof uses old coverage for $g_i \leq \Delta_i$ and new coverage for the recomputed endpoints. The observation lower bound holds for every realized permutation, and the received sources need not form a prefix.

Good epochs. Let $\mathcal{E}_m$ contain the current epoch's concentration event and mean coverage by the incoming intervals at confidence level λ_m. These intervals are recomputed from the previous epoch's retained samples when available, and are $[0,1]^K$ initially or after a test failure. Lemma 9 and the incoming-coverage argument above give an unconditional failure probability at most $e^{-\lambda_{\mathrm{off}}} N_m^{-3}$. Let $\mathcal{W}_T$ consist of the epochs started by T for which $\mathcal{E}_m$ holds and, when $m \geq 1$, $\mathcal{E}_{m-1}$ also holds.

Summing the delay-dependent terms. Corollary 17 charges epoch m the delay term $2\varepsilon_m(1 + \sigma_{\max}^{(m)})$ with $\varepsilon_m = \min\{\beta, \xi_{m,0}\}$. Once $N_{m-1} \geq 4(1+\sigma_{\max})$, Lemma 18 gives $\varepsilon_m = O(\sqrt{\rho\lambda_m/N_{m-1}})$, and since $N_{m+1} = N_m^2$, these square-root denominators grow at least twofold from one epoch to the next, so the delay terms of the long epochs form a geometric series and the delay cost is not paid per epoch.

Lemma 19 (Weighted pending-feedback sum)**.** *Let $N_0 \geq 4$, $N_{m+1} \geq 4N_m$, $\sigma_{\max}^{(m)} \leq \min\{N_m, \bar{\sigma}\}$, and $0 \leq \varepsilon_m \leq \varepsilon_{\max}$ with $\varepsilon_{\max} > 0$. Suppose, for $m \geq 1$,*

$$N_{m-1} \geq 4(1+\bar{\sigma}) \quad \Longrightarrow \quad \varepsilon_m \leq \sqrt{A/N_{m-1}}.$$

For every finite collection $\mathcal{J}$ of evaluated epochs,

$$\sum_{m\in\mathcal{J}} \varepsilon_m(1+\sigma_{\max}^{(m)}) \leq 10\varepsilon_{\max}(1+\bar{\sigma}) + \frac{2A}{\varepsilon_{\max}}. \tag{42}$$

It also holds for any subcollection satisfying the displayed hypotheses.

Proof. Set $M_0 = \max\{4(1+\bar{\sigma}), A/\varepsilon_{\max}^2\}$. The lengths $N_m \leq M_0$ sum to at most $4M_0/3$. Since $N_m \geq 4$, their contribution is at most $5\varepsilon_{\max}M_0/3$. There is at most one first epoch of length larger than M_0, contributing at most $\varepsilon_{\max}(1+\bar{\sigma})$. Every later epoch has $N_{m-1} > M_0 \geq 4(1+\bar{\sigma})$ and contributes at most $(1+\bar{\sigma})\sqrt{A/N_{m-1}}$. The square-root denominators grow by at least a factor two, so these terms sum to at most $2(1+\bar{\sigma})\sqrt{A/M_0} \leq 2\varepsilon_{\max}(1+\bar{\sigma})$. Finally use $M_0 \leq 4(1+\bar{\sigma}) + A/\varepsilon_{\max}^2$, which gives $\frac{29}{3}\varepsilon_{\max}(1+\bar{\sigma}) + \frac{5}{3}A/\varepsilon_{\max}$, and enlarge the constants to 10 and 2. □

For every graph class set $\varepsilon_m = \min\{\beta, \xi_{m,0}\}$; then $\varepsilon_m \leq 1$ because $\xi_{m,0} \leq 1$. The definition of β and Lemma 15 give $\rho/\beta \leq 16 H_G^{\mathrm{ac}}(\Delta)$. Apply the generic pending-feedback sum lemma with $\varepsilon_{\max} = \beta$. For an epoch in $\mathcal{W}_T$ with $m \geq 1$, at most $\sigma_{\max}$ of its previous sources are pending at its end, so Lemma 18 applies with $\bar{\sigma} = \sigma_{\max}$ if $N_{m-1} \geq 4(1+\sigma_{\max})$. With $A = c_{\mathrm{init}}^2 \rho\lambda_* = 1280\rho\lambda_*$, every $m \in \mathcal{W}_T$ therefore satisfies the hypothesis of Lemma 19, by Lemma 18 when $N_{m-1} \geq 4(1+\sigma_{\max})$ and vacuously otherwise. Applying Lemma 19 to $\mathcal{J} = \mathcal{W}_T$ gives

$$\sum_{m\in\mathcal{W}_T} \varepsilon_m(1+\sigma_{\max}^{(m)}) \leq C\beta(1+\sigma_{\max}) + C\frac{\rho\lambda_*}{\beta} \leq C\beta(1+\sigma_{\max}) + C H_G^{\mathrm{ac}}(\Delta)\lambda_*. \tag{43}$$

Epoch schedule bounds. For the actual squaring schedule the remaining elementary relations are

$$\lambda_* = O(\log(eKT) + \lambda_{\mathrm{off}}), \quad \sum_{m=0}^{m_T} \log(en_m) = O(\log(eT)). \tag{44}$$

Completed epoch logarithms form a geometric sequence; completed lengths grow by at least four. Add the last partially evaluated epoch separately. Its length is at most T^2, unless it is the first, which is bounded in terms of K.

Proof of the global stochastic bound.

Proof of the stochastic conclusions in Theorems 1 and 5 and Proposition 3. For stochastic losses, $\mathcal{E}_m$ already prevents a test failure regardless of previous gap-estimate accuracy. Quantitative incoming accuracy additionally requires $\mathcal{E}_{m-1}$. On $\mathcal{W}_T$ apply Corollary 17, whose b_0 equals ε_m in epoch m, followed by (43) and (44). This bounds $\sum_{m\in\mathcal{W}_T}\bar{G}_{n_m}$ pathwise on complete good epoch histories by $C[H_G^{\mathrm{ac}}(\Delta)\lambda_*\log(eT)+\beta\sigma_{\max}+1]$. On every decision, including fallback rounds, define $\bar{\gamma}_s=\mathbb{E}[\langle p_s,\Delta\rangle\mid\mathcal{F}_{s-1}]$; the fallback distribution is fixed by the pre-decision history, so this equals $\langle p_s,\Delta\rangle$ there. On each remaining epoch $0\le\bar{\gamma}_s\le 1$, so its contribution to $\mathbb{E}\sum_s\bar{\gamma}_s$ is at most

$$\sum_{m=0}^{m_T}N_m\mathbb{P}(\mathcal{E}_m^c)+\sum_{m=1}^{m_T}N_m\mathbb{P}(\mathcal{E}_{m-1}^c)\le e^{-\lambda_{\mathrm{off}}}\Big(\sum_{m=0}^{\infty}N_m^{-2}+\sum_{m=1}^{\infty}N_{m-1}^{-1}\Big)$$
$$=O(e^{-\lambda_{\mathrm{off}}}),$$

since $N_m=N_{m-1}^2$, $N_0\ge 4$ and the lengths grow at least fourfold. For the default rule, $\lambda_{\mathrm{off}}=0$, so $\lambda_*=O(\log(eKT))$ and the failure contribution is $O(1)$. For undirected graphs $\beta=\Delta_G^{\pi}\le 2\Delta_G$, which gives Theorem 1, and Proposition 3 follows since $\Delta_G^{\pi}\le 2\Delta_{\mathrm{avg}}$ for the self-loop-only and complete graphs (Appendix D.2). Taking unconditional expectations and using the tower property gives $R_T=\mathbb{E}\sum_{s=1}^{T}\langle p_s,\Delta\rangle=\mathbb{E}\sum_{s=1}^{T}\bar{\gamma}_s$.

□

E HIGH-PROBABILITY STOCHASTIC REGRET

For the high-probability version with parameter $\delta\in(0,1)$, take $\lambda_m=\lambda_{\delta,m}:=\log(16KN_m^5)+\log(64/\delta)$ in intervals, sampling and testing, including interval recomputation. Here s and t index global rounds, $\mathcal{F}_{s-1}$ is the global pre-permutation history, and $\bar{G}_T=\sum_{s=1}^{T}\bar{\gamma}_s$ is the cumulative sum over all rounds up to T, including fallback rounds, whereas $\bar{G}_t$ in Appendix C is a sum within one epoch. With $\lambda_{\mathrm{off}}=\log(64/\delta)$, each epoch failure event has probability $\mathbb{P}(\mathcal{E}_m^c)\le e^{-\lambda_{\mathrm{off}}}N_m^{-3}=(\delta/64)N_m^{-3}$. Since $N_m\ge 4^{m+1}$,

$$\mathbb{P}\Big(\bigcup_{m\ge 0}\mathcal{E}_m^c\Big)\le\frac{\delta}{64}\sum_{m\ge 0}64^{-(m+1)}=\frac{\delta}{64\cdot 63}\le\frac{\delta}{4}.$$

On $\bigcap_m\mathcal{E}_m$ no test fails, every started epoch lies in $\mathcal{W}_T$, and $\bar{G}_T=\sum_{m\le m_T}\bar{G}_{n_m}$ with the within-epoch sums on the right. The pathwise bound in the proof of the global stochastic bound in Appendix D.4 then applies for every T, with $\lambda_*=O(\log(eKT/\delta))$. This gives, on one event of probability at least $1-\delta/4$, simultaneously for all $T\ge 2$,

$$\bar{G}_T\le C[H_G^{\mathrm{ac}}(\Delta)\log(eKT/\delta)\log(eT)+\beta\sigma_{\max}+1].$$

For MAB and complete feedback, $\beta=\Delta_G^{\pi}\le 2\Delta_{\mathrm{avg}}$ under the same data-reuse rule.

The conversion to realized regret uses that the conditional second moment of the excess loss against i^* is at most $\bar{\gamma}_s/\Delta_{\min}$, proportional to its mean. This gives a Bernstein-type supermartingale whose deviation cost is an additive $\log(4/\delta)/\Delta_{\min}$ term, at most $H_G^{\mathrm{ac}}(\Delta)\log(4/\delta)$, instead of a $\sqrt{T}$ term. For actual losses put $Y_s=\ell_{s,I_s}-\ell_{s,i^*}$ and $\bar{\gamma}_s=\mathbb{E}[\langle p_s,\Delta\rangle\mid\mathcal{F}_{s-1}]$. Conditional on $\mathcal{F}_{s-1}^{+}$, the mean of Y_s is $\langle p_s,\Delta\rangle$ and its second moment is at most $\langle p_s,\Delta\rangle/\Delta_{\min}$. Averaging over π_s gives

$$\mathbb{E}_{s-1}Y_s=\bar{\gamma}_s\quad\text{and}\quad\mathbb{E}_{s-1}Y_s^2\le\mathbb{E}[1-p_{s,i^*}\mid\mathcal{F}_{s-1}]\le\bar{\gamma}_s/\Delta_{\min}.$$

Using $e^x\le 1+x+x^2$ for $|x|\le 1$ gives $\mathbb{E}_{s-1}e^{\Delta_{\min}Y_s}\le 1+2\Delta_{\min}\bar{\gamma}_s\le e^{2\Delta_{\min}\bar{\gamma}_s}$. Hence $\exp\{\Delta_{\min}(\sum_{s=1}^{t}Y_s-2\bar{G}_t)\}$ is a nonnegative supermartingale. Its maximal inequality bounds $\sum_{s=1}^{T}Y_s$ by $2\bar{G}_T+\log(4/\delta)/\Delta_{\min}$ for all T, outside probability $\delta/4$.

Separately, for each arm i put $X_{s,i}=\ell_{s,i^*}-\ell_{s,i}$, so that $X_{s,i^*}=0$ and the realized regret decomposes as

$$\sum_{s=1}^{T}\ell_{s,I_s}-\min_i\sum_{s=1}^{T}\ell_{s,i}=\sum_{s=1}^{T}Y_s+\max_i\sum_{s=1}^{T}X_{s,i},$$

where the term $i = i^*$ of the maximum is 0. For $i \neq i^*$, the mean of $X_{s,i}$ is $-\Delta_i$ and $|X_{s,i}| \leq 1$, so $\mathbb{E}_{s-1} e^{\Delta_i X_{s,i}} \leq 1 - \Delta_i^2 + \Delta_i^2 \mathbb{E}_{s-1} X_{s,i}^2 \leq 1$. Thus $\exp\{\Delta_i \sum_{s=1}^{t} X_{s,i}\}$ is a second nonnegative supermartingale. Its maximal inequality at level $4K/\delta$ and a union bound over the at most $K-1$ arms $i \neq i^*$ bound $\max_i \sum_{s=1}^{T} X_{s,i}$ by $\log(4K/\delta)/\Delta_{\min}$ for all T, outside probability $\delta/4$. The three exceptional events, for the epochs, for Y and for X, each have probability at most $\delta/4$, so all bounds hold simultaneously with probability at least $1 - 3\delta/4 \geq 1 - \delta$. Since $1/\Delta_{\min} \leq H_G^{\mathrm{ac}}(\Delta)$, the two logarithmic terms are absorbed. This proves Corollary 6; see Howard et al. (2021) for the standard maximal-inequality framework.

The same high-probability version of the algorithm has expected adversarial regret

$$O(\sqrt{\rho T \log(eKT)} + \sqrt{\rho T \log(64/\delta)} + \Psi_T + 1).$$

No high-probability adversarial bound or arbitrary-δ guarantee for the fixed $\lambda_m = \log(16KN_m^5)$ implementation is asserted.

F Adversarial regret

The proof splits each epoch at the first test failure τ. Up to τ, the last passed test (Proposition 10) and the upper-confidence sum (Proposition 13) bound the regret by confidence terms and the peak pending count $\sigma_{\max}(n)$; this uses count and excess-loss concentration but not coverage of a stochastic mean, so the initial intervals may be any nonempty box chosen from past observations. After τ, Lemma 20 shows that the remaining rounds form a fresh run of the independently initialized subroutine, controlled by Proposition 22, proved in Appendix G. The peak $\sigma_{\max}(n)$ is then converted into the skipping form of the delay term, since k sources outside U pending at the same time have distinct ages and hence truncated delays summing to at least $k(k-1)/2$.

F.1 Regret before and after test failure

For a fixed arm i, let $\mathrm{Reg}_n(i) = \sum_{s \leq n} (\ell_{s,I_s} - \ell_{s,i})$ be the regret against i over the first n rounds of an epoch. The fallback is analyzed through a restart at the first failure. Fix the pre-epoch history, an oblivious loss sequence and fixed delays. Realize the randomization of the epoch by independent variables: $\Xi_t = (\pi_t, \psi_t)$ for statistical round t, with ψ_t uniform on $[0,1]$ and $I_t = \iota(p_t, \psi_t)$ for the map ι of Lemma 24, and $\vartheta_1, \vartheta_2, \ldots$ uniform on $[0,1]$, where ϑ_u drives the draw in local round u of Algorithm 2. Let τ be the first round whose test fails, counting an empty interval as a failure, with $\tau = \infty$ if no test fails. With the randomization realized in this way and the losses fixed, the history $\mathcal{F}_t$ of Appendix C.2 through round t is $\mathcal{F}_t = \sigma(\Xi_1, \ldots, \Xi_t, (\vartheta_u \mathbf{1}\{\tau + u \leq t\})_{u \geq 1})$, so the two descriptions coincide.

Lemma 20 (Restart at the first test failure)**.** *In this setting, τ is a stopping time for $(\mathcal{F}_t)_{t \geq 0}$. Fix a deterministic $n \leq N$, an arm i and $U \subseteq [n]$. On $\{\tau < n\}$, conditional on $\mathcal{F}_\tau$, rounds $\tau + 1, \ldots, n$ are the first $n' = n - \tau$ rounds of Algorithm 2 started from the empty state on the losses $\ell'_{s'} = \ell_{\tau+s'}$ and delays $d'_{s'} = d_{\tau+s'}$, with seeds $\vartheta_1, \ldots, \vartheta_{n'}$ independent of $\mathcal{F}_\tau$; no feedback of rounds $s \leq \tau$ enters this run. Consequently, with $U' = \{s - \tau : s \in U,\ \tau < s \leq n\}$, on $\{\tau < n\}$,*

$$\mathbb{E}\Big[\sum_{s=\tau+1}^{n} (\ell_{s,I_s} - \ell_{s,i}) \,\Big|\, \mathcal{F}_\tau \Big] \leq 30 \Big[\sqrt{\rho n' \log K} + |U'| + 1 + \sqrt{\log K \sum_{s \in (\tau, n] \setminus U} \min\{d_s, n - s\}} \Big]. \tag{45}$$

Proof. Stopping time. Before the first failure, the state of Algorithm 1 after round t, consisting of its counters, sums, intervals, gap estimates, test statistics, and the stored records of rounds $s \leq t$ (the action I_s, the observation probabilities $Q_j(p_s)$ of its revealed coordinates j, and a_s), is a deterministic function of the state after round $t-1$, the fixed losses and delays, π_t and I_t. Here p_t is a function of the state after round $t-1$ and π_t, and $I_t = \iota(p_t, \psi_t)$. By induction over t, the event $\{\tau \geq t\}$ and, on it, the state after round $t-1$ are functions of $\Xi_1, \ldots, \Xi_{t-1}$. Hence $\{\tau = t\} = \{\tau \geq t\} \cap \{\text{the test fails at } t\}$ is $\sigma(\Xi_1, \ldots, \Xi_t)$-measurable, so $\{\tau = t\} \in \mathcal{F}_t$ and τ is a stopping time.

The history at τ. Let $A \in \mathcal{F}_\tau$ and $t < n$. Then $A \cap \{\tau = t\} \in \mathcal{F}_t$. On $E = \{\tau = t\}$, every generator $\vartheta_u \mathbf{1}\{\tau + u \leq t\}$ of $\mathcal{F}_t$ vanishes, and $E \in \sigma(\Xi_1, \ldots, \Xi_t)$. The sets $C \in \mathcal{F}_t$ with

$C \cap E \in \sigma(\Xi_1, \dots, \Xi_t)$ form a σ-field that contains $\sigma(\Xi_1, \dots, \Xi_t)$ and the preimages under each vanishing generator, which meet E in E or $\varnothing$. Hence $A \cap \{\tau = t\} \in \sigma(\Xi_1, \dots, \Xi_t)$, which is independent of $(\vartheta_u)_{u \geq 1}$.

The restarted run. On $\{\tau = t\}$, Algorithm 1 returns after round t and runs Algorithm 2 on rounds $t+1, \dots, N$. Algorithm 2 initializes $L^{\mathrm{obs}} = 0$, $\Sigma_0 = V_0 = B_0 = 0$ and $\mathcal{P} = \varnothing$, and registers only its own sources in line 37, so no feedback of rounds $s \leq t$ enters it. Its local round u is round $t+u$ of the epoch, with loss vector ℓ_{t+u}, delay d_{t+u} and seed ϑ_u, and its first $n-t$ rounds do not depend on when it is stopped. Since its other inputs are deterministic, there are deterministic maps $\Gamma_u^{(t)}$ with $I_{t+u} = \Gamma_u^{(t)}(\vartheta_1, \dots, \vartheta_u)$ on $\{\tau = t\}$. Let $R^{(t)} = \sum_{u=1}^{n-t} (\ell_{t+u,\Gamma_u^{(t)}(\vartheta_1,\dots,\vartheta_u)} - \ell_{t+u,i})$, a function of $(\vartheta_u)_{u \geq 1}$ alone. For $A \in \mathcal{F}_\tau$, the previous step gives

$$\mathbb{E}\Big[\mathbf{1}_A \mathbf{1}\{\tau = t\} \sum_{s=\tau+1}^{n} (\ell_{s,I_s} - \ell_{s,i})\Big] = \mathbb{E}\big[\mathbf{1}_{A \cap \{\tau = t\}} R^{(t)}\big] = \mathbb{P}(A \cap \{\tau = t\})\, \mathbb{E}R^{(t)}.$$

Since $\{\tau = t\} \in \mathcal{F}_\tau$, the conditional expectation in (45) equals $\mathbb{E}R^{(t)}$ on $\{\tau = t\}$. This is the expected regret $\mathbb{E}\,\mathrm{Reg}_{n'}^{\mathrm{DF}}(i)$ of Algorithm 2 over its first $n' = n - t$ rounds, from the empty state with independent randomization, on the oblivious losses ℓ' and fixed delays d'. Proposition 22 applies with the deterministic n' and $U' \subseteq [n']$. For $s = t + s'$ its truncated delays are $\min\{d'_{s'}, n' - s'\} = \min\{d_s, n - s\}$, and $[n'] \setminus U'$ corresponds to $(t, n] \setminus U$. This gives (45) on $\{\tau = t\}$ for every $t < n$. □

Proposition 21 (Regret within one epoch). *Fix $N \geq 2K$, $\lambda \geq \log(16KN^5)$, and any nonempty initial interval box selected before the epoch. For every deterministic $n \leq N$, fixed arm i, and $U \subseteq [n]$, the complete epoch satisfies*

$$\mathbb{E}\,\mathrm{Reg}_n(i) \leq C\big[\sqrt{\rho n \lambda} + |U| + 1 + \sqrt{\log K\, D_n(U)}\big], \quad \textit{where} D_n(U) = \sum_{s \in [n] \setminus U} \min\{d_s, n - s\}. \tag{46}$$

Under oblivious losses and delays this holds conditionally on the pre-epoch history and uniformly over the selected nonempty box. The box need not cover a stochastic mean.

Proof. Let τ be the first test failure. The conditional count and excess-loss event of Lemma 9 fails with probability at most $e^{-(\lambda - \log(16KN^5))} N^{-3} \leq N^{-3}$. On this event, let $t' = (\tau - 1) \wedge n$ be the last time at which the test was passed. The bound (30) of Proposition 10 at t', the confidence sum (36) of Proposition 13 with t' in place of n, the width bound $h_{t'}(v_{t'}) \leq 8\sqrt{\rho n \lambda} + 16\rho\lambda$, which follows from $\widehat{\rho}_{t'} \leq \rho$ and $v_{t'} \leq n$, and $\sigma_{t'} \leq \sigma_{\max}(n)$ give

$$M_{t'}(i) \leq C(\sqrt{\rho n \lambda} + \rho\lambda + \sigma_{\max}(n) + 1).$$

If $\tau \leq n$, the failing round τ adds $r_{\tau,i} \leq 1$, so $M_{\tau \wedge n}(i) \leq C(\sqrt{\rho n \lambda} + \rho\lambda + \sigma_{\max}(n) + 1) + 1$ on this event. On its complement use $M_{\tau \wedge n}(i) \leq n$. By its probability, this contributes at most $nN^{-3} \leq 1$ to the expectation, which is absorbed into C. Since $\tau \wedge n$ is a bounded stopping time, optional stopping in the form of Lemma 11 gives $\mathbb{E}\,\mathrm{Reg}_{\tau \wedge n}(i) = \mathbb{E}M_{\tau \wedge n}(i)$, and therefore

$$\mathbb{E}\,\mathrm{Reg}_{\tau \wedge n}(i) \leq C(\sqrt{\rho n \lambda} + \rho\lambda + \sigma_{\max}(n) + 1) + 1.$$

Fix U before taking expectations. Let $t \leq n$ be a peak of the pending count, and let k be the number of sources pending at t outside U, so $\sigma_{\max}(n) \leq |U| + k$. These sources have distinct nonnegative ages $t - s$, whose sum is at least $k(k-1)/2$. Each of them is pending at $t \leq n$, so $d_s > t - s$ and $n - s \geq t - s$, hence $\min\{d_s, n - s\} \geq t - s$. Thus $k(k-1) \leq 2D_n(U)$, and for $k \geq 1$, $(k-1)^2 \leq k(k-1) \leq 2D_n(U)$ gives $k \leq 1 + \sqrt{2D_n(U)}$. Therefore

$$\sigma_{\max}(n) \leq |U| + 1 + \sqrt{2D_n(U)}. \tag{47}$$

By Lemma 20, on $\{\tau < n\}$ the conditional expected regret of rounds $\tau+1, \dots, n$ given $\mathcal{F}_\tau$ is at most $30[\sqrt{\rho n \log K} + |U| + 1 + \sqrt{\log K\, D_n(U)}]$, since $n - \tau \leq n$, $|U'| \leq |U|$, and the truncated-delay sum in (45) is at most $D_n(U)$. Write $\mathrm{Reg}_n(i) = \mathrm{Reg}_{\tau \wedge n}(i) + \mathbf{1}\{\tau < n\} \sum_{s=\tau+1}^{n} (\ell_{s,I_s} - \ell_{s,i})$, where $\{\tau < n\} \in \mathcal{F}_\tau$. Taking the tower expectation and using $\lambda \geq \log K$ and $\log K \geq \log 2$ gives

$$\mathbb{E}\,\mathrm{Reg}_n(i) \leq C(\sqrt{\rho n \lambda} + \rho\lambda + |U| + 1 + \sqrt{\log K\, D_n(U)}).$$

Set $a = \rho\lambda$. If $n < a$, use $\mathbb{E}\,\mathrm{Reg}_n(i) \le n \le \sqrt{an}$; otherwise $a \le \sqrt{an}$. This removes the initialization term for the complete epoch and proves the claim. □

The coefficient $\log K$ in the delay term is independent of the statistical confidence parameter. Only the graph term changes when a confidence offset is prescribed.

F.2 Regret across epochs

Proof of the adversarial conclusions in Theorems 1 and 5. Fix the comparator arm i and one global set $U \subseteq [T]$ before taking expectations. Let m_T be the last epoch started by T, and let $\mathcal{T}_m$ and $n_m = |\mathcal{T}_m|$ denote the evaluated rounds and length of epoch m. Only $\mathcal{T}_{m_T}$ can be a strict prefix of its capacity. If $m_T \ge 2$, the earlier epochs have total length

$$T_{\mathrm{early}} = \sum_{m=0}^{m_T-2} N_m \le \tfrac{4}{3} N_{m_T-2} = \tfrac{4}{3}\sqrt{N_{m_T-1}} < \tfrac{4}{3}\sqrt{T}.$$

The sum bound follows by induction from $N_0 = 2K \ge 4$ and $N_{m+1} = N_m^2$: if the preceding sum is at most $\frac{4}{3}N_{q-1}$, then $N_q + \frac{4}{3}\sqrt{N_q} \le \frac{4}{3}N_q$. Set $T_{\mathrm{early}} = 0$ when $m_T < 2$. Since both the learner's loss and the fixed comparator's loss lie in $[0,1]$, regret on these early rounds is at most T_{early} pathwise.

Retain the epochs $\mathcal{J} = \{\max(0, m_T - 1), \dots, m_T\}$, at most two in number. For $m \in \mathcal{J}$, restrict U to $\mathcal{T}_m$ in local coordinates and call this set U_m; write $D_m(U_m)$ for the delay sum truncated at the end of that evaluated prefix. Their source sets are disjoint, and local truncation cannot exceed truncation at T, so

$$\sum_{m\in\mathcal{J}} |U_m| \le |U|, \quad \sum_{m\in\mathcal{J}} D_m(U_m) \le D_T(U), \quad \text{and} \quad \sum_{m\in\mathcal{J}} n_m \le T.$$

Every retained capacity satisfies $\log(eKN_m) \le 2\log(eKT)$. Indeed, for $m_T \ge 1$ starting epoch m_T requires $T > N_{m_T-1}$, hence $N_{m_T} = N_{m_T-1}^2 < T^2$; for $m_T = 0$, $N_0 = 2K$ and $\log(2eK^2) \le 2\log(eKT)$ for $T \ge 1$. Thus, with $\lambda_{\mathrm{off}} = 0$ for the default rule or $\lambda_{\mathrm{off}} = \log(64/\delta)$ for the confidence-offset rule, each retained $\lambda_m = \log(16KN_m^5) + \lambda_{\mathrm{off}} \le 5\log(eKN_m) + \lambda_{\mathrm{off}} \le 10\log(eKT) + \lambda_{\mathrm{off}}$.

The lengths n_m and sets U_m are fixed by the schedule, horizon, and the globally fixed U. Condition separately on each retained epoch's initial history, apply Proposition 21 to its deterministic evaluated prefix, and then take expectations. The same comparator i is used in every epoch. Cauchy–Schwarz and the preceding bounds give

$$\sum_{m\in\mathcal{J}} \sqrt{\rho n_m \lambda_m} \le \sqrt{2\rho T(10\log(eKT) + \lambda_{\mathrm{off}})} \le \sqrt{20\rho T\log(eKT)} + \sqrt{2\rho T\lambda_{\mathrm{off}}},$$

$$\sum_{m\in\mathcal{J}} \sqrt{\log K\, D_m(U_m)} \le \sqrt{2\log K\, D_T(U)}.$$

The cardinality terms sum to at most $|U|$ and the unit terms to at most 2. Combining the early pathwise bound with the conditional prefix bounds, for every fixed global U we obtain

$$\mathbb{E}\,\mathrm{Reg}_T(i) \le \tfrac{4}{3}\sqrt{T} + C\big[\sqrt{20\rho T\log(eKT)} + \sqrt{2\rho T\lambda_{\mathrm{off}}} + |U| + 2 + \sqrt{2\log K\, D_T(U)}\big].$$

Now minimize over global U after taking expectations. Since $\rho \ge 1$ and $\log(eKT) \ge 1$, the early term is absorbed into the graph term, while $|U| + \sqrt{2\log K\, D_T(U)} \le \sqrt{2}\big(|U| + \sqrt{\log K\, D_T(U)}\big)$. The resulting bound is $O(\sqrt{\rho T\log(eKT)} + \sqrt{\rho T\lambda_{\mathrm{off}}} + \Psi_T + 1)$, which gives the stated default and confidence-offset adversarial conclusions. □

G Delay-adaptive exponential weights with skipping

This section proves Proposition 22 for the adaptive fallback. Fix an oblivious loss sequence in $[0,1]^K$ and fixed source delays on a self-looped graph. Write $\rho = \mathrm{mas}(G)$, which equals $\alpha(G)$ for an undirected graph; this parameter is used only in the analysis. The run starts on the next round after a

Algorithm 2 Delay-adaptive exponential weights with skipping after a test failure

Input: Self-looped graph G with $K \geq 2$ vertices, independent randomness; runs until stopped at the epoch boundary

Initialize: $L^{\mathrm{obs}} \leftarrow 0, \Sigma_0, V_0, B_0 \leftarrow 0, \mathcal{P} \leftarrow \varnothing$
For each trial B, set $p_i(B) \propto \exp(-L_{t,i}^{\mathrm{obs}}/B)$.
Set $\theta_t(B) = \sum_i p_i(B)/Q_i(p(B))$.

for local rounds $t = 1, 2, \ldots$ until the enclosing epoch ends **do**
  $\mathcal{P}_t^- \leftarrow \mathcal{P}$; $\Sigma_t \leftarrow \Sigma_{t-1} + |\mathcal{P}_t^-|$ ▷ *count pending sources*
  $B \leftarrow \max\{B_{t-1}, \sqrt{(V_{t-1} + \Sigma_t + 1)/\log K}\}$
  while $B^2 \log K < V_{t-1} + \Sigma_t + \theta_t(B)$ **do**
    $B \leftarrow 2B$ ▷ *recompute the trial distribution and ratio*
  end while
  $B_t \leftarrow B$; $p_t^{\mathrm{DF}} \leftarrow p(B_t)$; $\theta_t \leftarrow \theta_t(B_t)$; $V_t \leftarrow V_{t-1} + \theta_t$
  $\mathcal{P}_t \leftarrow \{s \in \mathcal{P}_t^- : t - s \leq B_t\}$; $\mathcal{P} \leftarrow \mathcal{P}_t$ ▷ *skip before inspecting records*
  Draw $I_t \sim p_t^{\mathrm{DF}}$
  Save I_t and $Q_i(p_t^{\mathrm{DF}})$ for $i \in N^{\mathrm{out}}(I_t)$ with source t
  Register source t in $\mathcal{P}$
  for sources $s \in \mathcal{P}$ whose feedback arrives now **do**
    Form $\widehat{\ell}_s$ using saved source probabilities in (50)
    $L^{\mathrm{obs}} \leftarrow L^{\mathrm{obs}} + \widehat{\ell}_s$; remove s from $\mathcal{P}$
  end for
end for

test failure with an empty score, independent randomization, and local time starting at one. It uses no previous sources and ends at the enclosing epoch boundary.

Let L_t^{obs} be the sum of estimates processed before decision t, and let $\mathcal{P}_t^-$ contain the earlier sources that are pending (their feedback has not yet arrived) and not yet skipped. Initialize $\Sigma_0 = V_0 = B_0 = 0$; the algorithm counts $\Sigma_t = \Sigma_{t-1} + |\mathcal{P}_t^-|$ before choosing a scale. For each trial $B > 0$, write $p_i(B) \propto \exp(-L_{t,i}^{\mathrm{obs}}/B)$ and $\theta_t(B) = \sum_i p_i(B)/Q_i(p(B))$. The scale test starts at $B = \max\{B_{t-1}, \sqrt{(V_{t-1} + \Sigma_t + 1)/\log K}\}$ and doubles B while $B^2 \log K < V_{t-1} + \Sigma_t + \theta_t(B)$. The accepted values are

$$B_t = B, \qquad p_t^{\mathrm{DF}} = p(B_t), \qquad \theta_t = \theta_t(B_t), \qquad V_t = V_{t-1} + \theta_t. \tag{48}$$

After the scale test, the age test retains the sources $s \in \mathcal{P}_t^-$ with $t - s \leq B_t$, which form $\mathcal{P}_t$; the others are skipped. The action distribution is

$$p_{t,i}^{\mathrm{DF}} = \frac{\exp(-L_{t,i}^{\mathrm{obs}}/B_t)}{\sum_{j=1}^{K} \exp(-L_{t,j}^{\mathrm{obs}}/B_t)}. \tag{49}$$

An arriving source s is estimated using its source probabilities:

$$\widehat{\ell}_{s,i} = \frac{\ell_{s,i} \mathbf{1}\{i \in N^{\mathrm{out}}(I_s)\}}{Q_i(p_s^{\mathrm{DF}})}. \tag{50}$$

Self-loops and the positive exponential-weights distribution give $Q_i(p_s^{\mathrm{DF}}) \geq p_{s,i}^{\mathrm{DF}} > 0$. The source is registered before end-of-round arrivals, including when its delay is zero; the feedback of skipped sources is never processed.

The operative order is count, scale, skip, action, then arrivals. Trial distributions involve no draws, and both the scale test and skip decisions use the processed history without inspecting the stored action records of pending sources; all skips in an arriving batch therefore precede processing that batch. The random scale is nondecreasing and needs neither a graph parameter nor a horizon input. By Lemma 23, $1 \leq \theta_t(B) \leq \rho \leq K$, so the loop terminates. Its starting squared scale is at least $(V_{t-1} + \Sigma_t + 1)/\log K$, while each failed trial has $B^2 \log K < V_{t-1} + \Sigma_t + K$; thus each round has at most $O(\log K)$ doublings. Lemma 26 bounds the skipped sources.

For a fixed arm i, let $\mathrm{Reg}_n^{\mathrm{DF}}(i) = \sum_{t \leq n} (\ell_{t,I_t} - \ell_{t,i})$ be the regret of Algorithm 2 against i over its first n rounds.

Proposition 22 (Regret of the skipping subroutine). *Let Algorithm 2 start from an empty state with independent randomization, and let $\rho = \alpha$ for undirected graphs and $\rho = \mathrm{mas}(G)$, the size of a maximum acyclic induced subgraph, for directed graphs. For every deterministic n, fixed arm i, and fixed $U \subseteq [n]$, under oblivious losses and fixed delays,*

$$\mathbb{E}\,\mathrm{Reg}_n^{\mathrm{DF}}(i) \le 30\left[\sqrt{\rho n \log K} + |U| + 1 + \sqrt{\log K \sum_{s\in[n]\setminus U} \min\{d_s, n-s\}}\right]. \tag{51}$$

The proof has four steps. Be-the-leader on the estimates processed by round n, with a distribution p_t^+ that already includes these processed estimates through round t, costs $B_n \log K$. The played distribution lags p_t^+ by the pending estimates; since the estimates are nonnegative but can be large, we use only the one-sided comparison of Gyorgy & Joulani (2021), which costs a ratio term θ_t/B_t and a cross term with the pending sources, and Lemma 24 bounds the latter by $|\mathcal{P}_t|/B_t$ in expectation because the stored action records of pending sources are never read. The adaptive scale balances $B_t^2 \log K$ against the observed ratios and pending counts $V_t + \Sigma_t$, so these terms sum to at most $2B_n \log K$ without knowledge of ρ; it is chosen by doubling because $\theta_t(B)$ depends on B and B_t must be nondecreasing for the be-the-leader step. Finally, a source is skipped only after waiting more than B_t, so the total waiting grows quadratically in the number of skips (Lemma 26). Delay-aware skipping also appears in Zimmert & Seldin (2020); the moment and counting arguments for this rule are proved below.

G.1 The graph estimate and a one-sided comparison

We give the full expected-regret proof for (48)–(49).

Lemma 23 (Graph ratios and scale bounds). *For a self-looped directed graph and every positive distribution p, $1 \le \sum_{i=1}^K p_i/Q_i(p) \le \rho = \mathrm{mas}(G) \le K$. The accepted scale is nondecreasing and satisfies $B_t^2 \log K \ge V_t + \Sigma_t$ and*

$$B_t \le 2\sqrt{(\rho t + \Sigma_t)/\log K}. \tag{52}$$

Proof. Draw independent exponential clocks with rates p_i and select a vertex whose clock precedes every distinct incoming neighbor. Its selection probability is $p_i/Q_i(p)$. Selected vertices cannot form a directed cycle, since the clocks would then be strictly ordered around that cycle. Their number is at most ρ. Taking expectation proves the upper bound; the lower bound follows from $Q_i(p) \le 1$. For symmetric graphs the selected set is independent. This is the standard graph-ratio bound (Alon et al., 2017).

For every positive trial distribution, the graph-ratio bound gives $1 \le \theta_t(B) \le \rho$. The accepted scale is nondecreasing and satisfies the stated lower bound by the acceptance test. For the upper bound, if no doubling occurs, the inherited scale obeys this bound by induction, and the proposed square-root scale does too since $V_{t-1} \le \rho(t-1)$ and $\rho \ge 1$. If doubling occurs, the preceding trial $B_t/2$ failed, whence $(B_t/2)^2 \log K < V_{t-1} + \Sigma_t + \theta_t(B_t/2) \le \rho t + \Sigma_t$. No monotonicity of $\theta_t(B)$ is needed.

□

In this proof write $p_t = p_t^{\mathrm{DF}}$. Fix a deterministic evaluation time n. Let $\mathcal{S}_{\mathrm{drop}}$ consist of sources skipped by round n, together with those still pending after its end-of-round arrivals, and set $k_s = \mathbf{1}\{s \notin \mathcal{S}_{\mathrm{drop}}\}$. Both $\mathcal{S}_{\mathrm{drop}}$ and k_s may depend on the observed actions. Define the processed estimates $\widetilde{\ell}_s = k_s \widehat{\ell}_s$ and, for analysis only,

$$L_t^{\mathrm{pend}} = \sum_{s=1}^{t-1} \widetilde{\ell}_s - L_t^{\mathrm{obs}} \ge 0, \quad \text{and} p_{t,i}^+ = \frac{\exp(-(L_{t,i}^{\mathrm{obs}} + L_{t,i}^{\mathrm{pend}} + \widetilde{\ell}_{t,i})/B_t)}{\sum_{j=1}^K \exp(-(L_{t,j}^{\mathrm{obs}} + L_{t,j}^{\mathrm{pend}} + \widetilde{\ell}_{t,j})/B_t)}.$$

The vector L_t^{pend} sums the estimates of sources processed by round n (that is, with $k_s = 1$) that are still pending before decision t. Every such source belongs to $\mathcal{P}_t$. Neither L_t^{pend} nor p_t^+ is computed by the algorithm. In particular, p_t^+ includes the current estimate.

We realize the independent randomization as $I_s = \iota(p_s, \vartheta_s)$, where $\vartheta_1, \vartheta_2, \ldots$ are independent and uniform on $[0,1]$ and $\iota(p,x) = \min\{a \in [K] : \sum_{b\le a} p_b \ge x\}$, so that $\mathbb{P}(\iota(p,\vartheta_s) = a) = p_a$. A

source is processed when its estimate is added to L^{obs} in lines 38–40. For a decision t, let $\mathcal{R}_t$ be the set of sources processed in the arrival steps of rounds $1, \dots, t-1$, and let $\mathrm{Unproc}_t = [t-1] \setminus \mathcal{R}_t$ be the earlier sources not processed before decision t, whether pending or skipped. The decision history is

$$\mathcal{G}_t = \sigma(\Upsilon_t), \qquad \Upsilon_t = \big(\mathcal{R}_t, (I_s)_{s\in\mathcal{R}_t}\big).$$

Lemma 24 (Unprocessed actions are unrevealed)**.** *Under oblivious losses and fixed delays, the following hold for every decision t.*

(i) Every quantity that Algorithm 2 computes before drawing I_t is $\mathcal{G}_t$-measurable: L_u^{obs}, Σ_u, the trial ratios, B_u, θ_u, V_u, p_u, $\mathcal{P}_u^-$ and $\mathcal{P}_u$ for $u \le t$, and the round at which each source $s < t$ is processed or skipped, if this happens before decision t.

(ii) For every value z of Υ_t with $\mathbb{P}(\Upsilon_t = z) > 0$, every $a \in [K]$ and every $(a_s)_{s\in\mathrm{Unproc}_t}$ in $[K]^{\mathrm{Unproc}_t}$,

$$\mathbb{P}\big(I_t = a,\ I_s = a_s \text{ for all } s \in \mathrm{Unproc}_t \,\big|\, \Upsilon_t = z\big) = p_{t,a} \prod_{s\in\mathrm{Unproc}_t} p_{s,a_s},$$

where p_t and p_s are the values determined by z through (i).

(iii) Fix n and a source $s \le n$, and let $k_s = 1$ if s is processed in the arrival step of some round $t \le n$ and $k_s = 0$ otherwise. If $s + d_s > n$ then $k_s = 0$. Otherwise let $t_s = s + d_s$. Then k_s is $\mathcal{G}_{t_s}$-measurable, and for every $\mathcal{G}_{t_s}$-measurable random function $f : [K] \to \mathbb{R}$,

$$\mathbb{E}\big[k_s f(I_s) \,\big|\, \mathcal{G}_{t_s}\big] = k_s \sum_{a=1}^{K} p_{s,a} f(a).$$

Proof. For a deterministic sequence $x \in [K]^{t-1}$, run Algorithm 2 with the draw in line 35 replaced by $I_u = x_u$ for $u < t$, and stop after line 34 of round t. Since the losses and delays are fixed and trial distributions involve no draws, this defines deterministic functions $\mathcal{R}(x)$ and $\Phi(x)$, the latter collecting all quantities listed in (i). We claim that $\Phi(y) = \Phi(x)$ and $\mathcal{R}(y) = \mathcal{R}(x)$ whenever $y_s = x_s$ for all $s \in \mathcal{R}(x)$. We show by induction over $u \le t$ that the runs driven by x and y enter round u with the same L^{obs}, Σ_{u-1}, V_{u-1}, B_{u-1}, set $\mathcal{P}$ and list of processing and skip rounds. This holds at $u = 1$ since both start from the empty state. Lines 28–34 read only these quantities and the source indices in $\mathcal{P}$, never the stored action records, so both runs compute the same values in round u up to line 34; for $u = t$ this is the claim. For $u < t$, lines 35–37 store the record of source u and register u, so $\mathcal{P}$ remains equal. The sources processed in lines 38–40 are the members s of $\mathcal{P}$ with $s + d_s = u$, the same set in both runs because delays are fixed. Each of them lies in $\mathcal{R}(x)$, so $y_s = x_s$, and its estimate (50) depends only on I_s, the fixed ℓ_s, and p_s, which agree by the induction hypothesis at $u = s$. Hence both runs enter round $u + 1$ with the same state, which proves the claim.

Let $\zeta(x) = (\mathcal{R}(x), (x_s)_{s\in\mathcal{R}(x)})$, so that $\Upsilon_t = \zeta(I_1, \dots, I_{t-1})$. If $\zeta(x) = \zeta(y)$, then y agrees with x on $\mathcal{R}(x)$ and the claim gives $\Phi(y) = \Phi(x)$. Thus Φ factors through ζ, which proves (i). The claim also shows that for every attained value $z = (R, r)$ of Υ_t,

$$\{\Upsilon_t = z\} = \{I_s = r_s \text{ for all } s \in R\}. \tag{53}$$

For (ii), each p_u is a function of $I_1, \dots, I_{u-1}$, hence of $\vartheta_1, \dots, \vartheta_{u-1}$, and ϑ_u is independent of these. The chain rule gives $\mathbb{P}(I_u = x_u \text{ for } u \le t) = \prod_{u\le t} p_{u,x_u}(x)$, where $p_u(x)$ is the distribution computed in the run driven by x. Fix $z = (R, r)$ with $\mathbb{P}(\Upsilon_t = z) > 0$, a and $(a_s)_{s\in\mathrm{Unproc}_t}$, and let $x \in [K]^t$ be the unique sequence with $x_s = r_s$ for $s \in R$, $x_s = a_s$ for $s \in \mathrm{Unproc}_t = [t-1] \setminus R$, and $x_t = a$. By (53), the event in (ii) intersected with $\{\Upsilon_t = z\}$ is $\{I_u = x_u \text{ for } u \le t\}$, and by (i), $p_u(x)$ is the value p_u determined by z for every $u \le t$. Therefore

$$\mathbb{P}\big(\Upsilon_t = z,\ I_t = a,\ I_s = a_s \text{ for all } s \in \mathrm{Unproc}_t\big) = \prod_{s\in R} p_{s,r_s} \prod_{s\in\mathrm{Unproc}_t} p_{s,a_s}\, p_{t,a}.$$

Summing over a and $(a_s)_{s\in\mathrm{Unproc}_t}$ gives $\mathbb{P}(\Upsilon_t = z) = \prod_{s\in R} p_{s,r_s}$, and dividing proves (ii).

For (iii), source s can be processed only in the arrival step of round $s + d_s$, so $k_s = 0$ if $s + d_s > n$. If $d_s = 0$, source s is registered and processed in round s, so $k_s = 1$ and $t_s = s$. If $d_s \ge 1$, source s

is processed in round t_s exactly when it is retained by every age test up to decision t_s, that is, when $s \in \mathcal{P}_{t_s}$; this is $\mathcal{G}_{t_s}$-measurable by (i), and $s \in \mathrm{Unproc}_{t_s}$. In both cases, (ii) at $t = t_s$ shows that conditional on $\Upsilon_{t_s} = z$ the action I_s has law p_s. Since k_s and f are fixed on $\{\Upsilon_{t_s} = z\}$, the identity follows. □

Lemma 25 (Expected regret decomposition)**.** *For every deterministic n and fixed comparator i,*

$$\mathbb{E}\,\mathrm{Reg}_n^{\mathrm{DF}}(i) \le \mathbb{E}\big[|\mathcal{S}_{\mathrm{drop}}| + 3B_n \log K\big]. \tag{54}$$

Proof. Pathwise comparison. The exponential-weights comparison has two parts. For every realized trajectory and fixed comparator i,

$$\sum_{t=1}^{n} \langle p_t^+ - e_i, \widetilde{\ell}_t \rangle \le B_n \log K, \tag{55}$$

$$\langle p_t - p_t^+, \widetilde{\ell}_t \rangle \le \frac{1}{B_t} \sum_{j=1}^{K} p_{t,j} \widetilde{\ell}_{t,j}^2 + \frac{1}{B_t} \sum_{j=1}^{K} p_{t,j} \widetilde{\ell}_{t,j} L_{t,j}^{\mathrm{pend}}. \tag{56}$$

For the first comparison inequality let $R(p) = \log K + \sum_{j=1}^{K} p_j \log p_j \ge 0$, $B_0 = 0$, and $F_t(p) = \langle \sum_{s=1}^{t} \widetilde{\ell}_s, p \rangle + B_t R(p)$. For $t \ge 1$, p_t^+ minimizes F_t. At $t = 0$ every distribution minimizes F_0, so choose p_0^+ uniform. Evaluating F_{t-1} at p_t^+ gives

$$F_t(p_t^+) - F_{t-1}(p_{t-1}^+) \ge \langle p_t^+, \widetilde{\ell}_t \rangle + (B_t - B_{t-1}) R(p_t^+) \ge \langle p_t^+, \widetilde{\ell}_t \rangle.$$

Sum this inequality, use $F_0(p_0^+) = 0$, and compare $F_n(p_n^+) \le F_n(e_i) = \sum_{t=1}^{n} \widetilde{\ell}_{t,i} + B_n \log K$. This is the standard be-the-leader argument with nondecreasing coefficients.

For the one-sided inequality use nonnegativity:

$$\frac{p_{t,j}^+}{p_{t,j}} = \frac{\exp(-(\widetilde{\ell}_{t,j} + L_{t,j}^{\mathrm{pend}})/B_t)}{\sum_{k=1}^{K} p_{t,k} \exp(-(\widetilde{\ell}_{t,k} + L_{t,k}^{\mathrm{pend}})/B_t)} \ge \exp(-(\widetilde{\ell}_{t,j} + L_{t,j}^{\mathrm{pend}})/B_t).$$

Consequently $1 - p_{t,j}^+/p_{t,j} \le (\widetilde{\ell}_{t,j} + L_{t,j}^{\mathrm{pend}})/B_t$, by $1 - e^{-x} \le x$. Multiply by the nonnegative $p_{t,j}\widetilde{\ell}_{t,j}$ and sum. This is the comparison from Gyorgy & Joulani (2021), applied to the processed graph estimates.

Conditional moments and processing. The graph estimates satisfy the following expected moment bounds for every $t \le n$:

$$\mathbb{E}\frac{1}{B_t} \sum_{j=1}^{K} p_{t,j} \widetilde{\ell}_{t,j}^2 \le \mathbb{E}\frac{\theta_t}{B_t}, \quad \text{and} \quad \mathbb{E}\frac{1}{B_t} \sum_{j=1}^{K} p_{t,j} \widetilde{\ell}_{t,j} L_{t,j}^{\mathrm{pend}} \le \mathbb{E}\frac{|\mathcal{P}_t|}{B_t}.$$

Let $\mathcal{G}_t$ be the decision history of Lemma 24. By part (i) of that lemma, B_t, p_t, $\mathcal{P}_t$, and the source distributions p_s for $s < t$ are $\mathcal{G}_t$-measurable, and $\mathcal{P}_t \subseteq \mathrm{Unproc}_t$. By part (ii), conditional on $\mathcal{G}_t$, the current action has law p_t, and for each $s \in \mathcal{P}_t$ the pair (I_t, I_s) has law $p_t \otimes p_s$.

By nonnegativity, $\widetilde{\ell}_{t,j}^2 \le \widehat{\ell}_{t,j}^2$. Conditional on $\mathcal{G}_t$, the current action has law p_t, so $\mathbb{E}[\widehat{\ell}_{t,j}^2 \mid \mathcal{G}_t] = \ell_{t,j}^2/Q_j(p_t)$. Weighting by $p_{t,j}/B_t$ and summing gives the first bound.

For the cross term, nonnegativity allows us to drop the future processing indicators k_s before conditioning:

$$\widetilde{\ell}_{t,j} L_{t,j}^{\mathrm{pend}} \le \widehat{\ell}_{t,j} \sum_{s \in \mathcal{P}_t} \widehat{\ell}_{s,j}.$$

For each $s \in \mathcal{P}_t$, the product law of (I_t, I_s) gives $\mathbb{E}[\widehat{\ell}_{t,j}\widehat{\ell}_{s,j} \mid \mathcal{G}_t] = \ell_{t,j}\ell_{s,j} \le 1$. Weight by the $\mathcal{G}_t$-measurable $p_{t,j}/B_t$, sum over j and pending sources, and take expectations. The two source observation denominators cancel separately.

The indicator $k_s = \mathbf{1}\{s \notin \mathcal{S}_{\mathrm{drop}}\}$ is the one in Lemma 24(iii), since $k_s = 1$ exactly when source s is processed by round n. If $s + d_s > n$, then $k_s = 0$ and both sides of the next display vanish.

Otherwise apply that part with $f(a) = \ell_{s,a} - \ell_{s,i}$ and with $f(a) = \sum_{j=1}^{K}(p_{s,j} - \mathbf{1}\{j = i\})\ell_{s,j}\mathbf{1}\{j \in N^{\text{out}}(a)\}/Q_j(p_s)$, both $\mathcal{G}_{t_s}$-measurable; since $\sum_a p_{s,a}\mathbf{1}\{j \in N^{\text{out}}(a)\} = Q_j(p_s)$, both conditional means equal $k_s\langle p_s - e_i, \ell_s\rangle$. Hence, for every source $s \le n$,

$$\mathbb{E}\big[k_s(\ell_{s,I_s} - \ell_{s,i})\big] = \mathbb{E}\langle p_s - e_i, \widetilde{\ell}_s\rangle.$$

Combination. Each source in $\mathcal{S}_{\text{drop}}$ contributes at most one actual regret pathwise. The comparison and expected moment bounds yield

$$\mathbb{E}\,\text{Reg}_n^{\text{DF}}(i) \le \mathbb{E}\left[|\mathcal{S}_{\text{drop}}| + B_n \log K + \sum_{t=1}^{n} \frac{\theta_t + |\mathcal{P}_t|}{B_t}\right].$$

The increment of $V_t + \Sigma_t$ is $\theta_t + |\mathcal{P}_t^-| \ge \theta_t + |\mathcal{P}_t|$, and $B_t \ge \sqrt{(V_t + \Sigma_t)/\log K}$. Pathwise,

$$\sum_{t=1}^{n} \frac{\theta_t + |\mathcal{P}_t|}{B_t} \le \sqrt{\log K} \sum_{t=1}^{n} \frac{\theta_t + |\mathcal{P}_t^-|}{\sqrt{V_t + \Sigma_t}} \le 2\sqrt{(V_n + \Sigma_n)\log K} \le 2B_n \log K.$$

Pure exponential weights are positive at every finite round in real arithmetic, so every estimate is finite. The moment calculations above establish integrability without a uniform positive probability floor. □

G.2 A SINGLE AGE CONDITION CONTROLS SKIPPED OBSERVATIONS

Lemma 26 (Waiting times and skipped sources)**.** *Let w_s be the number of rounds $t \le n$ in which source s belongs to $\mathcal{P}_t^-$. Let $\mathcal{S}_{\text{skip}} \subseteq \mathcal{S}_{\text{drop}}$ be the sources skipped by the age test by round n, excluding those that are merely still pending at the evaluation time. Then*

$$\Sigma_n = \sum_{s=1}^{n} w_s, \qquad w_s \le \min\{d_s, n - s, B_n + 1\}, \tag{57}$$

$$|\mathcal{S}_{\text{skip}}| \le 2\sqrt{\Sigma_n \log K}, \qquad |\mathcal{S}_{\text{drop}}| \le 2\sqrt{\Sigma_n \log K} + B_n + 1. \tag{58}$$

For every fixed $U \subseteq [n]$,

$$\Sigma_n \le D_n(U) + |U|(B_n + 1), \quad \textit{where} \quad D_n(U) = \sum_{s \in [n]\setminus U} \min\{d_s, n - s\}. \tag{59}$$

Proof. Membership of a source in the pre-skip pending set $\mathcal{P}_t^-$ is consecutive, starting at decision $s + 1$. It ends on its arrival or skip, or at evaluation. Hence its count does not exceed $\min\{d_s, n - s\}$. If t is its last decision with $s \in \mathcal{P}_t^-$ and it has waited at least two decisions, it was retained by the previous age test, so $t - 1 - s \le B_{t-1} \le B_n$. If it waited just one decision the bound by $B_n + 1$ is immediate. Double counting proves the first equality.

Order the actually skipped sources by skip time, resolving simultaneous skips arbitrarily; denote their waiting counts by $w'_1, \dots, w'_m$ and set $W_j = \sum_{k=1}^{j} w'_k$. A source skipped at t_j has $w'_j = t_j - s_j > B_{t_j}$. All these W_j units of waiting have already been counted in Σ_{t_j}, so

$$w'_j > \sqrt{W_j/\log K}, \quad \text{and} \sqrt{W_j} - \sqrt{W_{j-1}} = \frac{w'_j}{\sqrt{W_j} + \sqrt{W_{j-1}}} > \frac{1}{2\sqrt{\log K}}.$$

Summing and using $W_m \le \Sigma_n$ proves the first count bound. Before the action at n, the retained sources have distinct positive integer ages at most B_n, so there are at most B_n of them. Adding the new source shows that at most $B_n + 1$ sources are still pending after that round's arrivals. This proves the bound on $\mathcal{S}_{\text{drop}}$. Finally split the sum of w_s over U and its complement, using the respective two bounds in (57). □

Proof of Proposition 22. Write $x = B_n \log K$ and $m = |U|$. Lemma 26 is pathwise for the realized nondecreasing scale. Since $\log K \ge \log 2$, it and Lemma 25 give $\mathbb{E}\,\text{Reg}_n^{\text{DF}}(i) \le \mathbb{E}[7x + 1]$. The waiting-time bound and (52) imply

$$x^2 \le 4 \log K\{\rho n + D_n(U) + m\} + 4mx.$$

Solving this quadratic and using $m \le n \le \rho n$ gives

$$x \le 4m + 2\sqrt{\log K\{\rho n + D_n(U) + m\}} \le 4m + 2\sqrt{2\rho n \log K} + 2\sqrt{\log K\, D_n(U)}.$$

The deterministic right side bounds x pathwise; substituting into $\mathbb{E}[7x+1]$ proves (51). The statement is for each deterministic evaluation time; Lemma 20 applies it after a random test failure. □

H A CONNECTED-GRAPH LIMIT ON THE ARITHMETIC-AVERAGE BENCHMARK

This lower bound concerns every policy, including policies given the horizon and delay. For a stochastic instance ν, define the proposed uniform benchmark

$$\mathrm{Bench}_T(\nu) = H_G(\Delta)\log(eKT)\log(eT) + \Delta_{\mathrm{avg}}\sigma_{\max} + 1. \tag{60}$$

A uniform bound would use one constant for all graphs, gaps, horizons, and delays. The construction rests on an information barrier. The loss of a leaf v is revealed only by playing v or a single neighbor b, both of which are costly on the instances where v is suboptimal, so small regret on these instances allows few leaf observations. Revealing each block's feedback at its end, which only helps the learner, turns the problem into M batches in which each block can use only the leaf observations of earlier blocks, as in batched bandits. Even after all M blocks the learner cannot tell whether v is optimal, so on the instance where v is optimal it misses v in most rounds, and the ratio to the benchmark is of order $M = \Theta(\log\log T)$. The blind-prefix comparison is given at the end of this appendix.

H.1 A FINITE FAMILY ON A CONNECTED GRAPH

Fix integers $d = K \ge 4$ and $M \ge 2$, and put

$$T = Md, \quad L^2 = \log(eKT)\log(eT), \quad \varepsilon = \frac{L}{\sqrt{d}}, \quad \text{and} \quad \Lambda_0 = L\sqrt{d} = \varepsilon d. \tag{61}$$

We assume $\varepsilon \le 1/4$. Vertices $1, \dots, K-1$ form a clique, $b = K-1$ is joined to the leaf $v = K$, and every vertex has a self-loop. The graph is connected, has diameter two, b is universal, and $\alpha(G) = 2$. Playing a clique arm reveals the clique losses; playing b also reveals the leaf loss, while playing v reveals the losses of v and b.

All clique losses are deterministic and common to the family:

$$\ell_{t,1} = \frac{1}{2}, \quad \ell_{t,i} = \frac{1}{2} + \varepsilon \quad (2 \le i \le K-2), \quad \text{and} \quad \ell_{t,b} = 1.$$

The leaf has independent Bernoulli losses over time. Define

$$\begin{aligned} \nu_- &: \quad \mu_v = \frac{1}{2} - \varepsilon, \\ \nu_\varphi &: \quad \mu_v = \frac{1}{2} + \varphi, \quad \text{where} \quad \varepsilon \le \varphi \le \frac{1}{4}. \end{aligned}$$

These are bounded i.i.d. loss-vector instances with unique optimal arms: v is optimal under ν_- and arm 1 under ν_φ. Every source has the constant delay d, with feedback arriving at the end of round $s + d$, so $\sigma_{\max} = d$. The learner may know G, T, d, and all clique means; only the leaf mean is unknown.

Direct calculation of this family gives

$$\begin{aligned} H_G(\nu_\varphi) &= \frac{1}{\varepsilon} + \frac{1}{\varphi}, & \Delta_{\mathrm{avg}}(\nu_\varphi) &= \frac{(K-3)\varepsilon + 1/2 + \varphi}{K}, \\ H_G(\nu_-) &= \frac{1}{\varepsilon}, & \Delta_{\mathrm{avg}}(\nu_-) &= \frac{(2K-4)\varepsilon + 1/2}{K}. \end{aligned}$$

Whenever $\Lambda_0 \ge 2$ and $\varepsilon \le \varphi \le 1/4$,

$$\mathrm{Bench}_T(\nu_\varphi) \le 4\Lambda_0 \quad \text{and} \quad \mathrm{Bench}_T(\nu_-) \le 4\Lambda_0. \tag{62}$$

An independent set contains at most one clique vertex and may contain v with any clique vertex other than b. Under ν_φ, excluding optimal arm 1, the largest inverse-gap sum uses a middle clique arm and

v, giving $1/\varepsilon + 1/\varphi$. Under ν_-, excluding optimal v, the largest inverse-gap weight is $1/\varepsilon$ from arm 1. Directly summing the gaps gives the stated arithmetic averages. As $K = d$, the positive-instance learning contribution is at most $2L^2/\varepsilon = 2\Lambda_0$ and its delay contribution is at most $\Lambda_0 + 3/4$. The negative-instance contributions are at most Λ_0 and $2\Lambda_0 + 1/2$. Adding 1 and using $\Lambda_0 \geq 2$ proves (62). Thus the choice of ε balances the learning and delay terms of the benchmark at order Λ_0, and $K = d$ keeps the contribution of the large gaps of b and v to $\Delta_{\text{avg}}\sigma_{\max}$ constant.

H.2 Information available before each block ends

The next lemma relates the number of informative actions in one instance to the regret in a comparison instance. Partition time into M blocks of d rounds. For the lower bound, enlarge the learner's information: after the last decision in each block, reveal every graph-feedback batch generated in that block. For a source s in block j, $s \leq jd < s + d$, so this extra feedback arrives earlier than its constant-delay feedback. Every original policy can be run in the enlarged experiment by storing early observations until their original arrival times; its action and regret laws are unchanged. A lower bound for all policies in the enlarged experiment therefore applies to the original model.

No outcome generated in block j is visible before the last decision of that block. Let N_j count selections of either v or b in the first j blocks, and put $x_j = \mathbb{E}_- N_j$, $x_0 = 0$. Let $\mathcal{T}_j$ contain all actions through block j and all feedback from the first $j-1$ blocks, before outcomes from block j are revealed. This history determines N_j.

Lemma 27 (Information comparison)**.** *For any policy* alg *in the enlarged experiment and* $\varepsilon \leq \varphi \leq 1/4$*, if* $R_T(\mathsf{alg}, \nu_\varphi) \leq \bar{R}$*, then*

$$x_j \leq \frac{\bar{R}}{\varphi} + 4jd\,\varphi\sqrt{x_{j-1}}. \tag{63}$$

Proof. The histories satisfy

$$\mathrm{KL}(\mathbb{P}_-^{\mathcal{T}_j}, \mathbb{P}_\varphi^{\mathcal{T}_j}) = x_{j-1}\,\mathrm{kl}\big(\frac{1}{2} - \varepsilon, \frac{1}{2} + \varphi\big) \leq 32\varphi^2 x_{j-1}. \tag{64}$$

At a fixed observed history, the action kernel is the same in both instances. A clique observation is deterministic and common to the family. An observed leaf loss contributes the displayed Bernoulli relative entropy, whether it was obtained by playing v or by playing b. Before the outcomes of block j are revealed, exactly the selections of v or b in blocks $1, \ldots, j-1$ have supplied leaf observations. The relative-entropy chain rule gives the equality in (64), including for randomized policies.

For Bernoulli parameters p, q, $\mathrm{kl}(p, q) \leq (p-q)^2/[q(1-q)]$. Here $q(1-q) \geq 3/16$ and $|p - q| = \varepsilon + \varphi \leq 2\varphi$, so the relative entropy per leaf observation is below $32\varphi^2$. Pinsker's inequality gives $\mathrm{TV}(\mathbb{P}_-^{\mathcal{T}_j}, \mathbb{P}_\varphi^{\mathcal{T}_j}) \leq 4\varphi\sqrt{x_{j-1}}$. Since $0 \leq N_j \leq jd$, its expectations differ by at most jd times this total variation distance, proving

$$x_j \leq \mathbb{E}_\varphi N_j + 4jd\,\varphi\sqrt{x_{j-1}}. \tag{65}$$

Under ν_φ, every choice counted by N_j incurs gap at least φ: the gap of v is φ and that of b is $1/2$. Hence $R_T(\mathsf{alg}, \nu_\varphi) \leq \bar{R}$ gives $\mathbb{E}_\varphi N_j \leq \mathbb{E}_\varphi N_M \leq \bar{R}/\varphi$. Combine this with (65) to obtain (63). □

Under ν_-, every action outside $\{v, b\}$ incurs gap at least ε. Ignoring the positive regret of b gives the conservative inequality $R_T(\mathsf{alg}, \nu_-) \geq \varepsilon(T - x_M)$.

H.3 A finite lower bound and its consequence

For $j = 0, \ldots, M$ put $\eta_j = 8^{-j}$, and for $j \geq 1$ choose

$$\varphi_j = d^{-1/2+2\eta_j}. \tag{66}$$

The comparison instances share the graph, clique losses, horizon, and delay. The recursion $\eta_{j-1} = 8\eta_j$ limits the expected plays of v or b under ν_- through block j to $d^{1-\eta_j}$ when the positive instances have regret at most $M\Lambda_0/4$, and the choice $M = \Theta(\log\log d)$ in Corollary 29 ensures the condition $d^{\eta_M} \geq M(L/4 + 4)$ of Proposition 28.

Proposition 28 (Connected-graph lower bound)**.** *Suppose* (61) *gives* $\Lambda_0 \geq 2$ *and*

$$\varepsilon \leq \varphi_M \leq \varphi_1 \leq \frac{1}{4}, \quad \textit{and} \quad d^{\eta_M} \geq M\big(\frac{L}{4}+4\big). \tag{67}$$

For every policy alg*, even one given the horizon and delay, some* $\nu \in \{\nu_-, \nu_{\varphi_1}, \dots, \nu_{\varphi_M}\}$ *satisfies*

$$R_T(\mathsf{alg}, \nu) \geq \frac{M}{16}\mathrm{Bench}_T(\nu). \tag{68}$$

Proof. Work in the enlarged experiment; every original policy is included by simulation. If any positive instance has regret greater than $M\Lambda_0/4$, (62) gives (68). Otherwise all positive instances have regret at most $\bar{R} = M\Lambda_0/4 = ML\sqrt{d}/4$. We prove by induction that $x_j \leq d^{1-\eta_j}$ for $0 \leq j \leq M$. The base case is $x_0 = 0$. Assume the bound at $j-1$ and apply (63) with φ_j. Since $\eta_{j-1} = 8\eta_j$,

$$\frac{\bar{R}}{\varphi_j} = \frac{ML}{4}\, d^{1-2\eta_j},$$

$$4jd\,\varphi_j\sqrt{x_{j-1}} \leq 4j\, d^{1+(-1/2+2\eta_j)+(1-8\eta_j)/2} = 4j\, d^{1-2\eta_j}.$$

The condition in (67) then yields $x_j \leq (ML/4+4j)d^{1-2\eta_j} \leq d^{\eta_M} d^{1-2\eta_j} \leq d^{1-\eta_j}$. On ν_-, the conservative regret bound above gives

$$R_T(\mathsf{alg}, \nu_-) \geq \varepsilon(T - x_M) \geq \Lambda_0(M - d^{-\eta_M}) \geq M\Lambda_0/2.$$

By (62), its ratio to the benchmark is at least $M/8$, stronger than (68). □

Corollary 29 (No uniform arithmetic-average benchmark on connected graphs; Proposition 4 restated)**.** *No policy and universal constant* C *satisfy* $R_T \leq C\,\mathrm{Bench}_T(\nu)$ *for every known connected undirected self-looped graph, bounded i.i.d. loss instance with unique optimum, horizon, and constant delay, even if the horizon and delay are supplied to the policy. Along a sequence with* $\alpha(G) = 2$*, graph diameter two, and* $T \geq K$*, every policy has an instance on which the ratio is* $\Omega(\log\log T)$*.*

Proof. For sufficiently large integer d, choose $M = \lfloor \frac{1}{2}\log_8(\log d)\rfloor \geq 2$. Then $L = \Theta(\log d)$, $M = \Theta(\log\log d)$, and $d^{\eta_M} \geq \exp(\sqrt{\log d})$, which eventually exceeds $M(L/4+4)$. Also $\varphi_1 = d^{-1/4} \leq 1/4$ and $\varphi_M/\varepsilon = d^{2\eta_M}/L \to \infty$. Thus Proposition 28 applies. Since $T = Md$, $\log\log T = \Theta(\log\log d)$. □

Corollary 30 (Fixed logarithmic learning factors)**.** *For every fixed* $p \geq 0$*, no policy has a uniform constant-factor bound of the form* $R_T \leq C[H_G(\Delta)\log^p(eKT) + \Delta_{\mathrm{avg}}\sigma_{\max} + 1]$ *over this class of connected graphs and bounded i.i.d. instances, even with known horizon and delay. The worst-case ratio along a sequence of these instances is* $\Omega(\log\log T)$*.*

Proof. Repeat the finite construction with $L_p = \log^{p/2}(eKT)$, $\varepsilon = L_p/\sqrt{d}$, and $\Lambda_0 = L_p\sqrt{d}$ in place of $L, \varepsilon, \Lambda_0$. The complexity cap and induction use only $L_p^2 = \log^p(eKT)$ and condition $d^{\eta_M} \geq M(L_p/4+4)$. With $M = \lfloor \frac{1}{2}\log_8(\log d)\rfloor$, $L_p = \Theta((\log d)^{p/2})$ for fixed p, so the condition and gap restrictions hold for all sufficiently large d. The same $M/16$ ratio follows. □

The graph, gaps, and delay vary with the horizon in this family. Proposition 28 compares regret with the benchmark containing both the learning and delay terms, uniformly over the family. Whether $\Delta_G^{\pi}\sigma_{\max}$ is necessary remains open.

H.4 An unavoidable blind prefix

The permutation argument of Lancewicki et al. (2021, Appendix A.4 of the full version) also applies to any fixed feedback graph when the gaps are given as an unassigned multiset.

Proposition 31 (Blind-prefix arithmetic-average cost, cited in Section 5)**.** *Fix a known self-looped feedback graph on* K *vertices, a feasible multiset of gaps* $\{\Delta_1, \dots, \Delta_K\}$ *with one zero gap, a horizon* T*, and a constant nonnegative integer delay* d*. For every learner uninformed of the assignment of gaps to vertices, there is an assignment with deterministic losses such that*

$$R_T \geq \min\{T, d+1\}\Delta_{\mathrm{avg}}, \quad \textit{where} \quad \Delta_{\mathrm{avg}} = K^{-1}\sum_{i=1}^{K}\Delta_i.$$

Proof. Choose a uniform random permutation of the fixed loss levels, with minimum zero, across the graph vertices. Feedback generated in round s arrives only at the end of round $s + d$, so none is available before the first $T_0 = \min\{T, d + 1\}$ decisions. Conditional on the learner's randomization and on all information available for these decisions, each vertex has expected gap Δ_{avg} under the permutation. The expected regret over permutations in the first T_0 rounds is therefore $T_0\Delta_{\text{avg}}$. Later regret is nonnegative for every assignment. Averaging gives at least one deterministic assignment with the stated expected regret. □

This observation uses no graph property beyond the absence of feedback in the first T_0 decisions and does not compare the graph-dependent learning term or establish an additive joint lower bound.

I Implementation and storage

The statistical phase maintains K raw counts and sums, interval endpoints, gap estimates, test statistics, and the graph sampling vector. At every statistical decision, draw a fresh uniform permutation π_t, sort by $(g_{t-1,i}, \pi_t(i))$, and greedily cover out-neighborhoods. This costs $O(K \log K + |E|)$ per decision for either graph class. Computing all $Q_i(p_t)$ costs $O(K + |E|)$ after the permutation and cover have fixed p_t. Updating intervals and gap estimates after arrivals at a round costs $O(K)$, in addition to processing the revealed coordinates. The intervals are recomputed from retained empirical sums and counts in $O(K)$ at an epoch boundary.

The fresh permutation may change $\mathcal{C}_t$, J_t, and $\widehat{\rho}_t$ even when no feedback arrives and the gap estimates stay fixed. Rebuild the cover and evaluate the statistical test at every round. The sampling distribution uses the current t in its exploration rates. The test and the observation probabilities use the realized permutation, conditional on $\mathcal{F}^+_{t-1}$.

For each current-epoch source, the statistical phase stores its epoch and source indices, the scalar a_s fixed before drawing I_s, and the observation probabilities computed from the realized p_s for coordinates revealed by its action. The term $\ell_{s,I_s} - a_s$ is common to all residual coordinates and is stored as one shared offset; only the importance-weighted part requires coordinate updates. Arrival processing traverses those revealed coordinates once and costs $O(|N^{\text{out}}(I_s)|)$ for source s. A source-indexed queue can register and remove each source-round feedback record in constant amortized bookkeeping time, apart from coordinate processing. Algorithm 2 evaluates a softmax and graph observation probabilities in $O(K + |E|)$ time per trial. It makes at most $O(\log K)$ doubling trials per decision, and its age test removes each pending source at most once. Across T evaluated decisions, total work beyond external feedback delivery is therefore

$$O\Big(T(K + |E|) \log K + \sum_{s=1}^{T} |N^{\text{out}}(I_s)|\Big).$$

Let $\mathcal{P}_t^{\text{stored}}$ denote the current branch's source records still retained immediately after registration and before that round's arrivals. The algorithmic working memory beyond storage of the graph is

$$O\Big(K + \max_{t \le T} \sum_{s \in \mathcal{P}_t^{\text{stored}}} (1 + |N^{\text{out}}(I_s)|)\Big).$$

The constant term per statistical record includes a_s. The source record counts once regardless of how many coordinates its feedback reveals. This bound uses the branch's retained source-round records, not the original pending-feedback peak $\sigma_{\max}$ in the regret theorem; old-epoch records are discarded and records skipped by Algorithm 2 are removed. The record for the current action is included before a zero-delay arrival is processed.