NEARLY OPTIMAL BEST-OF-BOTH-WORLDS GUARANTEES FOR DELAYED MULTI-ARMED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study multi-armed bandits with delayed feedback. Our goal is to provide best-of-both-worlds regret guarantees, which is to say, guarantee optimal performance in both stochastic and adversarial environments. To this end, we propose a new version of Follow-the-Regularized-Leader (FTRL) with a time-varying hybrid regularizer combining Tsallis entropy and KL divergence. The novelty of the algorithm is in the learning rate for the KL divergence, which tracks the drift of the iterates of the algorithm. This allows the amount of regularization to respond directly to the drift induced by delayed observations. In the stochastic setting, our algorithm guarantees a regret of $O\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right)$, where $\Delta(i)$ denotes the suboptimality gap of arm i, T is the number of rounds, $d_{\max}$ is the maximum delay, and K is the number of actions. On the other hand, in the adversarial setting, the same algorithm guarantees a regret of $O\left(\sqrt{KT} + \sqrt{D\log K} + d_{\max}\sqrt{K}\right)$, where D is the total delay. The algorithm requires neither prior knowledge of the total delay D nor knowledge of whether the environment is stochastic or adversarial. In both the stochastic and adversarial settings, these guarantees are optimal up to the $d_{\max}\sqrt{K}$ term. Additionally, we compare our new algorithm with prior state-of-the-art algorithms through synthetic experiments.

1 INTRODUCTION

Delayed feedback is ubiquitous in sequential decision making problems. In applications such as medicine (Eick, 1988), feedback is not immediately available after making a decision. Consequently, algorithms have to issue a new decision before feedback on prior decisions is received. With bandit feedback, one has the additional challenge of only receiving feedback for the actions one chooses. In the multi-armed bandit setting, these challenges have been studied in a long line of work. Cesa-Bianchi et al. (2019); Thune et al. (2019); Bistritz et al. (2019); Zimmert & Seldin (2020) all consider the adversarial setting, in which both the losses and delays are chosen by an adversary, with Zimmert & Seldin (2020) providing an algorithm with optimal regret guarantees. If one assumes that the losses and delays are i.i.d. (called the stochastic setting), one can provide stronger guarantees on the performance of algorithms (see Joulani et al., 2013; Lancewicki et al., 2021; Wu & Wager, 2022).

However, algorithms designed for the adversarial setting may not obtain the stronger guarantees in the stochastic setting and, likewise, algorithms designed for the stochastic setting may not provide any guarantee at all in the adversarial setting. Unfortunately, in most sequential decision making problems we do not know a priori whether or not one is in an adversarial or stochastic setting. Best-Of-Both-Worlds (BOBW) algorithms aim to provide strong guarantees in whatever type of environment the algorithm is deployed in. In this paper, we continue a line of work on BOBW algorithms for sequential decision making problems with bandit and delayed feedback. We propose a new algorithm that improves upon the state-of-the-art regret guarantees.

The exact setting that we study is called multi-armed bandits with delayed feedback. Formally, this is a sequential setting where in each round $t = 1, \dots, T$, the learner picks an action $I_t \in \{1, \dots, K\} = [K]$, suffers loss $\ell_t(I_t)$, and subsequently observes losses $\ell_s(I_s)$ with s such that $s + d_s = t$ with $d_s \in [0, d_{\max}]$. We assume that $\ell_t(i) \in [0, 1]$ for all $i \in [K]$. The losses are either generated by an oblivious adversary or sampled i.i.d. from a fixed distribution. The former is referred to as the adversarial setting and the latter as the stochastic setting. In the stochastic setting, we denote

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.

by $\mu(i) = \mathbb{E}[\ell_t(i)]$ and we assume that the best arm is unique, i.e., there exists an $i^\star$ such that $\Delta(i) = \mu(i) - \mu(i^\star) > 0$ for all $i \in [K] \setminus i^\star$. The goal is to control the regret:

$$\mathcal{R}_T = \max_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_t(I_t) - \ell_t(i)\right)\right],$$

where the expectation is over the randomness of the learner and in the stochastic setting also over the randomness of the losses.

Our main contribution is a new best-of-both-worlds algorithm, Adaptive Tsallis-KL FTRL . In the adversarial setting, Adaptive Tsallis-KL FTRL guarantees that

$$\mathcal{R}_T = O\left(\sqrt{KT} + \sqrt{D \log K} + d_{\max}\sqrt{K}\right),$$

where $D = \sum_{t=1}^{T} d_t$. In the stochastic setting, Adaptive Tsallis-KL FTRL also guarantees that

$$\mathcal{R}_T = O\left(\sum_{i \neq i^\star} \frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right).$$

In both the adversarial and stochastic settings, our upper bounds are optimal up to the $d_{\max}\sqrt{K}$ term.

Our work is most closely related to Masoudian et al. (2022; 2024) and Schlisselberg et al. (2025b) (see also Table 1). Masoudian et al. (2022) develop a version of FTRL which in the adversarial setting guarantees a regret bound of order $\sqrt{KT} + \sqrt{D\log(K)} + d_{\max}K^{1/3}\log(K)$ and in the stochastic setting a regret bound of order $\sum_{i \neq i^\star}\left(\frac{\log T}{\Delta(i)} + \frac{\sigma_{\max}}{\Delta(i)\log(K)}\right) + d_{\max}K^{1/3}\log(K)$, where $\sigma_{\max}$ denotes the maximum number of outstanding observations. Masoudian et al. (2024) provide an algorithm with improved robustness to excessive delays. When disregarding this additional robustness, their algorithm guarantees $\mathcal{R}_T = O(\sqrt{KT} + \sqrt{D\log(K)} + \sigma_{\max})K$ in the adversarial setting and $\mathcal{R}_T = O(\sum_{i \neq i^\star}\left(\frac{\log T}{\Delta(i)} + \frac{\sigma_{\max}}{\Delta(i)\log(K)}\right) + \sigma_{\max}K)$ in the stochastic setting. We emphasize that the algorithm of Masoudian et al. (2024) provides considerably better guarantees in the presence of infrequent excessive delays, which we do not discuss for brevity. The regret bounds of Masoudian et al. (2022; 2024) for the adversarial setting are optimal, as shown by the lower bounds in Cesa-Bianchi et al. (2019); Masoudian et al. (2024). However, the lower bound for the stochastic setting is $\Omega(\sum_{i \neq i^\star}\left(\frac{\log T}{\Delta(i)} + \frac{d_{\max}\Delta(i)}{K}\right)$ (for constant delay $d_t = d_{\max}$) (Lancewicki et al., 2021). Hence, the $\frac{\sigma_{\max}}{\Delta(i)\log(K)}$ term in the upper bounds of Masoudian et al. (2022; 2024) is suboptimal.

Schlisselberg et al. (2025b) develop an algorithm which guarantees $\mathcal{R}_T = O(\sqrt{KT\log(T)} + \sqrt{D\log(K)} + \sigma_{\max}\log(K))$ in the adversarial setting and $\mathcal{R}_T = O(\sum_{i \neq i^\star}\left(\frac{\log T}{\Delta(i)} + \frac{\Delta(i)\sigma_{\max}}{K}\right))$ in the stochastic setting. This bound is optimal in the stochastic setting, but incurs an additional $\sqrt{\log T}$ factor in the leading adversarial term, which makes it suboptimal.

The algorithm of Schlisselberg et al. (2025b) is based on the SAPO algorithm of Auer & Chiang (2016), which initially assumes that the losses are stochastic and continuously monitors for deviations from this assumption. However, the tests for checking the assumption are quite involved, making the algorithm complicated. On the other hand, the algorithms of Masoudian et al. (2022; 2024) are instances of Follow-the-Regularized-Leader (FTRL) and are considerably simpler, at the cost of weaker guarantees in the stochastic setting. Schlisselberg et al. (2025b) explicitly ask whether the simpler FTRL framework can also yield (nearly) optimal guarantees in the stochastic setting. Our algorithm answers this question affirmatively: Adaptive Tsallis-KL FTRL is also an instance of FTRL and inherits its relative simplicity while providing nearly optimal guarantees in both the stochastic and adversarial settings. For sufficiently large T in the adversarial setting or sufficiently small $\Delta(i)$ in the stochastic setting, the regret bounds of Adaptive Tsallis-KL FTRL match or improve upon the previous state of the art for best-of-both-worlds multi-armed bandits with delayed feedback.

The main difficulty in the analysis of FTRL with delayed feedback is that the distribution over the actions at the time a loss is observed may differ from the distribution used when the corresponding action was sampled. Prior analyses of the FTRL algorithm controlled this drift in the action distribution via a conservative learning rate, which led to suboptimal results in the stochastic setting. Taking

Table 1: State-of-the-art best-of-both-worlds regret guarantees for delayed multi-armed bandits. Here, $\sigma_{\max}$ is the maximum number of outstanding observations.

Paper	Stochastic regret	Adversarial regret
Best-of-both-worlds bandits with delayed feedback		
Masoudian et al. (2022)	$O\left(\sum_{i\neq i^\star}\left(\frac{\log T}{\Delta(i)}+\frac{\sigma_{\max}}{\Delta(i)\log K}\right)+d_{\max}K^{1/3}\log(K)\right)$	$O\left(\sqrt{KT}+\sqrt{D\log K}+d_{\max}K^{1/3}\log(K)\right)$
Masoudian et al. (2024)	$O\left(\sum_{i\neq i^\star}\left(\frac{\log T}{\Delta(i)}+\frac{\sigma_{\max}}{\Delta(i)\log K}\right)+K\sigma_{\max}\right)$	$O\left(\sqrt{KT}+\sqrt{D\log K}+K\sigma_{\max}\right)$
Schlisselberg et al. (2025b)	$O\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)}+\frac{\sigma_{\max}}{K}\sum_{i\neq i^\star}\Delta(i)\right)$	$O\left(\sqrt{KT\log T}+\sqrt{D\log K}\right)$
This work	$O\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)}+d_{\max}\sqrt{K}\right)$	$O\left(\sqrt{KT}+\sqrt{D\log K}+d_{\max}\sqrt{K}\right)$

inspiration from AdaHedge (De Rooij et al., 2014) and its adaptation to the full information setting with delayed feedback (Flaspohler et al., 2021), we instead adaptively track the drift of the algorithm through its learning rate. Roughly speaking, the contribution of the delay to the regret is given by

$$\mathbb{E}[\lambda_{T+d_{\max}+1}\log(K)]+\mathbb{E}[\sum_{t=1}^{T}\underbrace{\langle p_t-p_t^\star,\widehat{\ell}_t\rangle}_{\text{drift}}], \tag{1}$$

where $\lambda_{T+d_{\max}+1}^{-1}$ is the learning rate for the KL part of our regularizer, $\widehat{\ell}_t$ is an importance weighted estimator of ℓ_t, and $p_t^\star$ is a clairvoyant iterate which incorporates all importance weighted losses before round t (which may not be available to use in round t). Normally, the analysis proceeds by showing that $\mathbb{E}[\sum_{t=1}^{T}\langle p_t-p_t^\star,\widehat{\ell}_t\rangle]\leq\sum_{t=1}^{T}\lambda_t^{-1}|m_t|$, where m_t is the set of indices of missing losses in round t. Now, a conservative tuning of λ_t here leads to optimal bounds in the adversarial setting and suboptimal bounds in the stochastic setting, as in Zimmert & Seldin (2020). Instead, we design a learning rate schedule that equalizes the two terms in equation (1) such that

$$\mathbb{E}[\lambda_{T+d_{\max}+1}\log(K)]+\mathbb{E}[\sum_{t=1}^{T}\langle p_t-p_t^\star,\widehat{\ell}_t\rangle]\approx 2\,\mathbb{E}[\sum_{t=1}^{T}\langle p_t-p_t^\star,\widehat{\ell}_t\rangle].$$

The induced learning rate is considerably less conservative than that of Masoudian et al. (2022), which allows us to develop improved guarantees in the stochastic setting. Intuitively, the regularization automatically increases only when the delayed feedback creates drift, rather than being chosen conservatively. The main technical challenge is that λ_t does not allow for a simple upper bound in terms of $\sqrt{D}$ as is the case for prior work. In the stochastic regime, we resolve this by using the self-bounding property of the regret together with a martingale argument, showing that the accumulated drift can be controlled by the regret itself and standard gap-dependent terms. In the adversarial setting, we rely on arguments similar to those in (De Rooij et al., 2014; Orabona & Pál, 2015; Flaspohler et al., 2021).

We complement our theoretical results with extensive synthetic experiments comparing Adaptive Tsallis-KL FTRL against previous state-of-the-art algorithms. Across most experimental settings, Adaptive Tsallis-KL FTRL achieves the best empirical performance.

The rest of the paper is organized as follows. In the remainder of the introduction, we discuss further related work and introduce the necessary notation for the rest of the paper. In Section 2 we introduce our algorithm and state the main result of the paper formally. Section 3 presents the analysis. We first establish the shared regret decomposition and stability properties that underlie both regimes, and then specialize the analysis to the stochastic and adversarial settings. Section 4 presents our experimental evaluation. Section 5 concludes with directions for future work.

Further Related Work The challenges induced by delayed feedback have been tackled in many different settings, which we review below.

In the full information setting with a discrete finite action set, (Weinberger & Ordentlich, 2002; Joulani et al., 2013) provide worst case guarantees. Similar worst case guarantees for convex losses

Algorithm 1: Adaptive Tsallis-KL FTRL

Input: Number of arms K, maximum delay $d_{\max}$

Set $\theta_1 \leftarrow 100\, d_{\max}(K-1)^{1/3}$, $\theta_2 \leftarrow 10^4 d_{\max}^2$, $p_1(i) = p_0(i) \leftarrow \frac{1}{K}\ \forall\, i \in [K]$, $S_1 \leftarrow 0$, and $\mathcal{O}_1 \leftarrow \emptyset$;

for $t = 1$ **to** T **do**

Sample arm $I_t \sim p_t$ and play it;

Set $\mathcal{O}_{t+1} \leftarrow \mathcal{O}_t \cup \{s \le t : s + d_s = t\}$;

for *each* s *satisfying* $s + d_s = t$ **do**

Observe $\ell_s(I_s)$ and set $\widehat{\ell}_s(i) \leftarrow \dfrac{\ell_s(i)\mathbf{1}\{I_s = i\}}{p_s(i)}$, for all $i \in [K]$;

Update $L_{t+1}^{\text{obs}} \leftarrow \sum_{s \in \mathcal{O}_{t+1}} \widehat{\ell}_s$;

Compute $S_{t+1} \leftarrow \sum_{s \in \mathcal{O}_{t+1}} (p_s - p_s^\star)^\top \widehat{\ell}_s \mathbb{1}[\max_{s' \le s} s' + d_{s'} \le t]$, where

$$L_s^\star \leftarrow \sum_{r=1}^{s-1} \widehat{\ell}_r \qquad\qquad p_s^\star \leftarrow \arg\min_{p \in \Delta^K} \left\{ p^\top L_s^\star + R_s(p) \right\};$$

Update $\lambda_{t+1} \leftarrow \theta_1 + \dfrac{1}{\log K} \max_{1 \le r \le t+1} S_r$ and $\gamma_{t+1} \leftarrow \sqrt{\theta_2 + t + 1}$;

Set $R_{t+1}(p) \leftarrow \lambda_{t+1} \mathrm{KL}(p, p_0) - 2\gamma_{t+1} \sum_{i=1}^K \sqrt{p(i)}$;

Update $p_{t+1} \leftarrow \arg\min_{p \in \Delta^K} \left\{ p^\top L_{t+1}^{\text{obs}} + R_{t+1}(p) \right\}$

and convex action sets are obtained by Quanrud & Khashabi (2015); Joulani et al. (2016). More adaptive guarantees for both action sets are obtained by Flaspohler et al. (2021).

In the stochastic bandit setting, delayed feedback has been studied under many different action sets: in Markov decision processes (Howson et al., 2023a), linear bandits (Vernade et al., 2020; Howson et al., 2023b; Schlisselberg et al., 2025a), and many other variations of (multi-armed) bandits (Agarwal & Duchi, 2012; Dudik et al., 2011; Vernade et al., 2017; Pike-Burke et al., 2018; Zhou et al., 2019).

In the adversarial bandit setting, delayed feedback has also been studied under different action sets: in MDPs (Katsikopoulos & Engelbrecht, 2003; Walsh et al., 2009; Dai et al., 2022; Lancewicki et al., 2023; Zierahn et al., 2025), linear bandits (Ito et al., 2020; Huang et al., 2023; Zierahn et al., 2025), and many other variations (Neu et al., 2014; Cesa-Bianchi et al., 2022; Levy et al., 2026)

Another line of work considers settings in which the delays are coupled with the losses or rewards; see (Gael et al., 2020; Van der Hoeven & Bianchi, 2022; Schlisselberg et al., 2025a; Zhang et al., 2025). These works are both on adversarial and stochastic losses.

Notation We define the set of observations available before round t by $O_t := \{s \in 1, \dots, t-1 : s + d_s < t\}$. The corresponding set of pending observations is $P_t := \{s \in 1, \dots, t-1 : s + d_s \ge t\}$. Thus $O_t \cup P_t = 1, \dots, t-1$, and under the bounded-delay assumption, $|P_t| \le d_{\max}$. We define $\Phi_t(p, L) := \langle p, L \rangle + R_t(p)$, with $p \in \Delta^K$.. $\Delta^K = \{p : p(i) > 0, \sum_{i=1}^K p(i) = 1\}$ denotes the simplex Furthermore, we define: $L_t^{\text{obs}} := \sum_{s \in O_t} \widehat{\ell}_s$ and $L_t^\star := \sum_{s=1}^{t-1} \widehat{\ell}_s = L_t^{\text{obs}} + \sum_{s \in P_t} \widehat{\ell}_s$., where $\widehat{\ell}_s$ is the standard importance weighted estimator. Let $(\mathcal{F}_t)_{t \ge 1}$ denote the filtration containing the learner's history up to the beginning of round t, together with all loss vectors ℓ_s, $s < t$, including those whose feedback has not yet been revealed to the learner.

2 Algorithm

Our algorithm (Algorithm 1) is based on FTRL with importance-weighted loss estimates. At round t, the learner samples $I_t \sim p_t$, where p_t is the distribution

$$p_t = \arg\min_{p \in \Delta^K} \left\{ p^\top L_t^{\text{obs}} + R_t(p) \right\},$$

L_t^{obs} contains the importance weighted loss estimates for rounds s such that $s + d_s < t$ (i.e. rounds for which the feedback has arrived before round t), and $R_t(p)$ is the regularizer (specified below). The learner then receives the feedback from rounds s such that $s + d_s = t$ and forms the standard importance-weighted estimator $\widehat{\ell}_s(i) = \frac{\ell_s(i)\mathbf{1}\{I_s=i\}}{p_s(i)}$. We use the regularizer

$$R_t(p) = \lambda_t \mathrm{KL}(p, p_0) - 2\gamma_t \sum_{i=1}^{K} \sqrt{p(i)}\,,$$

which combines KL regularization with Tsallis entropy. The Tsallis term provides the curvature needed to control the variance of the bandit loss estimates, while the KL term is used to control the additional instability caused by delayed feedback. The main difference from standard FTRL algorithms designed to obtain BOBW results is the choice of the learning rate of the regularizer. Indeed, our algorithm is strongly related to the algorithms of Zimmert & Seldin (2020); Masoudian et al. (2022). The main difference from these algorithms is the choice of the learning rate for the KL term in the regularizer. Specifically, where prior works used $\lambda_t \approx \sqrt{\sum_{s=1}^{t} |P_s| / \log(K)}$, we explicitly account for the drift induced by delayed feedback, allowing for a less conservative learning rate. To understand our choice of learning rate, we need to introduce the clairvoyant iterate $p_t^\star = \arg\min_{p\in\Delta^K} \left\{p^\top L_t^\star + R_t(p)\right\}$, where $L_t^\star$ additionally contains the importance weighted loss estimates that are not yet observed at the beginning of round t due to delayed feedback. Thus, $p_t^\star$ represents the distribution that FTRL would use at round t if all feedback from previous rounds were available. The discrepancy between p_t and $p_t^\star$ therefore isolates the effect of delayed information. For every round whose feedback has arrived, we measure this discrepancy through $X_s = (p_s - p_s^\star)^\top \widehat{\ell}_s$, known in the literature as the drift. This drift is then used in the learning rate as follows:

$$\lambda_{t+1} = \theta_1 + \frac{1}{\log K} \max_{1\le r\le t+1} S_r \;\text{ where }\; S_{t+1} \leftarrow \sum_{s\in\mathcal{O}_{t+1}} (p_s - p_s^\star)^\top \widehat{\ell}_s \mathbb{1}[\max_{s'\le s} s' + d_{s'} \le t]\,,$$

and $\theta_1 = O(d_{\max}(K-1)^{1/3})$. In other words, λ_t is chosen to be the running maximum of the accumulated observable drift. Consequently, periods of large delay-induced instability increase λ_t, which strengthens the KL regularization and makes subsequent FTRL iterates more stable. Thus, the key feature is that the additional regularization is determined by the instability actually experienced by the algorithm, rather than by a learning-rate schedule based only on time or cumulative delay.

This adaptive learning rate is the crucial new ingredient required to derive our main result below.

Theorem 1 (Main Result)**.** *Algorithm 1 guarantees that*

$$\mathcal{R}_T \le \begin{cases} O\left(\sum_{i\neq i^\star} \frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right) & \textit{stochastic setting} \quad (2) \\ O\left(\sqrt{KT} + \sqrt{D\log K} + d_{\max}\sqrt{K}\right) & \textit{adversarial setting} \quad (3) \end{cases}$$

Theorem 1 shows that Algorithm 1 achieves nearly optimal regret guarantees simultaneously in the stochastic and adversarial regimes. In the stochastic setting, the leading gap-dependent term $O\left(\sum_{i\neq i^\star} \frac{\log T}{\Delta(i)}\right)$ matches the classical logarithmic dependence in the no-delay setting, with delayed feedback contributing only an additive term of $O(d_{\max}\sqrt{K})$. In particular, this removes the additional $\sigma_{\max}/\Delta(i)$-dependent terms appearing in the FTRL-based guarantees of Masoudian et al. (2022; 2024). In the adversarial setting, the bound $O\left(\sqrt{KT} + \sqrt{D\log K} + d_{\max}\sqrt{K}\right)$ retains the optimal leading $\sqrt{KT}$ and total-delay dependence achieved by previous FTRL approaches, while avoiding the additional $\sqrt{\log T}$ factor in the leading term of Schlisselberg et al. (2025b). Thus, up to the additive $d_{\max}\sqrt{K}$ term, a single FTRL algorithm recovers the optimal dependence in both regimes. This answers the question raised by Schlisselberg et al. (2025b) of whether the comparatively simple FTRL framework can provide nearly optimal BOBW guarantees under delayed bandit feedback.

3 Analysis

The analysis consists of three parts. In the first, we present a regret bound for our algorithm that will serve as the basis for the rest of the analysis and show that the iterates of Algorithm 1 are

multiplicatively close when there are $d_{\max}$ rounds between them. In the second, we prove equation (2). In the third, we prove equation (3). Throughout the analysis, we assume $d_{\max} \geq 1$, which is used only to simplify bounds involving the parameters $\theta_1 = 100 d_{\max}(K-1)^{1/3}$ and $\theta_2 = 10^4 d_{\max}^2$.

3.1 Shared Regret Bound and Stability

In this section, we develop the shared regret bound which will be used for the remainder of the analysis. We first show that the iterates of the algorithm are close in the following sense.

Corollary 2. *For any $i \in [K]$ and $s, t \in [T]$ such that $s \leq t$ and $t - s \leq d_{\max}$, we have that $p_t(i) \leq \frac{6}{5} p_s(i)$. Furthermore, for any $i \in [K]$ we have that $p^\star_{t,i} \leq \frac{6}{5} p_t(i)$.*

This corollary is a consequence of Lemmas 6 and 12 in Appendix A.1.1. These Lemmas are variations of the stability Lemmas of Masoudian et al. (2022). Corollary 2 shows that the action probabilities remain comparable while feedback is pending. This is crucial because $\widehat{\ell}_s$ is importance weighted using p_s, but may only be incorporated when the learner uses a later distribution p_t. The result therefore controls the instability introduced by delayed feedback. Likewise, the actual delayed iterate p_t remains close, coordinate-wise, to the clairvoyant iterate $p^\star_t$, which has access to all losses generated before round t. Hence, the information missing due to delayed feedback cannot cause the learner's distribution to drift too far from its immediate-feedback counterpart.

The following Lemma serves as the basis of our analysis.

Lemma 3 (Main Regret Decomposition)**.** *For every comparator $u \in \Delta^K$, Algorithm 1 guarantees that*

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(p_t - u)^\top \ell_t\right] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(p_t - p_t^\star)^\top \ell_t\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\Big(\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\Big)\right]}_{\text{(II)}} \\
\underbrace{+\,\mathbb{E}[\Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star)] + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}}_{\text{(III)}} \\
\underbrace{+\,\mathbb{E}\left[\sum_{t=1}^{T}\Big(R_t(p_{t+1}^\star) - R_{t+1}(p_{t+1}^\star)\Big)\right] - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}}_{\text{(IV)}}\,. \qquad (4)
\end{aligned}$$

Lemma 3 decomposes the regret into four terms that can be controlled separately. Term (I) captures the additional error caused by delayed feedback and is known as the drift, term (II) measures the one-step stability of the clairvoyant iterates, term (III) is a boundary term, and term (IV) accounts for the change in the time-varying regularizer. This decomposition isolates the additional regret caused by delayed feedback from the terms that would also appear in the standard FTRL analysis.

Let u be a point mass on an arbitrary $j \in [K]$. For each term in Lemma 3 we find the following bounds:

$$\begin{aligned}
(I) &\leq \mathbb{E}[\lambda_{T+d_{\max}+1}\log K] && \text{(Definition of } \lambda_{T+d_{\max}+1}) \\
(II) &\leq 25\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\neq j}\sqrt{\frac{p_t(i)}{t}}\right], && \text{(Lemma 13)} \\
(III) &\leq \mathbb{E}[\lambda_{T+1}\log K] + 2\sqrt{\theta_2+1}\sqrt{K}\,, && \text{(Lemma 14)} \\
(IV) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{i=1}^{K}\sqrt{p_{t+1}^\star(i)}\right] - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}\,. && \text{(Lemma 15)}
\end{aligned}$$

These Lemmas and their proofs are relatively standard and can be found, for example, in (Zimmert & Seldin, 2021; Masoudian et al., 2022; Lee & Orabona, 2025). As a consequence of these bounds, we

have that

$$\mathbb{E}\Big[\sum_{t=1}^{T}(p_t-u)^\top \ell_t\Big] \le \mathbb{E}[2\lambda_{T+d_{\max}+1}\log K] + \frac{131}{5}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\neq j}\sqrt{\frac{p_t(i)}{t}}\right] + 2\sqrt{\theta_2+1}\sqrt{K}$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sqrt{p^\star_{t+1}(j)}\right] - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}$$
$$\le \mathbb{E}[2\lambda_{T+d_{\max}+1}\log K] + \frac{131}{5}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\neq j}\sqrt{\frac{p_t(i)}{t}}\right] + 2\sqrt{\theta_2+1}\sqrt{K}\,, \tag{5}$$

where we used that u is a point mass on j, $p^\star_t(j)\le 1$, $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$, $2\gamma_{T+1}\ge 2\sqrt{T}$, and that $p^\star_{t+1}(i)\le \frac{36}{25}p_t(i)$ by Corollary 2, and therefore $\sqrt{p^\star_{t+1}(i)}\le \frac{6}{5}\sqrt{p_t(i)}$.

The term $\frac{131}{5}\mathbb{E}\left[\sum_{i\neq j}\sqrt{\frac{p_t(i)}{t}}\right]$ can be controlled using standard arguments in both the stochastic and adversarial setting. The term $2\sqrt{\theta_2+1}\sqrt{K}$ is of order $d_{\max}\sqrt{K}$. The remaining term, $\mathbb{E}[2\lambda_{T+d_{\max}+1}\log K]$, captures the drift due to delayed feedback and requires the most effort in the analyses that follow.

3.2 Stochastic Setting (Proof of equation (2))

Recall that in the stochastic setting there exists a unique $i^\star\in[K]$ such that for all $i\neq i^\star$ and all t, $\mu_t(i)-\mu_t(i^\star)=\Delta(i)>0$, where $\mu_t(i)=\mathbb{E}[\ell_t(i)]$. Our strategy is to show that the r.h.s. of equation (5) is of the form $c\mathbb{E}\Big[\sum_{t=1}^{T}(p_t-e_{i^\star})^\top\ell_t\Big]+B = c\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i\neq i^\star}p_t(i)\Delta(i)\Big]+B$, where $c\in(0,1)$, B will be determined later and $e_{i^\star}$ is a basis vector in direction $i^\star$. As a consequence of this bound, we have that the pseudoregret is bounded by $\frac{B}{1-c}$. We first present a bound on $\mathbb{E}[\lambda_{T+d_{\max}+1}\log(K)]$.

Lemma 4. *In the stochastic setting, we have that*

$$\mathbb{E}[\lambda_{T+d_{\max}+1}\log(K)] \le \log(K)\theta_1 + \frac{24}{\log 2}\sum_{i\neq i^\star}\frac{\log(T)}{\Delta(i)} + \sum_{t=1}^{T}\frac{1}{4}\mathbb{E}[\sum_{i\neq i^\star}p_t(i)\Delta(i)] + 4d_{\max}$$

The proof is deferred to Appendix A.2. Essentially, the proof follows from a particular application of Doob's maximal inequality (Williams, 1991, Section 14.11, p. 143) combined with showing that $\mathbb{E}\Big[\sum_{t=1}^{T}(p_t-p\star)^\top\ell_t\Big]\le\frac{1}{6}\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i\neq i^\star}p_t(i)\Delta(i)\Big]$. Next, we have that

$$\frac{131}{5}\sum_{t=1}^{T}\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}} \le \sum_{t=1}^{T}\sum_{i\neq i^\star}\frac{2^{12}}{\Delta(i)t} + \frac{1}{12}\sum_{t=1}^{T}\sum_{i\neq i^\star}p_s(i)\Delta(i) \qquad \text{(Young's inequality)}$$
$$\le \sum_{i\neq i^\star}\frac{2^{12}(1+\log(T))}{\Delta(i)} + \frac{1}{12}\sum_{t=1}^{T}\sum_{i\neq i^\star}p_s(i)\Delta(i)\,,$$

where the last inequality is a standard bound on the harmonic number. Thus, by applying the preceding upper bounds to the right hand side of equation (5) and using that $\mathbb{E}[\ell_t(i)]=\mu(i)$ and $\sum_{i=1}^{K}(p_t(i)-u(i))\ell_t(i^\star)=0$, we can see that

$$\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i\neq i^\star}p_s(i)\Delta(i)\Big] = \mathbb{E}\Big[\sum_{t=1}^{T}(p_t-u)^\top\ell_t\Big]$$
$$\le \sum_{i\neq i^\star}\frac{2^{13}(1+\log(T))}{\Delta(i)} + \frac{7}{12}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^\star}p_s(i)\Delta(i)\right] + 4d_{\max} + \log(K)\theta_1 + 2\sqrt{\theta_2+1}\sqrt{K}\,.$$

which implies that

$$\mathbb{E}\Big[\sum_{t=1}^{T}(p_t-u)^\top \ell_t\Big] \le \sum_{i\neq i^\star}\frac{2^{14}(1+\log(T))}{\Delta(i)} + 12d_{\max} + 3\log(K)\theta_1 + 6\sqrt{\theta_2+1}\sqrt{K}\,.$$

When we substitute θ_2 and θ_1 with their definitions, we have reached the conclusion of the proof of equation (2). A more detailed proof can be found in Appendix A.2.1.

3.3 Adversarial Setting (Proof of equation (3))

In this setting, the main challenge is bounding $\lambda_{T+d_{\max}+1}$, which we do in the following Lemma.

Lemma 5. *We have that* $\mathbb{E}[\lambda_{T+d_{\max}+1}\log(K)] \le \lambda_1\log(K) + 2\sqrt{\log(K)D}\,.$

The proof is deferred to Appendix A.3. The proof essentially follows the same steps as in the derivation of the regret bound of AdaHedge (De Rooij et al., 2014), the regret bound of AdaHedgeD (Flaspohler et al., 2021), and the regret bounds of (Orabona & Pál, 2015). Since $\sum_{i\neq j}\sqrt{p_{t+1}(i)} \le \sqrt{K}$ and $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$, we have that $\sum_{t=1}^{T}\sum_{i\neq j}\sqrt{\frac{p_t(i)}{t}} \le 2\sqrt{KT}$. Consequently, continuing from equation (5), we have that

$$\mathbb{E}\Big[\sum_{t=1}^{T}(p_t-u)^\top \ell_t\Big] \le 33\sqrt{KT} + 2\lambda_1\log(K) + 4\sqrt{\log(K)D} + 2\sqrt{(\theta_2+1)K}\,.$$

When we substitute θ_2 and θ_1 with their definitions, we have reached the conclusion of the proof of equation (3). A more detailed proof can be found in Appendix A.3.2.

4 Experiments

In this section, we compare state-of-the-art best-of-both-worlds algorithms for bandits with delayed feedback with Algorithm 1. We first describe our experimental setup and then discuss the results. All experiments mimic the experiments of Zimmert & Seldin (2021), but with various forms of delay.

Setup For all experiments we use $T = 50000$. We set up two types of loss generating processes and run experiments for $K \in \{2, 8, 16, 32\}$. The first process is a purely stochastic process (labeled as stochastic environment in the figures), where the optimal arm has mean loss $(1-\Delta)/2$ and every suboptimal arm has mean loss $(1+\Delta)/2$, so that all gaps equal Δ and the means are centered at $1/2$. We vary $\Delta \in \{0.03125, 0.0625, 0.125, 0.25\}$. The losses are sampled from a Bernoulli distribution. The second loss generating process is nearly a stochastic setting (labeled *alternating environment* in the figures), alternating between two stochastic settings with the same optimal arm. In the first, expected losses are 0 for the optimal arm and Δ for suboptimal arms. In the second, they are $1-\Delta$ and 1, respectively. We again vary $\Delta \in \{0.03125, 0.0625, 0.125, 0.25\}$. The first phase lasts 100 rounds and the time between consecutive alternations increases by a factor of 1.6 after each switch. Since the identity of the best arm and the gaps are invariant across phases, the per-round pseudo-regret of any action is the same as in the stochastic regime, while the loss sequence is non-stationary.

There are three delay generating processes. In the first, the delay is constant $d_t = 10$. In the second, the delay is sampled uniformly at random from $\{1, \dots, 50\}$. In the third, the delay is 5 with probability 0.9 and 50 with probability 0.1, which we call two-point delays.

For Algorithm 1 and the algorithms of Masoudian et al. (2022; 2024), we set all tuning constants in the parameters of these algorithms to 1. For the algorithm of Schlisselberg et al. (2025b), we made considerable efforts to find constants that make the algorithm competitive with Algorithm 1 and the algorithms of Masoudian et al. (2022; 2024), but were unsuccessful. We therefore omit its results. In the figures below, MZSDelayedFTRL and RobustDelayedBOBW denote the algorithms of Masoudian et al. (2022) and Masoudian et al. (2024), respectively.

Results The results are displayed in Figure 1 in the main body and Figures 2 to 6 in Appendix B. The results for the stochastic and alternating environments are not very different. The main difference

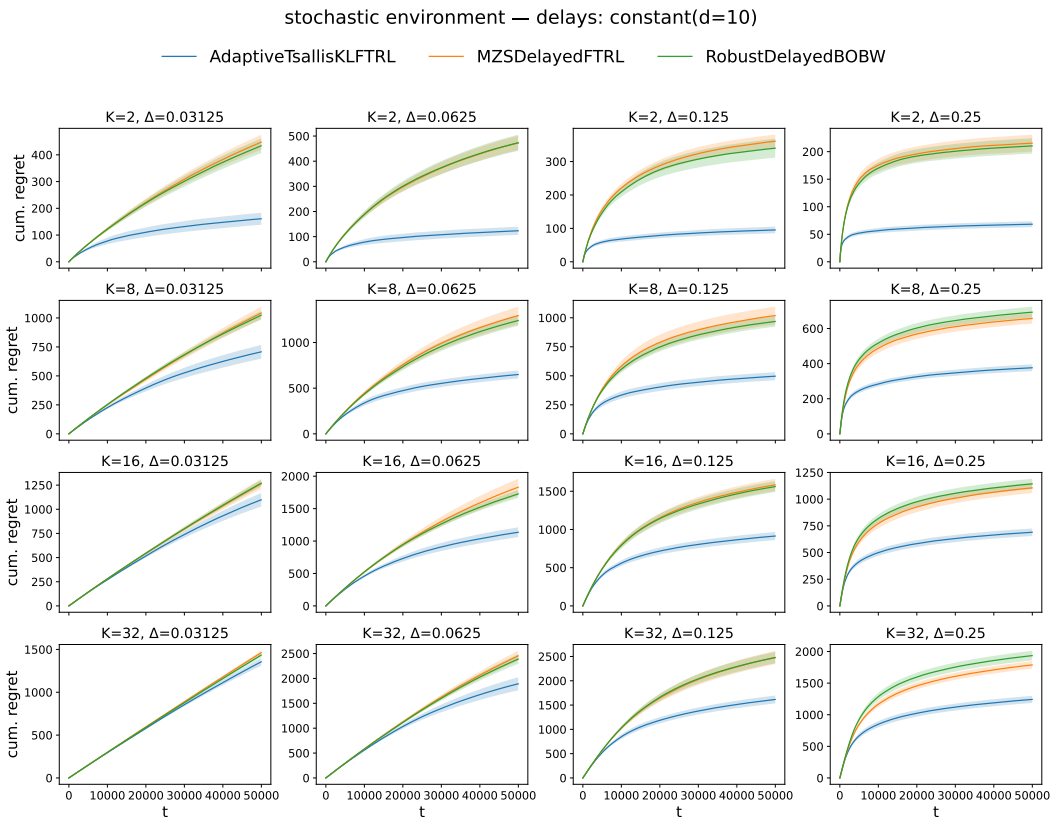


Figure 1: Results for the stochastic environment with constant ($d_t = 10$) delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.

between the results comes in the delay process. For constant (Figures 1 and 4) and uniform (Figures 2 and 5), the algorithms of Masoudian et al. (2022; 2024) have very close performances and Adaptive Tsallis-KL FTRL has considerably better performance. For example, with constant delay, at $T = 50000$ for $K = 16$, and $\Delta = 0.25$, Adaptive Tsallis-KL FTRL has about 2 times less estimated regret. That the algorithms of Masoudian et al. (2022; 2024) have very close performances is not surprising: Masoudian et al. (2024) set out to make the algorithm of Masoudian et al. (2022) robust to excessive delays, which does not happen for these delay processes. For the two-point delay process, the robustness of the algorithm of Masoudian et al. (2024) does help, as in some cases it outperforms the algorithm of Masoudian et al. (2022). For $K \in \{8, 16, 32\}$ and $\Delta = 0.25$, Adaptive Tsallis-KL FTRL is on par with the algorithm of Masoudian et al. (2024), but in all other cases Adaptive Tsallis-KL FTRL outperforms the other algorithms.

5 Conclusion and Future Work

In this work, we introduced Adaptive Tsallis-KL FTRL, a best-of-both-worlds algorithm for multi-armed bandits with delayed feedback. The algorithm combines Tsallis entropy with KL regularization and adapts the KL learning rate to the observed drift between the delayed and clairvoyant FTRL iterates. This allows the regularization to respond directly to instability caused by delayed feedback. We established nearly optimal regret guarantees in both stochastic and adversarial environments. Several questions remain open. Most importantly, both regret guarantees contain the additive term $d_{\max}\sqrt{K}$. It remains to determine whether this dependence is necessary or can be removed through a sharper stability analysis or a different choice of regularizer. A related direction is to remove the algorithm's prior dependence on $d_{\max}$ and instead adapt directly to the observed delay sequence. We suspect that techniques used by Masoudian et al. (2024) may be useful for this endeavor. Finally, it would be interesting to extend the drift-adaptive regularization principle beyond multi-armed bandits.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs (of the form: "carefully check this proof"), feedback on written work, and assistance in Python code generation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for transcribing of handwritten notes to LaTeX. We have reviewed all AI-assisted work: LLM-generated code was verified and tested for correctness by one of the authors.

We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Clear explanations of any assumptions and a complete proof of the claims are either in the Appendix or the main body of the paper. All experimental details, the implementation of Adaptive Tsallis-KL FTRL and the comparison baselines, and the code required to reproduce all results can be found in the supplementary material. A complete description of the data processing steps is in the paper (Section 4).

REFERENCES

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *IEEE Conference on Decision and Control*, pp. 5451–5452, 2012.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 116–120. PMLR, 2016.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose H. Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11345–11354. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper/2019/hash/ae2a2db40a12ec0131d48acc1218d2ef-Abstract.html.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38, 2019.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. *Journal of Machine Learning Research*, 23(277):1–24, 2022.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial Markov Decision Processes with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11437–11449. Curran Associates, Inc., 2022.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.

Stephen G. Eick. The two-armed bandit with delayed responses. *The Annals of Statistics*, 1988.

Genevieve E Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *International Conference on Machine Learning*, pp. 3363–3373. PMLR, 2021.

Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pp. 3348–3356, 2020.

Dirk van der Hoeven and Nicolò Cesa Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Optimism and delays in episodic reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6061–6094, 2023a.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *International Conference on Artificial Intelligence and Statistics*, pp. 6095–6119, 2023b.

Jiatai Huang, Yan Dai, and Longbo Huang. Banker online mirror descent: A universal approach for delayed online bandit learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13814–13844. PMLR, 23–29 Jul 2023.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In *Advances in Neural Information Processing Systems*, pp. 4872–4883, 2020.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *ICML*, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *AAAI Conference on Artificial Intelligence*, 2016.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5969–5978. PMLR, 2021. URL https://proceedings.mlr.press/v139/lancewicki21a.html.

Tal Lancewicki, Aviv Rosenberg, and Dmitry Sotnikov. Delay-adapted policy optimization and improved regret for adversarial MDP with delayed bandit feedback. In *International Conference on Machine Learning*, pp. 18482–18534. PMLR, 2023.

Wei-Cheng Lee and Francesco Orabona. A best-of-both-worlds proof for tsallis-inf without fenchel conjugates. *arXiv preprint arXiv:2511.11211*, 2025.

Orin Levy, Liad Erez, Alon Peled-Cohen, and Yishay Mansour. Regret bounds for adversarial contextual bandits with general function approximation and delayed feedback. *Advances in Neural Information Processing Systems*, 38:31674–31705, 2026.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. (arXiv:2206.14906), 2022. doi: 10.48550/arXiv.2206.14906. URL http://arxiv.org/abs/2206.14906. arXiv:2206.14906 [cs].

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 37 (NeurIPS)*. Neural Information Processing Systems Foundation, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/ff54269c8c51ae379822f5f655b0f9a7-Abstract-Conference.html.

Dragoslav S. Mitrinović and Petar M. Vasić. *Analytic Inequalities*, volume 165 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1970.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov Decision Processes under bandit feedback. *IEEE Trans. Automat. Contr.*, 59(3):676–691, 2014.

Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. In *International Conference on Algorithmic Learning Theory*, pp. 287–301. Springer, 2015.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113, 2018.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, pp. 1270–1278, 2015.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the Leader if you can, Hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

Ofir Schlisselberg, Ido Cohen, Tal Lancewicki, and Yishay Mansour. Delay as payoff in mab. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20310–20317, 2025a.

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 38 (NeurIPS)*. Neural Information Processing Systems Foundation, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02f0ac0a323dc17d964d4bbf8a62e01b-Abstract-Conference.html.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pp. 9712–9721, 2020.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83–105, 2009.

Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Han Wu and Stefan Wager. Thompson sampling with unrestricted delays. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 937–955, 2022.

Mengxiao Zhang, Yingfei Wang, and Haipeng Luo. Contextual linear bandits with delay as payoff. *arXiv preprint arXiv:2502.12528*, 2025.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pp. 5197–5208, 2019.

Lukas Zierahn, Dirk Van Der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolo Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. *Journal of Machine Learning Research*, 26(104):1–60, 2025.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3285–3294. PMLR, 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.

A Deferred Proofs and Technical Results

A.1 Shared regret bound and stability

A.1.1 Stability

The proof of Lemma 6 separates the change from p_u to p_{u+1} into two effects. We first control the change caused by incorporating newly observed losses while keeping the regularizer fixed, and then control the change caused by updating the regularizer while keeping the cumulative loss vector fixed. Combining these two bounds over at most $d_{\max}$ consecutive rounds yields the desired stability result.

Lemma 6. *For any $i \in [K]$ and $s, t \in [T]$ such that $s \le t$ and $t - s \le d_{\max}$, we have that*

$$p_t(i) \le \frac{6}{5} p_s(i).$$

Proof. We proceed by strong induction on t. Fix $t \ge 1$ and assume that for every $t' < t$, every $s' \le t'$ satisfying $t' - s' \le d_{\max}$, and every $i \in [K]$,

$$p_{t',i} \le \frac{6}{5} p_{s',i}.$$

Fix $s \le t$ such that $t - s \le d_{\max}$. If $s = t$, the claim follows directly. For the base case, $t = 1$, necessarily $s = 1$; the claim is immediate.

For every $u \in \{s, \dots, t-1\}$, define

$$A_u := O_{u+1} \setminus O_u.$$

Fix $q \in A_u$. By the definition of the observation sets, $q + d_q = u$. Hence

$$q \le u \qquad \text{and} \qquad u - q = d_q \le d_{\max}.$$

Since $u < t$, the induction hypothesis applied with $t' = u$ and $s' = q$ gives

$$p_u(j) \le \frac{6}{5} p_q(j), \qquad \forall j \in [K].$$

Thus the assumptions of Lemma 8 are satisfied for every $u \in \{s, \dots, t-1\}$.

For $u \ge 1$, we define

$$C_u^{\mathrm{reg}} := \left(\frac{\lambda_{u+1} - \lambda_u}{\gamma_u} + \frac{\gamma_{u+1}}{\gamma_u} \right)^2.$$

By Lemma 8,

$$\bar{p}_{u+1}(i) \le \frac{1}{1 - \frac{1}{\lambda_u} B_u} p_u(i),$$

while Lemma 10 gives

$$p_{u+1}(i) \le C_u^{\mathrm{reg}} \bar{p}_{u+1}(i).$$

Therefore,

$$p_{u+1}(i) \le \frac{C_u^{\mathrm{reg}}}{1 - \frac{1}{\lambda_u} B_u} p_u(i).$$

Dividing by $p_u(i) > 0$ and multiplying over $u = s, \dots, t-1$ gives

$$\frac{p_t(i)}{p_s(i)} = \prod_{u=s}^{t-1} \frac{p_{u+1}(i)}{p_u(i)} \le \prod_{u=s}^{t-1} \frac{C_u^{\mathrm{reg}}}{1 - \frac{B_u}{\lambda_u}}.$$

Therefore,

$$\frac{p_t(i)}{p_s(i)} \le \left(\prod_{u=s}^{t-1} \frac{1}{1 - \frac{B_u}{\lambda_u}} \right) \left(\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} \right).$$

By Lemma 9,

$$\prod_{u=s}^{t-1} \frac{1}{1-\frac{1}{\lambda_u}B_u} \le \frac{125}{119},$$

and by Lemma 11,

$$\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} < 1.06.$$

Consequently,

$$\frac{p_t(i)}{p_s(i)} < \frac{125}{119} \cdot 1.06 < \frac{6}{5}.$$

Therefore,

$$p_t(i) \le \frac{6}{5} p_s(i).$$

This closes the induction.

□

The proof follows the same general strategy as Lemma 4 of Masoudian et al. (2022). This lemma is an analogue of their Lemma 17.

Lemma 7. *Assume that for some $\alpha \ge 1$, $s \le t$ and $t - s \le d_{\max}$ we have*

$$p_t(i) \le \alpha p_s(i), \qquad \forall i \in [K].$$

Then we have that

$$\frac{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_s(j)}{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1}} \le 2\alpha(K-1)^{1/3}.$$

Proof. We first compute the curvature of the regularizer. Since

$$r_{t,i}(x) = \lambda_t \mathrm{KL}(x, p_0) - 2\gamma_t\sqrt{x},$$

we have

$$r''_{t,i}(x) = \frac{\lambda_t}{x} + \frac{\gamma_t}{2x^{3/2}}.$$

Define

$$a = \frac{\gamma_t}{2}, \qquad b = \lambda_t.$$

Then

$$r''_{t,i}(x) = ax^{-3/2} + bx^{-1}.$$

Let I_s be the arm selected at round s. Since

$$\widehat{\ell}_{s,j} = \frac{\ell_{s,j}\mathbf{1}\{I_s = j\}}{p_s(j)},$$

only the coordinate $j = I_s$ contributes. Let

$$z = p_t(I_s).$$

Then

$$\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_{s,j}}{\sum_j r''_{t,j}(p_t(j))^{-1}} = \frac{r''_{t,I_s}(z)^{-1}\ell_{s,I_s}p_s(I_s)^{-1}}{\sum_j r''_{t,j}(p_t(j))^{-1}}.$$

Using $\ell_{s,I_s} \le 1$ and the drift assumption

$$p_t(I_s) \le \alpha p_s(I_s),$$

we obtain

$$p_s(I_s)^{-1} \le \alpha z^{-1}.$$

Hence

$$\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_{s,j}}{\sum_j r''_{t,j}(p_t(j))^{-1}} \le \frac{\alpha r''_t(z)^{-1}z^{-1}}{\sum_j r''_t(p_t(j))^{-1}}.$$

We first verify that the function $x \mapsto r''_t(x)^{-1}$ is convex. Recall that

$$r''_t(x) = ax^{-3/2} + bx^{-1},$$

where $a, b \ge 0$. Therefore

$$r''_t(x)^{-1} = \frac{1}{ax^{-3/2} + bx^{-1}} = \frac{x^{3/2}}{a + b\sqrt{x}}.$$

Differentiating twice yields

$$\left(r''_t(x)^{-1}\right)'' = \frac{a(3a + b\sqrt{x})}{4\sqrt{x}(a + b\sqrt{x})^3}.$$

Since $x > 0$ and $a, b \ge 0$, both the numerator and denominator are nonnegative, and the denominator is strictly positive. Hence

$$\left(r''_t(x)^{-1}\right)'' \ge 0.$$

Thus $x \mapsto r''_t(x)^{-1}$ is convex on $(0, \infty)$. Since $x \mapsto r''_t(x)^{-1}$ is convex for $r''_t(x) = ax^{-3/2} + bx^{-1}$, Jensen's inequality gives

$$\sum_{j \ne I_s} r''_t(p_t(j))^{-1} \ge (K-1) r''_t\left(\frac{1-z}{K-1}\right)^{-1}.$$

This gives us the following:

$$\begin{aligned}
\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_{s,j}}{\sum_j r''_{t,j}(p_t(j))^{-1}} &= \frac{r''_{t,I_s}(p_t(I_s))^{-1}\,\ell_s(I_s)\,p_s(I_s)^{-1}}{\sum_j r''_{t,j}(p_t(j))^{-1}} && (6)\\
&\le \frac{r''_{t,I_s}(p_t(I_s))^{-1}\,p_s(I_s)^{-1}}{\sum_j r''_{t,j}(p_t(j))^{-1}} && (7)\\
&\le \frac{\alpha\, r''_{t,I_s}(z)^{-1}z^{-1}}{\sum_j r''_{t,j}(p_t(j))^{-1}} && (8)\\
&= \frac{\alpha}{(az^{-1/2} + b)\sum_j r''_{t,j}(p_t(j))^{-1}}, && (9)
\end{aligned}$$

For the denominator, we have:

$$\begin{aligned}
\sum_j r''_{t,j}(p_t(j))^{-1} &= r''_t(z)^{-1} + \sum_{j \ne I_s} r''_t(p_t(j))^{-1}\\
&\ge r''_t(z)^{-1} + (K-1) r''_t\left(\frac{1-z}{K-1}\right)^{-1}\\
&= \frac{1}{az^{-3/2} + bz^{-1}} + \frac{K-1}{a\left(\frac{1-z}{K-1}\right)^{-3/2} + b\left(\frac{1-z}{K-1}\right)^{-1}}\\
&= \frac{z}{az^{-1/2} + b} + \frac{1-z}{a\sqrt{K-1}(1-z)^{-1/2} + b}.
\end{aligned}$$

Substituting this into (9):

$$\begin{aligned}
\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_{s,j}}{\sum_j r''_{t,j}(p_t(j))^{-1}} &\le \frac{\alpha}{(az^{-1/2} + b)\left[\frac{z}{az^{-1/2}+b} + \frac{1-z}{a\sqrt{K-1}(1-z)^{-1/2}+b}\right]}\\
&= \alpha\left[z + (1-z)\frac{az^{-1/2} + b}{a\sqrt{K-1}(1-z)^{-1/2} + b}\right]^{-1}.
\end{aligned}$$

We now distinguish two cases.

Case 1: $z \le 1/K$**.** Then

$$\sqrt{K-1}(1-z)^{-1/2} \le z^{-1/2}.$$

Hence

$$\frac{az^{-1/2}+b}{a\sqrt{K-1}(1-z)^{-1/2}+b} \ge 1.$$

Thus the bracketed term is at least 1, and therefore

$$\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_{s,j}}{\sum_j r''_{t,j}(p_t(j))^{-1}} \le \alpha$$

Case 2: $z \ge 1/K$**.** Since $z \ge 1/K$, we have

$$z^{-1/2} \le \sqrt{K-1}\,(1-z)^{-1/2}.$$

Therefore, using $b \ge 0$,

$$\frac{az^{-1/2}+b}{a\sqrt{K-1}(1-z)^{-1/2}+b} \ge \frac{az^{-1/2}}{a\sqrt{K-1}(1-z)^{-1/2}} = \frac{(1-z)^{1/2}}{\sqrt{K-1}\,z^{1/2}}.$$

Hence the bracketed term is at least

$$z + (1-z)\frac{(1-z)^{1/2}}{\sqrt{K-1}\,z^{1/2}} = z + \frac{(1-z)^{3/2}}{\sqrt{K-1}\,z^{1/2}}.$$

Thus

$$\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_s(j)}{\sum_j r''_{t,j}(p_t(j))^{-1}} \le \alpha\left(z + \frac{(1-z)^{3/2}}{\sqrt{K-1}\,z^{1/2}}\right)^{-1}.$$

We now lower bound the bracket. If

$$z \ge \frac{1}{(K-1)^{1/3}+1},$$

then

$$\left(z + \frac{(1-z)^{3/2}}{\sqrt{K-1}\,z^{1/2}}\right)^{-1} \le z^{-1} \le (K-1)^{1/3}+1 \le 2(K-1)^{1/3}.$$

Otherwise,

$$z \le \frac{1}{(K-1)^{1/3}+1}.$$

Since $K-1 \ge 1$, this implies $1-z \ge 1/2$. Hence

$$z + \frac{(1-z)^{3/2}}{\sqrt{K-1}\,z^{1/2}} \ge z + \frac{1}{\sqrt{8}}(K-1)^{-1/2}z^{-1/2}.$$

By the weighted AM–GM inequality,

$$z + \frac{1}{\sqrt{8}}(K-1)^{-1/2}z^{-1/2} \ge c_0(K-1)^{-1/3},$$

for a universal constant $c_0 > 1/2$. Therefore,

$$z + \frac{1}{\sqrt{8}}(K-1)^{-1/2}z^{-1/2} \ge \frac{1}{2}(K-1)^{-1/3}.$$

Consequently,

$$\left(z + \frac{(1-z)^{3/2}}{\sqrt{K-1}\,z^{1/2}}\right)^{-1} \le 2(K-1)^{1/3}.$$

Combining the two subcases gives

$$\frac{\sum_j r''_{t,j}(p_t(j))^{-1}\widehat{\ell}_s(j)}{\sum_j r''_{t,j}(p_t(j))^{-1}} \le 2\alpha(K-1)^{1/3}.$$

□

Lemma 8. *Suppose that, for every $q \in A_u$,*

$$p_u(j) \le \frac{6}{5} p_q(j), \qquad \forall j \in [K].$$

Then, for every $i \in [K]$,

$$\bar{p}_{u+1}(i) \le \frac{p_u(i)}{1 - \frac{1}{\lambda_u} B_u},$$

where

$$B_u := \frac{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1} \sum_{q \in A_u} \widehat{\ell}_q(j)}{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1}}$$

satisfies

$$B_u \le \frac{12}{5} |A_u| (K-1)^{1/3}.$$

Proof. For every $u \ge 1$, define

$$A_u := O_{u+1} \setminus O_u$$

and the intermediate iterate

$$\bar{p}_{u+1} := \arg\min_{p \in \Delta^K} \left\{ \langle p, L^{\mathrm{obs}}_{u+1} \rangle + R_u(p) \right\}.$$

Since

$$L^{\mathrm{obs}}_{u+1} - L^{\mathrm{obs}}_u = \sum_{q \in A_u} \widehat{\ell}_q,$$

the iterates p_u and $\bar{p}_{u+1}$ use the same regularizer R_u and differ only in their cumulative loss vectors. By the KKT conditions, there exist $\mu, \bar{\mu} \in \mathbb{R}$ such that

$$r'_{u,i}(p_u(i)) = -L^{\mathrm{obs}}_u(i) + \mu$$

and

$$r'_{u,i}(\bar{p}_{u+1}(i)) = -L^{\mathrm{obs}}_{u+1}(i) + \bar{\mu}.$$

Subtracting gives

$$r'_{u,i}(p_u(i)) - r'_{u,i}(\bar{p}_{u+1}(i)) = \sum_{q \in A_u} \widehat{\ell}_q(i) + \mu - \bar{\mu}. \tag{10}$$

Since $r'_{u,i}$ is concave,

$$r'_{u,i}(\bar{p}_{u+1}(i)) \le r'_{u,i}(p_u(i)) + r''_{u,i}(p_u(i)) \big(\bar{p}_{u+1}(i) - p_u(i)\big). \tag{11}$$

Combining equation (11) with equation (10) gives

$$\bar{p}_{u+1}(i) - p_u(i) \ge r''_{u,i}(p_u(i))^{-1} \left(\bar{\mu} - \mu - \sum_{q \in A_u} \widehat{\ell}_q(i) \right).$$

Summing over i and using $\sum_i \bar{p}_{u+1}(i) = \sum_i p_u(i) = 1$ yields

$$\bar{\mu} - \mu \le \frac{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1} \sum_{q \in A_u} \widehat{\ell}_q(j)}{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1}} = B_u.$$

If $\bar{p}_{u+1}(i) \le p_u(i)$, the desired inequality is immediate since $B_u \ge 0$. Hence suppose that $\bar{p}_{u+1}(i) > p_u(i)$. By concavity of $r'_{u,i}$,

$$r'_{u,i}(\bar{p}_{u+1}(i)) - r'_{u,i}(p_u(i)) \ge r''_{u,i}(\bar{p}_{u+1}(i))\big(\bar{p}_{u+1}(i) - p_u(i)\big).$$

Together with the KKT identities, this gives

$$r''_{u,i}(\bar{p}_{u+1}(i))\big(\bar{p}_{u+1}(i) - p_u(i)\big) \le \bar{\mu} - \mu - \sum_{q \in A_u} \widehat{\ell}_q(i) \le B_u.$$

Since

$$r''_{u,i}(x) \ge \frac{\lambda_u}{x},$$

we obtain

$$\bar{p}_{u+1}(i) - p_u(i) \le \frac{1}{\lambda_u} B_u \bar{p}_{u+1}(i).$$

Therefore,

$$\bar{p}_{u+1}(i) \le \frac{p_u(i)}{1 - \frac{1}{\lambda_u} B_u}.$$

Note that since $|A_u| \le d_{\max}$ and

$$\lambda_u \ge \theta_1 = 100 d_{\max} (K-1)^{1/3},$$

we have

$$\frac{B_u}{\lambda_u} \le \frac{12}{500} < 1.$$

Thus $1 - B_u/\lambda_u > 0$.

It remains to bound B_u. By linearity,

$$B_u = \sum_{q \in A_u} \frac{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1} \widehat{\ell}_q(j)}{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1}}.$$

For every $q \in A_u$, by assumption,

$$p_u(j) \le \frac{6}{5} p_q(j), \qquad \forall j \in [K].$$

Moreover,

$$q \le u, \qquad u - q \le d_{\max}.$$

Hence Lemma 7 applies with $t = u$, $s = q$, and $\alpha = 6/5$, and yields

$$\frac{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1} \widehat{\ell}_q(j)}{\sum_{j=1}^{K} r''_{u,j}(p_u(j))^{-1}} \le \frac{12}{5}(K-1)^{1/3}.$$

Summing over $q \in A_u$ gives

$$B_u \le \frac{12}{5} |A_u| (K-1)^{1/3}.$$

□

Lemma 9. *Let $s \le t$ with $t - s \le d_{\max}$. Then*

$$\prod_{u=s}^{t-1} \frac{1}{1 - \frac{1}{\lambda_u} B_u} \le \frac{125}{119}.$$

Proof. By Lemma 8,

$$B_u \le \frac{12}{5}|A_u|(K-1)^{1/3}.$$

Since

$$\eta_u = \frac{1}{\lambda_u} \le \frac{1}{\theta_1} = \frac{1}{100 d_{\max}(K-1)^{1/3}},$$

we have

$$\frac{1}{\lambda_u} B_u \le \frac{12}{500}\frac{|A_u|}{d_{\max}}.$$

Therefore,

$$\sum_{u=s}^{t-1} \frac{1}{\lambda_u} B_u \le \frac{12}{500 d_{\max}} \sum_{u=s}^{t-1} |A_u|.$$

Since

$$A_u = O_{u+1} \setminus O_u$$

and the observation sets are nested, the sets A_u are pairwise disjoint and

$$\bigcup_{u=s}^{t-1} A_u = O_t \setminus O_s.$$

Hence

$$\sum_{u=s}^{t-1} |A_u| = |O_t \setminus O_s|.$$

Moreover,

$$O_t \setminus O_s \subseteq \{s, \dots, t-1\} \cup P_s,$$

and therefore

$$|O_t \setminus O_s| \le (t-s) + |P_s| \le 2d_{\max}.$$

Thus

$$\sum_{u=s}^{t-1} \frac{1}{\lambda_u} B_u \le \frac{24}{500}.$$

Finally, using the Weierstrass product inequality (Mitrinović & Vasić, 1970, p. 210),

$$\prod_{u=s}^{t-1}(1-x_u) \ge 1 - \sum_{u=s}^{t-1} x_u$$

for $x_u \in [0,1)$. Since each $B_u \ge 0$ and

$$\sum_{u=s}^{t-1} \frac{B_u}{\lambda_u} \le \frac{24}{500} < 1,$$

we also have

$$0 \le \frac{B_u}{\lambda_u} < 1$$

for every $u \in \{s, \dots, t-1\}$. We obtain

$$\prod_{u=s}^{t-1} \frac{1}{1-\frac{1}{\lambda_u}B_u} \le \frac{1}{1-\sum_{u=s}^{t-1}\frac{1}{\lambda_u}B_u} \le \frac{1}{1-24/500} = \frac{125}{119}.$$

□

Lemma 10. *For every $i \in [K]$ and $u \ge 1$,*

$$p_{u+1}(i) \le C_u^{\mathrm{reg}} \bar{p}_{u+1}(i),$$

where

$$C_u^{\mathrm{reg}} := \left(\frac{\lambda_{u+1}-\lambda_u}{\gamma_u} + \frac{\gamma_{u+1}}{\gamma_u}\right)^2.$$

Proof. Recall that

$$p_{u+1} = \arg\min_{p \in \Delta^K} \left\{ \langle p, L_{u+1}^{\text{obs}} \rangle + R_{u+1}(p) \right\},$$

whereas, by definition of the intermediate iterate,

$$\bar{p}_{u+1} = \arg\min_{p \in \Delta^K} \left\{ \langle p, L_{u+1}^{\text{obs}} \rangle + R_u(p) \right\}.$$

Thus, the two iterates use the same cumulative loss vector L_{u+1}^{obs} and differ only in their regularizers; $\bar{p}_{u+1}$ uses R_u, whereas p_{u+1} uses R_{u+1}.

Since p_0 is uniform, the term $-\lambda_u x \log p_{0,i}$ is linear in x with the same coefficient for every coordinate and therefore contributes only a constant on the simplex. Likewise, $\lambda_u \sum_i p(i) = \lambda_u$ is constant on Δ^K. Hence, without changing the minimizer, we may use the coordinate-wise regularizer

$$r_{u,i}(x) = \lambda_u x(\log x - 1) - 2\gamma_u \sqrt{x}.$$

Hence

$$r'_{u,i}(x) = \lambda_u \log x - \frac{\gamma_u}{\sqrt{x}}.$$

In particular, $r'_{u,i}$ is strictly increasing on $(0, 1]$.

By the KKT conditions, there exist multipliers $\mu, \bar{\mu} \in \mathbb{R}$ such that, for every $i \in [K]$,

$$r'_{u+1,i}(p_{u+1}(i)) = -L_{u+1}^{\text{obs}}(i) + \mu$$

and

$$r'_{u,i}(\bar{p}_{u+1}(i)) = -L_{u+1}^{\text{obs}}(i) + \bar{\mu}.$$

We first show that

$$\mu \le \bar{\mu}.$$

Since both $(\lambda_v)_v$ and $(\gamma_v)_v$ are non-decreasing and $\log x \le 0$ for every $x \in (0, 1]$, we have

$$r'_{u+1,i}(x) - r'_{u,i}(x) = (\lambda_{u+1} - \lambda_u) \log x - \frac{\gamma_{u+1} - \gamma_u}{\sqrt{x}} \le 0.$$

Therefore,

$$r'_{u+1,i}(x) \le r'_{u,i}(x), \qquad x \in (0, 1].$$

Suppose, for contradiction, that $\mu > \bar{\mu}$. Then, for every $i \in [K]$,

$$r'_{u+1,i}(p_{u+1}(i)) = -L_{u+1}^{\text{obs}}(i) + \mu > -L_{u+1}^{\text{obs}}(i) + \bar{\mu} = r'_{u,i}(\bar{p}_{u+1}(i)) \ge r'_{u+1,i}(\bar{p}_{u+1}(i)).$$

Since $r'_{u+1,i}$ is strictly increasing, this implies

$$p_{u+1}(i) > \bar{p}_{u+1}(i), \qquad \forall i \in [K],$$

which contradicts

$$\sum_{i=1}^{K} p_{u+1}(i) = \sum_{i=1}^{K} \bar{p}_{u+1}(i) = 1.$$

Hence

$$\mu \le \bar{\mu}.$$

Consequently,

$$r'_{u+1,i}(p_{u+1}(i)) \le r'_{u,i}(\bar{p}_{u+1}(i)).$$

Using the expression for the derivatives gives

$$\lambda_{u+1} \log p_{u+1}(i) - \frac{\gamma_{u+1}}{\sqrt{p_{u+1}(i)}} \le \lambda_u \log \bar{p}_{u+1}(i) - \frac{\gamma_u}{\sqrt{\bar{p}_{u+1}(i)}}.$$

If

$$p_{u+1}(i) \le \bar{p}_{u+1}(i),$$

the desired result follows immediately since $C_u^{\text{reg}} \ge 1$.

Hence suppose that

$$p_{u+1}(i) > \bar{p}_{u+1}(i)$$

and define

$$\alpha := \frac{p_{u+1}(i)}{\bar{p}_{u+1}(i)} > 1.$$

Substituting

$$p_{u+1}(i) = \alpha \bar{p}_{u+1}(i)$$

into the previous inequality and rearranging gives

$$\frac{1}{\sqrt{\alpha}} \geq \frac{\gamma_u}{\gamma_{u+1}} + 2\sqrt{\bar{p}_{u+1}(i)} \log\left(\sqrt{\bar{p}_{u+1}(i)}\right) \frac{\lambda_{u+1} - \lambda_u}{\gamma_{u+1}} + \log(\alpha) \frac{\lambda_{u+1}}{\gamma_{u+1}} \sqrt{\bar{p}_{u+1}(i)}.$$

We first consider the case

$$\lambda_{u+1} = \lambda_u.$$

Then

$$\frac{1}{\sqrt{\alpha}} \geq \frac{\gamma_u}{\gamma_{u+1}} + \log(\alpha) \frac{\lambda_{u+1}}{\gamma_{u+1}} \sqrt{\bar{p}_{u+1}(i)} \geq \frac{\gamma_u}{\gamma_{u+1}},$$

where the second inequality follows from $\alpha > 1$. Hence

$$\alpha \leq \left(\frac{\gamma_{u+1}}{\gamma_u}\right)^2 \leq C_u^{\mathrm{reg}}.$$

It remains to consider the case

$$\lambda_{u+1} > \lambda_u.$$

Set

$$z := \sqrt{\bar{p}_{u+1}(i)} \in (0, 1].$$

Then

$$\frac{1}{\sqrt{\alpha}} \geq \frac{\gamma_u}{\gamma_{u+1}} + \min_{0 < z \leq 1} \left\{ 2z \log(z) \frac{\lambda_{u+1} - \lambda_u}{\gamma_{u+1}} + z \log(\alpha) \frac{\lambda_{u+1}}{\gamma_{u+1}} \right\}.$$

The function inside the minimum is convex. For its unconstrained minimizer $z^\star$ we have

$$0 < z^\star = \frac{1}{e} \left(\frac{1}{\sqrt{\alpha}}\right)^{\frac{\lambda_{u+1}}{\lambda_{u+1} - \lambda_u}} < \frac{1}{e} < 1,$$

so $z^\star$ is also the minimizer over $(0, 1]$. Therefore,

$$\frac{1}{\sqrt{\alpha}} \geq \frac{\gamma_u}{\gamma_{u+1}} - \frac{2}{e} \frac{\lambda_{u+1} - \lambda_u}{\gamma_{u+1}} \left(\frac{1}{\sqrt{\alpha}}\right)^{\frac{\lambda_{u+1}}{\lambda_{u+1} - \lambda_u}}.$$

Since $\alpha \geq 1$ and

$$\frac{\lambda_{u+1}}{\lambda_{u+1} - \lambda_u} \geq 1,$$

we have

$$\left(\frac{1}{\sqrt{\alpha}}\right)^{\frac{\lambda_{u+1}}{\lambda_{u+1} - \lambda_u}} \leq \frac{1}{\sqrt{\alpha}}.$$

Moreover, $2/e \leq 1$. Thus

$$\frac{1}{\sqrt{\alpha}} \geq \frac{\gamma_u}{\gamma_{u+1}} - \frac{\lambda_{u+1} - \lambda_u}{\gamma_{u+1}} \frac{1}{\sqrt{\alpha}}.$$

Rearranging gives

$$\frac{1}{\sqrt{\alpha}} \left(1 + \frac{\lambda_{u+1} - \lambda_u}{\gamma_{u+1}}\right) \geq \frac{\gamma_u}{\gamma_{u+1}},$$

and therefore

$$\sqrt{\alpha} \leq \frac{\gamma_{u+1} + \lambda_{u+1} - \lambda_u}{\gamma_u}.$$

Squaring yields

$$\alpha \le \left(\frac{\lambda_{u+1} - \lambda_u}{\gamma_u} + \frac{\gamma_{u+1}}{\gamma_u} \right)^2 = C_u^{\mathrm{reg}}.$$

Thus, in either case,

$$p_{u+1}(i) \le C_u^{\mathrm{reg}} \bar{p}_{u+1}(i).$$

□

Lemma 11. *Let $s \le t$ with $t - s \le d_{\max}$. Then*

$$\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} \le \exp\left(2\left(\frac{\gamma_t}{\gamma_s} - 1 \right) + 2\frac{\lambda_t - \lambda_s}{\gamma_s} \right) < 1.06.$$

Proof. Define

$$a_u := \frac{\gamma_{u+1} - \gamma_u}{\gamma_u}, \qquad b_u := \frac{\lambda_{u+1} - \lambda_u}{\gamma_u}.$$

Then

$$C_u^{\mathrm{reg}} = (1 + a_u + b_u)^2.$$

Since $a_u, b_u \ge 0$ and $1 + x \le e^x$,

$$C_u^{\mathrm{reg}} \le \exp(2a_u + 2b_u).$$

Consequently,

$$\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} \le \exp\left(2\sum_{u=s}^{t-1} a_u + 2\sum_{u=s}^{t-1} b_u \right).$$

Since $\gamma_u \ge \gamma_s$,

$$\sum_{u=s}^{t-1} a_u \le \frac{1}{\gamma_s} \sum_{u=s}^{t-1} (\gamma_{u+1} - \gamma_u) = \frac{\gamma_t - \gamma_s}{\gamma_s} = \frac{\gamma_t}{\gamma_s} - 1.$$

Similarly,

$$\sum_{u=s}^{t-1} b_u \le \frac{1}{\gamma_s} \sum_{u=s}^{t-1} (\lambda_{u+1} - \lambda_u) = \frac{\lambda_t - \lambda_s}{\gamma_s}.$$

Therefore,

$$\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} \le \exp\left(2\left(\frac{\gamma_t}{\gamma_s} - 1 \right) + 2\frac{\lambda_t - \lambda_s}{\gamma_s} \right).$$

Let $X_q = (p_q - p_q^\star)^\top \widehat{\ell}_q$. Since the running maximum can increase only through newly incorporated feedback, we have

$$\lambda_t - \lambda_s \le \frac{1}{\log K} \sum_{q \in O_t \setminus O_s} (X_q)_+.$$

Moreover, since $\widehat{\ell}_q$ has only one nonzero coordinate and $\ell_q(I_q) \in [0, 1]$,

$$X_q = \frac{p_q(I_q) - p_q^\star(I_q)}{p_q(I_q)} \ell_q(I_q) \le 1,$$

and hence $(X_q)_+ \le 1$. Therefore,

$$\lambda_t - \lambda_s \le \frac{|O_t \setminus O_s|}{\log K}.$$

As in Lemma 9,

$$|O_t \setminus O_s| \le (t - s) + |P_s| \le 2d_{\max},$$

so

$$\lambda_t - \lambda_s \le \frac{2d_{\max}}{\log K}.$$

Moreover, by definition,

$$\gamma_s = \sqrt{\theta_2 + s}, \qquad \theta_2 = 10^4 d_{\max}^2.$$

Hence

$$\gamma_s \geq \sqrt{\theta_2} = 100 d_{\max}.$$

Thus,

$$\frac{\lambda_t - \lambda_s}{\gamma_s} \leq \frac{1}{50 \log K} \leq \frac{1}{50 \log 2}.$$

Furthermore,

$$\left(\frac{\gamma_t}{\gamma_s}\right)^2 \leq \frac{\theta_2 + t}{\theta_2 + s} = 1 + \frac{t - s}{\theta_2 + s} \leq 1 + \frac{d_{\max}}{10^4 d_{\max}^2} \leq 1.0001.$$

Hence

$$\prod_{u=s}^{t-1} C_u^{\mathrm{reg}} \leq \exp\left(2(\sqrt{1.0001} - 1) + \frac{1}{25 \log 2}\right) < 1.06.$$

□

Lemma 12. *For any $i \in [K]$, we have that*

$$p_{t,i}^\star \leq \frac{6}{5} p_t(i).$$

Proof. By the KKT conditions, there exist $\mu, \mu^\star \in \mathbb{R}$ such that

$$r'_{t,i}(p_t(i)) = -L_t^{\mathrm{obs}}(i) + \mu$$

and

$$r'_{t,i}(p_t^\star(i)) = -L_t^\star(i) + \mu^\star.$$

Hence

$$r'_{t,i}(p_t(i)) - r'_{t,i}(p_t^\star(i)) = U_t(i) + \mu - \mu^\star. \tag{12}$$

Since $r'_{t,i}$ is concave, we have

$$\big(p_t(i) - p_t^\star(i)\big) r''_{t,i}(p_t(i)) \leq r'_{t,i}(p_t(i)) - r'_{t,i}(p_t^\star(i)).$$

Combining this with equation (12) gives

$$p_t(i) - p_t^\star(i) \leq r''_{t,i}(p_t(i))^{-1} \big(U_t(i) + \mu - \mu^\star\big).$$

Summing over $i \in [K]$ and using

$$\sum_{i=1}^K p_t(i) = \sum_{i=1}^K p_t^\star(i) = 1,$$

we obtain

$$0 \leq \sum_{i=1}^K r''_{t,i}(p_t(i))^{-1} \big(U_t(i) + \mu - \mu^\star\big).$$

Rearranging yields

$$\mu^\star - \mu \leq \frac{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1} U_t(j)}{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1}}. \tag{13}$$

For every $s \in P_t$, we have $t - s \leq d_{\max}$, so Lemma 6 gives

$$p_t(j) \leq \frac{6}{5} p_s(j), \qquad j \in [K].$$

Applying Lemma 7 with $\alpha = 6/5$ therefore yields

$$\frac{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1} \widehat{\ell}_s(j)}{\sum_{j=1}^K r''_{t,j}(p_t(j))^{-1}} \leq \frac{12}{5}(K-1)^{1/3}.$$

Since $|P_t| \le d_{\max}$, (13) implies

$$\mu^\star - \mu \le \frac{12}{5} d_{\max}(K-1)^{1/3} =: B_t. \tag{14}$$

By concavity of $r'_{t,i}$,

$$r'_{t,i}(p_t(i)) - r'_{t,i}(p_t^\star(i)) \le \big(p_t(i) - p_t^\star(i)\big) r''_{t,i}(p_t^\star(i)).$$

Combining this with (12) gives

$$\big(p_t^\star(i) - p_t(i)\big) r''_{t,i}(p_t^\star(i)) \le \mu^\star - \mu - U_t(i) \le B_t,$$

where we used $U_t(i) \ge 0$. Since

$$r''_{t,i}(x) \ge \frac{\lambda_t}{x},$$

we obtain

$$p_t^\star(i) - p_t(i) \le \frac{1}{\lambda_t} B_t \, p_t^\star(i).$$

Therefore

$$p_t(i) \ge (1 - \frac{1}{\lambda_t} B_t) p_t^\star(i).$$

Finally,

$$\frac{1}{\lambda_t} B_t \le \frac{\frac{12}{5} d_{\max}(K-1)^{1/3}}{100 d_{\max}(K-1)^{1/3}} = \frac{12}{500} < \frac{1}{6}.$$

Hence

$$p_t(i) \ge \frac{5}{6} p_t^\star(i),$$

or equivalently,

$$p_t^\star(i) \le \frac{6}{5} p_t(i).$$

□

A.1.2 Regret decomposition and bounds

Lemma 3 (restated). *For every comparator $u \in \Delta^K$, Algorithm 1 guarantees that*

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^T (p_t - u)^\top \ell_t\Big] \le{}& \underbrace{\mathbb{E}\Big[\sum_{t=1}^T (p_t - p_t^\star)^\top \ell_t\Big]}_{\text{(I)}} + \underbrace{\mathbb{E}\Big[\sum_{t=1}^T \Big(\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\Big)\Big]}_{\text{(II)}} \\
&+ \underbrace{\mathbb{E}[\Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star)] + 2\gamma_{T+1} \sum_{i=1}^K \sqrt{u(i)}}_{\text{(III)}} \\
&+ \underbrace{\mathbb{E}\Big[\sum_{t=1}^T \Big(R_t(p_{t+1}^\star) - R_{t+1}(p_{t+1}^\star)\Big)\Big] - 2\gamma_{T+1} \sum_{i=1}^K \sqrt{u(i)}}_{\text{(IV)}}.
\end{aligned} \tag{4}$$

Proof. We decompose regret as

$$\mathbb{E}\left[\sum_{t=1}^T (p_t - u)^\top \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^T (p_t - p_t^\star)^\top \ell_t\right] + \mathbb{E}\left[\sum_{t=1}^T (p_t^\star - u)^\top \ell_t\right]. \tag{15}$$

Since $\mathbb{E}[\widehat{\ell}_t \mid \mathcal{F}_t] = \ell_t$,

$$\mathbb{E}\left[\sum_{t=1}^T (p_t^\star - u)^\top \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^T (p_t^\star - u)^\top \widehat{\ell}_t\right].$$

We add and subtract a telescoping sequence of potentials in order to isolate a one-step stability term.

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}(p_t^\star - u)^\top \ell_t] = \mathbb{E}[\sum_{t=1}^{T}(p_t^\star - u)^\top \widehat{\ell}_t] \\
&= \mathbb{E}[\sum_{t=1}^{T}(p_t^\star - u)^\top \widehat{\ell}_t + \sum_{t=1}^{T}\Big(\Phi_t(p_t^\star, L_t^\star) - \Phi_{t+1}(p_{t+1}^\star, L_{t+1}^\star)\Big) - \Phi_1(p_1^\star, L_1^\star) + \Phi_{T+1}(p_{T+1}^\star, L_{T+1}^\star)] \\
&= \mathbb{E}[\sum_{t=1}^{T}\Big({p_t^\star}^\top \widehat{\ell}_t + \Phi_t(p_t^\star, L_t^\star) - \Phi_{t+1}(p_{t+1}^\star, L_{t+1}^\star)\Big) + \Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star) + \\
&\quad \Phi_{T+1}(p_{T+1}^\star, L_{T+1}^\star) - \Phi_{T+1}(u, L_{T+1}^\star)] \\
&\le \mathbb{E}[\sum_{t=1}^{T}\Big({p_t^\star}^\top \widehat{\ell}_t + \Phi_t(p_t^\star, L_t^\star) - \Phi_{t+1}(p_{t+1}^\star, L_{t+1}^\star)\Big) + \Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star)] \\
&= \mathbb{E}[\sum_{t=1}^{T}\Big({p_t^\star}^\top \widehat{\ell}_t + \Phi_t(p_t^\star, L_t^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\Big) + \Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star) \\
&\quad + \sum_{t=1}^{T}\big(R_t(p_{t+1}^\star) - R_{t+1}(p_{t+1}^\star)\big)] \\
&= \mathbb{E}[\sum_{t=1}^{T}\Big(\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\Big) + \Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star) \\
&\quad + \sum_{t=1}^{T}\big(R_t(p_{t+1}^\star) - R_{t+1}(p_{t+1}^\star)\big)] + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)} - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)} \qquad (16)
\end{aligned}
$$

Note that the initial condition is unchanged: since $L_1^{\text{obs}} = L_1^\star = 0$ and p_0 is uniform, we have $p_1 = p_1^\star$.

Combining (15) and (16) yields

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(p_t - u)^\top \ell_t\right] \le\ & \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(p_t - p_t^\star)^\top \ell_t\right]}_{\text{(I)}} \\
& + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\Big(\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\Big)\right]}_{\text{(II)}} \\
& + \underbrace{\mathbb{E}[\Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star)] + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}}_{\text{(III)}} \\
& + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\Big(R_t(p_{t+1}^\star) - R_{t+1}(p_{t+1}^\star)\Big)\right] - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}}_{\text{(IV)}}.
\end{aligned}
$$

□

Using the Tsallis curvature lower bound (Zimmert & Seldin, 2021),

$$\nabla^2\Phi_t(p)_{ii} \ge \frac{1}{4}\frac{\sqrt{t}}{p(i)^{3/2}},$$

we can use the following Lemma.

Lemma 13. *We find the following bound on term (II).*

$$(II) = \mathbb{E}\left[\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\right] \le 25\mathbb{E}\left[\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}}\right].$$

Proof. Define the auxiliary one-step iterate

$$\widetilde{p}_{t+1}^\star \in \arg\min_{p\in\Delta^K} \Phi_t(p, L_{t+1}^\star) = \arg\min_{p\in\Delta^K}\left\{p^\top(L_t^\star + \widehat{\ell}_t) + R_t(p)\right\}.$$

Notice that $\widetilde{p}_{t+1}^\star$ uses the same regularizer R_t as $p_t^\star$. Since $\widetilde{p}_{t+1}^\star$ minimizes $\Phi_t(\cdot, L_{t+1}^\star)$,

$$\Phi_t(\widetilde{p}_{t+1}^\star, L_{t+1}^\star) \le \Phi_t(p_{t+1}^\star, L_{t+1}^\star).$$

Consequently,

$$\begin{aligned}&\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\\ &\quad\le \Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(\widetilde{p}_{t+1}^\star, L_{t+1}^\star).\end{aligned} \tag{17}$$

Define

$$F_t(p) := \Phi_t(p, L_t^\star) = p^\top L_t^\star + R_t(p).$$

Then $p_t^\star$ minimizes F_t over Δ^K, whereas $\widetilde{p}_{t+1}^\star$ minimizes $F_t(p) + p^\top\widehat{\ell}_t$.

Let

$$\delta_t := p_t^\star - \widetilde{p}_{t+1}^\star.$$

Since both vectors belong to the simplex,

$$\delta_t^\top \mathbf{1} = 0. \tag{18}$$

We first establish the coordinate-wise comparison needed to control the curvature along the segment joining $p_t^\star$ and $\widetilde{p}_{t+1}^\star$. Write

$$r_{t,i}(x) = \lambda_t x(\log x - 1) - 2\gamma_t\sqrt{x},$$

so that

$$r'_{t,i}(x) = \lambda_t \log x - \frac{\gamma_t}{\sqrt{x}}, \qquad r''_{t,i}(x) = \frac{\lambda_t}{x} + \frac{\gamma_t}{2x^{3/2}}.$$

By the KKT conditions, there exist $\mu, \widetilde{\mu} \in \mathbb{R}$ such that

$$r'_{t,i}(p_t^\star(i)) = -L_t^\star(i) + \mu$$

and

$$r'_{t,i}(\widetilde{p}_{t+1}^\star(i)) = -L_t^\star(i) - \widehat{\ell}_t(i) + \widetilde{\mu}.$$

Hence

$$r'_{t,i}(p_t^\star(i)) - r'_{t,i}(\widetilde{p}_{t+1}^\star(i)) = \widehat{\ell}_t(i) + \mu - \widetilde{\mu}. \tag{19}$$

Since $r'_{t,i}$ is concave,

$$\widetilde{p}_{t+1}^\star(i) - p_t^\star(i) \ge r''_{t,i}(p_t^\star(i))^{-1}\left(\widetilde{\mu} - \mu - \widehat{\ell}_t(i)\right).$$

Summing over i and using that both iterates belong to the simplex gives

$$\widetilde{\mu} - \mu \le B_t^\star,$$

where

$$B_t^\star := \frac{\displaystyle\sum_{j=1}^K r''_{t,j}(p_t^\star(j))^{-1}\widehat{\ell}_t(j)}{\displaystyle\sum_{j=1}^K r''_{t,j}(p_t^\star(j))^{-1}}.$$

Then by a simple variation of Lemma 7 we have that

$$B_t^\star \le \frac{12}{5}(K-1)^{1/3}\,.$$

By definition of λ_t,

$$\frac{B_t^\star}{\lambda_t} \le \frac{12}{500}\,.$$

Hence

$$1 - \frac{B_t^\star}{\lambda_t} > 0\,.$$

If $\widetilde{p}_{t+1}^\star(i) \le p_t^\star(i)$, the desired comparison is immediate. Otherwise, concavity of $r'_{t,i}$ and (19) give

$$r''_{t,i}(\widetilde{p}_{t+1}^\star(i))\left(\widetilde{p}_{t+1}^\star(i) - p_t^\star(i)\right) \le \widetilde{\mu} - \mu - \widehat{\ell}_t(i) \le B_t^\star.$$

Since

$$r''_{t,i}(x) \ge \frac{\lambda_t}{x},$$

we obtain

$$\widetilde{p}_{t+1}^\star(i) \le \frac{p_t^\star(i)}{1 - B_t^\star/\lambda_t} \le \frac{125}{122}p_t^\star(i) < \frac{25}{23}p_t^\star(i).$$

Therefore,

$$\widetilde{p}_{t+1}^\star(i) \le \frac{25}{23}p_t^\star(i).$$

Moreover, Lemma 12 gives

$$p_t^\star(i) \le \frac{6}{5}p_t(i).$$

Hence

$$\widetilde{p}_{t+1}^\star(i) \le \frac{30}{23}p_t(i).$$

Therefore, every point z on the segment joining $p_t^\star$ and $\widetilde{p}_{t+1}^\star$ satisfies

$$z(i) \le \frac{30}{23}p_t(i), \qquad i \in [K]. \tag{20}$$

By the Tsallis curvature lower bound,

$$\nabla^2 F_t(z)_{ii} = \nabla^2 R_t(z)_{ii} \ge \frac{\sqrt{t}}{4z(i)^{3/2}}.$$

Using (20),

$$\nabla^2 F_t(z)_{ii} \ge \frac{1}{4(30/23)^{3/2}}\frac{\sqrt{t}}{p_t(i)^{3/2}}.$$

Define

$$H_t := \operatorname{diag}\left(\frac{\sqrt{t}}{p_t(1)^{3/2}}, \dots, \frac{\sqrt{t}}{p_t(K)^{3/2}}\right)$$

and

$$c_0 := \frac{1}{4(30/23)^{3/2}}.$$

Thus, along the entire segment,

$$\nabla^2 F_t(z) \succeq c_0 H_t. \tag{21}$$

Since $p_t^\star$ minimizes F_t over the simplex, its first-order optimality condition implies that for some scalar α_t,

$$\nabla F_t(p_t^\star) = \alpha_t \mathbf{1}.$$

Together with (18), this gives

$$\nabla F_t(p_t^\star)^\top \left(\widetilde{p}_{t+1}^\star - p_t^\star\right) = 0.$$

Using (21) along the segment therefore yields

$$F_t(\widetilde{p}_{t+1}^\star) - F_t(p_t^\star) \geq \frac{c_0}{2}\|\delta_t\|_{H_t}^2. \tag{22}$$

Since $L_{t+1}^\star = L_t^\star + \widehat{\ell}_t$, we obtain

$$\begin{aligned}
&\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(\widetilde{p}_{t+1}^\star, L_{t+1}^\star) \\
&= F_t(p_t^\star) - F_t(\widetilde{p}_{t+1}^\star) + \widehat{\ell}_t^\top \left(p_t^\star - \widetilde{p}_{t+1}^\star\right) \\
&\leq \widehat{\ell}_t^\top \delta_t - \frac{c_0}{2}\|\delta_t\|_{H_t}^2.
\end{aligned} \tag{23}$$

Let $b \in \mathbb{R}$ be arbitrary. Since $\delta_t^\top \mathbf{1} = 0$,

$$\widehat{\ell}_t^\top \delta_t = (\widehat{\ell}_t - b\mathbf{1})^\top \delta_t.$$

Young's inequality in the H_t-norm gives

$$(\widehat{\ell}_t - b\mathbf{1})^\top \delta_t - \frac{c_0}{2}\|\delta_t\|_{H_t}^2 \leq \frac{1}{2c_0}\|\widehat{\ell}_t - b\mathbf{1}\|_{H_t^{-1}}^2.$$

Since

$$\frac{1}{2c_0} = 2\left(\frac{30}{23}\right)^{3/2} < 3,$$

we conclude that, for every $b \in \mathbb{R}$,

$$\begin{aligned}
&\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star) \\
&\qquad \leq \frac{3}{\sqrt{t}} \sum_{i=1}^K p_t(i)^{3/2} \big(\widehat{\ell}_t(i) - b\big)^2.
\end{aligned} \tag{24}$$

Choose

$$b = b_t := \ell_t(i^\star).$$

Condition on $\mathcal{F}_t$ and ℓ_t, and take expectation over $I_t \sim p_t$.

For every $i \neq i^\star$,

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\ell}_t(i) - b_t)^2 \mid \mathcal{F}_t, \ell_t\right] &= p_t(i)\left(\frac{\ell_t(i)}{p_t(i)} - b_t\right)^2 + (1 - p_t(i))b_t^2 \\
&= \frac{\ell_t(i)^2}{p_t(i)} - 2\ell_t(i)b_t + b_t^2 \\
&\leq \frac{1}{p_t(i)} + 1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&p_t(i)^{3/2}\mathbb{E}\left[(\widehat{\ell}_t(i) - b_t)^2 \mid \mathcal{F}_t, \ell_t\right] \\
&\qquad \leq \sqrt{p_t(i)} + p_t(i)^{3/2} \leq 2\sqrt{p_t(i)}.
\end{aligned} \tag{25}$$

For the optimal arm,

$$\mathbb{E}\left[(\widehat{\ell}_t(i^\star) - b_t)^2 \mid \mathcal{F}_t, \ell_t\right] = \ell_t(i^\star)^2\left(\frac{1}{p_t(i^\star)} - 1\right).$$

Hence

$$\begin{aligned}
&p_t(i^\star)^{3/2}\mathbb{E}\left[(\widehat{\ell}_t(i^\star) - b_t)^2 \mid \mathcal{F}_t, \ell_t\right] \\
&= \ell_t(i^\star)^2\sqrt{p_t(i^\star)}(1 - p_t(i^\star)) \\
&\leq 1 - p_t(i^\star) = \sum_{i \neq i^\star} p_t(i) \leq \sum_{i \neq i^\star} \sqrt{p_t(i)}.
\end{aligned} \tag{26}$$

Combining (25) and (26),

$$\mathbb{E}\left[\sum_{i=1}^{K} p_t(i)^{3/2}(\widehat{\ell}_t(i)-b_t)^2 \mid \mathcal{F}_t, \ell_t\right] \le 3\sum_{i\neq i^\star}\sqrt{p_t(i)}.$$

Taking expectations and substituting into (24) gives

$$\mathbb{E}\left[\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\right] \le 9\mathbb{E}\left[\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}}\right].$$

In particular,

$$\mathbb{E}\left[\Phi_t(p_t^\star, L_{t+1}^\star) - \Phi_t(p_{t+1}^\star, L_{t+1}^\star)\right] \le 25\mathbb{E}\left[\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}}\right].$$

Summing over $t = 1, \dots, T$ yields

$$(II) \le 25\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}}\right].$$

□

Lemma 14. *The bound on term III is given by:*

$$(III) \le \mathbb{E}[\lambda_{T+1}]\log K + 2\sqrt{\theta_2+1}\sqrt{K}.$$

Consequently,

$$(III) \le \mathbb{E}[\lambda_{T+1}]\log K + O(d_{\max}\sqrt{K}).$$

Proof. Recall that: Let

$$R_t(p) = \lambda_t \mathrm{KL}(p, p_0) - 2\gamma_t\sum_{i=1}^{K}\sqrt{p(i)},$$

where

$$\gamma_t = \sqrt{\theta_2 + t}, \qquad \theta_2 = 10^4 d_{\max}^2.$$

Since $L_1^\star = 0$, we have

$$\begin{aligned}(III) &= \Phi_{T+1}(u, L_1^\star) - \Phi_1(p_1^\star, L_1^\star) + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)} \\ &= R_{T+1}(u) - R_1(p_1^\star) + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)}.\end{aligned}$$

Expanding the regularizers gives

$$\begin{aligned}(III) &= \lambda_{T+1}\mathrm{KL}(u, p_0) - \lambda_1\mathrm{KL}(p_1^\star, p_0) \\ &\quad - 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)} + 2\gamma_1\sum_{i=1}^{K}\sqrt{p_1^\star(i)} + 2\gamma_{T+1}\sum_{i=1}^{K}\sqrt{u(i)} \\ &\le \lambda_{T+1}\mathrm{KL}(u, p_0) - \lambda_1\mathrm{KL}(p_1^\star, p_0) + 2\gamma_1\sum_{i=1}^{K}\sqrt{p_1^\star(i)}.\end{aligned}$$

Since $L_1^\star = 0$ and p_0 is uniform, we have

$$p_1^\star = p_0,$$

and hence

$$\mathrm{KL}(p_1^\star, p_0) = 0.$$

Therefore,

$$(III) \leq \lambda_{T+1}\mathrm{KL}(u, p_0) + 2\gamma_1 \sum_{i=1}^{K} \sqrt{p_0(i)}.$$

Since $p_0(i) = 1/K$,

$$\sum_{i=1}^{K} \sqrt{p_0(i)} = \sqrt{K}.$$

Furthermore,

$$\mathrm{KL}(u, p_0) \leq \log K.$$

Thus

$$(III) \leq \lambda_{T+1} \log K + 2\gamma_1 \sqrt{K}.$$

Using $\gamma_1 = \sqrt{\theta_2 + 1}$ gives

$$\text{(III)} \leq \lambda_{T+1} \log K + 2\sqrt{\theta_2 + 1}\sqrt{K}.$$

Finally, since $\theta_2 = 10^4 d_{\max}^2$,

$$\sqrt{\theta_2 + 1} = O(d_{\max} + 1),$$

and therefore

$$\text{(III)} \leq \lambda_{T+1} \log K + O(d_{\max}\sqrt{K}).$$

□

Lemma 15. *The bound on term (IV) is given by:*

$$\text{(IV)} \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i=1}^{K} \sqrt{p^\star_{t+1}(i)}\right] - 2\gamma_{T+1} \sum_{i=1}^{K} \sqrt{u(i)}.$$

Proof. Recall that

$$\text{(IV)} = \mathbb{E}\left[\sum_{t=1}^{T} \left(R_t(p^\star_{t+1}) - R_{t+1}(p^\star_{t+1})\right)\right] - 2\gamma_{T+1} \sum_{i=1}^{K} \sqrt{u(i)}.$$

Note that λ_t is non-decreasing and recall that

$$\gamma_t = \sqrt{\theta_2 + t}\,.$$

For every $p \in \Delta^K$,

$$R_t(p) - R_{t+1}(p) = (\lambda_t - \lambda_{t+1})\mathrm{KL}(p, p_0) + 2(\gamma_{t+1} - \gamma_t) \sum_{i=1}^{K} \sqrt{p(i)}.$$

Since λ_t is non-decreasing,

$$\lambda_t - \lambda_{t+1} \leq 0,$$

and since $\mathrm{KL}(p, p_0) \geq 0$,

$$(\lambda_t - \lambda_{t+1})\mathrm{KL}(p, p_0) \leq 0.$$

Therefore,

$$R_t(p) - R_{t+1}(p) \leq 2(\gamma_{t+1} - \gamma_t) \sum_{i=1}^{K} \sqrt{p(i)}.$$

Applying this with $p = p^\star_{t+1}$ and summing over t gives

$$\text{(IV)} \leq 2\sum_{t=1}^{T} (\gamma_{t+1} - \gamma_t) \sum_{i=1}^{K} \sqrt{p^\star_{t+1}(i)} - 2\gamma_{T+1} \sum_{i=1}^{K} \sqrt{u(i)}.$$

Next,

$$\begin{aligned}\gamma_{t+1} - \gamma_t &= \sqrt{\theta_2 + t + 1} - \sqrt{\theta_2 + t} \\ &= \frac{1}{\sqrt{\theta_2 + t + 1} + \sqrt{\theta_2 + t}}.\end{aligned}$$

Hence

$$\gamma_{t+1} - \gamma_t \leq \frac{1}{2\sqrt{\theta_2 + t}} \leq \frac{1}{2\sqrt{t}}.$$

It follows that

$$\text{(IV)} \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i=1}^{K} \sqrt{p^\star_{t+1}(i)}\right] - 2\gamma_{T+1} \sum_{i=1}^{K} \sqrt{u(i)}.$$

□

A.2 Stochastic setting

Lemma 4 (restated)**.** *In the stochastic setting, we have that*

$$\mathbb{E}[\lambda_{T+d_{\max}+1} \log(K)] \leq \log(K)\theta_1 + \frac{24}{\log 2} \sum_{i \neq i^\star} \frac{\log(T)}{\Delta(i)} + \sum_{t=1}^{T} \frac{1}{4} \mathbb{E}[\sum_{i \neq i^\star} p_t(i)\Delta(i)] + 4d_{\max}$$

Proof. By definition of λ_t,

$$\lambda_{T+d_{\max}+1} = \theta_1 + \frac{1}{\log K} \max_{u \leq T+d_{\max}+1} \left(\sum_{s \in O_u} X_s \mathbb{1}[\max_{s' \leq s} s' + d_{s'} \leq u - 1] \right),$$

where, we recall it,

$$X_s := (p_s - p^\star_s)^\top \widehat{\ell}_s \,.$$

We rewrite

$$X_s = \mathbb{E}[X_s \mid \mathcal{F}_s] + M_s, \qquad M_s := X_s - \mathbb{E}[X_s \mid \mathcal{F}_s] \,.$$

Recall that p_s and $p^\star_s$ are $\mathcal{F}_s$-measurable, while $I_s \mid \mathcal{F}_s \sim p_s$. Moreover, in the stochastic setting, ℓ_s is independent of $\mathcal{F}_s$. Hence, by unbiasedness of the importance-weighted estimator,

$$\begin{aligned}\mathbb{E}[X_s \mid \mathcal{F}_s] &= (p_s - p^\star_s)^\top \mu \\ &= \sum_{i \neq i^\star} \big(p_s(i) - p^\star_s(i)\big)\Delta(i),\end{aligned}$$

where the second equality follows from $(p_s - p^\star_s)^\top \mathbf{1} = 0$.

Now, by Lemma 17, we have that

$$\sum_{i \neq i^\star} \big(p_s(i) - p^\star_s(i)\big)\Delta(i) \leq \sum_{i \neq i^\star} \lambda_s^{-1} p_s(i)\Delta(i) \sum_{s' \in P_s} \widehat{\ell}_{s'}(i) \,.$$

Hence, we have that

$$\begin{aligned}&\mathbb{E}[\max_{u \leq T+d_{\max}+1} \left(\sum_{s \in O_u} X_s \mathbb{1}[\max_{s' \leq s} s' + d_{s'} \leq u - 1] \right)] \\ &\leq \mathbb{E}[\max_{u \leq T+d_{\max}+1} \left(\sum_{s \in O_u} M_s \mathbb{1}[\max_{s' \leq s} s' + d_{s'} \leq u - 1] \right)] + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^\star} \lambda_s^{-1} p_s(i)\Delta(i) \sum_{s' \in P_s} \widehat{\ell}_{s'}(i) \right].\end{aligned}$$

By unbiasedness of the importance weighted estimator we have that

$$\begin{aligned}\mathbb{E}\left[\sum_{i\neq i^\star}\lambda_s^{-1}p_s(i)\Delta(i)\sum_{s'\in P_s}\widehat{\ell}_{s'}(i)\right] &= \mathbb{E}[\sum_{i\neq i^\star}\lambda_s^{-1}p_s(i)\Delta(i)\sum_{s'\in P_s}\ell_{s'}(i)]\\ &\leq \mathbb{E}[\sum_{i\neq i^\star}\lambda_s^{-1}p_s(i)\Delta(i)d_{\max}] \qquad \text{(since } \ell_t(i)\in[0,1]\text{)}\\ &\leq \mathbb{E}[\frac{1}{6}\sum_{i\neq i^\star}p_s(i)\Delta(i)]\,. \qquad \text{(definition of } \lambda_t\text{)}\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathbb{E}[\max_{u\leq T+d_{\max}+1}\left(\sum_{s\in O_u}X_s\mathbb{1}[\max_{s'\leq s}s'+d_{s'}\leq u-1]\right)]\\ &\leq \mathbb{E}[\max_{u\leq T+d_{\max}+1}\left(\sum_{s\in O_u}M_s\mathbb{1}[\max_{s'\leq s}s'+d_{s'}\leq u-1]\right)]+\mathbb{E}[\frac{1}{6}\sum_{i\neq i^\star}p_s(i)\Delta(i)]\,.\end{aligned}$$

By Lemma 12, $p_s^\star(i)\leq\frac{6}{5}p_s(i)$, and consequently $|M_s|\leq 2$. Recall that $O_u\cup P_u=[u-1]$. Hence,

$$\max_{u\leq T+d_{\max}+1}\left(\sum_{s\in O_u}M_s\mathbb{1}[\max_{s'\leq s}s'+d_{s'}\leq u-1]\right)\leq\max_{u\leq T}\left(\sum_{s\leq u}M_s\right)+4d_{\max}\,.$$

To apply Doob's maximal inequality (Williams, 1991, Section 14.11, p. 143), we make the filtration indexing explicit. By construction,

$$\mathbb{E}[M_s\mid\mathcal{F}_s]=0.$$

Moreover, X_s, and hence M_s, is $\mathcal{F}_{s+1}$-measurable. Therefore, defining

$$Z_n:=\sum_{s=1}^{n}M_s,$$

the process $(Z_n)_{n\geq 0}$ is a martingale with respect to the shifted filtration $(\mathcal{F}_{n+1})_{n\geq 0}$. Indeed,

$$\mathbb{E}[Z_n\mid\mathcal{F}_n]=Z_{n-1}+\mathbb{E}[M_n\mid\mathcal{F}_n]=Z_{n-1}.$$

Thus, no additional filtration containing different information is required; the shift only accounts for the fact that the randomness generated in round n is contained in $\mathcal{F}_{n+1}$.

We first have

$$\mathbb{E}\left[\max_{n\leq T}Z_n\right]\leq\mathbb{E}\left[\max_{n\leq T}|Z_n|\right].$$

By the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\max_{n\leq T}|Z_n|\right]\leq\sqrt{\mathbb{E}\left[\max_{n\leq T}|Z_n|^2\right]}.$$

Since $(Z_n)_{n\geq 0}$ is a square-integrable martingale with respect to $(\mathcal{F}_{n+1})_{n\geq 0}$, Doob's L^2 maximal inequality (Williams, 1991, Section 14.11, p. 143) gives

$$\mathbb{E}\left[\max_{n\leq T}|Z_n|^2\right]\leq 4\,\mathbb{E}[|Z_T|^2].$$

Consequently,

$$\mathbb{E}\left[\max_{n\leq T}\sum_{s=1}^{n}M_s\right]\leq 2\sqrt{\mathbb{E}\left[\left(\sum_{s=1}^{T}M_s\right)^2\right]}.$$

Furthermore, the martingale-difference property implies that the cross terms vanish. Hence,

$$\mathbb{E}\left[\left(\sum_{s=1}^{T} M_s\right)^2\right] = \sum_{s=1}^{T} \mathbb{E}[M_s^2].$$

By Lemma 16,

$$\begin{aligned}\sum_{s=1}^{T} \mathbb{E}[M_s^2] &\le 2 \sum_{s=1}^{T} \sum_{i \neq i^\star} \mathbb{E}[p_s(i)] \\ &= 2 \sum_{i \neq i^\star} \frac{1}{\Delta(i)} \mathbb{E}\left[\sum_{s=1}^{T} p_s(i)\Delta(i)\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\max_{n \le T} \sum_{s=1}^{n} M_s\right] &\le 2\sqrt{2} \sqrt{\sum_{i \neq i^\star} \frac{1}{\Delta(i)} \mathbb{E}\left[\sum_{s=1}^{T} p_s(i)\Delta(i)\right]} \\ &\le 2\sqrt{2} \sum_{i \neq i^\star} \sqrt{\frac{1}{\Delta(i)} \mathbb{E}\left[\sum_{s=1}^{T} p_s(i)\Delta(i)\right]}.\end{aligned}$$

Using Young's inequality

$$2\sqrt{2}\sqrt{ab} \le \frac{1}{12} a + 24b,$$

we obtain

$$\mathbb{E}\left[\max_{n \le T} \sum_{s=1}^{n} M_s\right] \le \frac{1}{12} \mathbb{E}\left[\sum_{s=1}^{T} \sum_{i \neq i^\star} p_s(i)\Delta(i)\right] + 24 \sum_{i \neq i^\star} \frac{1}{\Delta(i)}.$$

Combined with the preceding inequalities, this implies that

$$\mathbb{E}[\log(K)\lambda_{T+d_{\max}+1}] \le \log(K)\theta_1 + \frac{24}{\log 2} \sum_{i \neq i^\star} \frac{\log(T)}{\Delta(i)} + \frac{3}{12} \mathbb{E}[\sum_{i \neq i^\star} p_s(i)\Delta(i)] + 4d_{\max}\,.$$

□

Lemma 16. *We have that*

$$\mathbb{E}[M_t^2 \mid \mathcal{F}_t] \le 2 \sum_{i \neq i^\star} p_t(i),$$

Proof. Recall that

$$X_t = (p_t - p_t^\star)^\top \widehat{\ell}_t, \qquad M_t = X_t - \mathbb{E}[X_t \mid \mathcal{F}_t].$$

Since the conditional variance is bounded by the conditional second moment,

$$\mathbb{E}[M_t^2 \mid \mathcal{F}_t] \le \mathbb{E}[X_t^2 \mid \mathcal{F}_t].$$

This lemma uses the same filtration $(\mathcal{F}_t)_{t \ge 1}$ defined earlier. Recall that p_t and $p_t^\star$ are $\mathcal{F}_t$-measurable, while $I_t \mid \mathcal{F}_t \sim p_t$. Moreover, in the stochastic setting, ℓ_t is independent of $\mathcal{F}_t$.

By definition of the importance-weighted estimator,

$$X_t = \sum_{i=1}^{K} \big(p_t(i) - p_t^\star(i)\big) \frac{\ell_t(i)\mathbf{1}\{I_t = i\}}{p_t(i)}.$$

Conditioning on $\mathcal{F}_t$, and using that $\mathbb{P}(I_t = i \mid \mathcal{F}_t) = p_t(i)$ and that I_t is conditionally independent of ℓ_t, we obtain

$$\begin{aligned}\mathbb{E}[X_t^2 \mid \mathcal{F}_t] &= \sum_{i=1}^{K} \frac{(p_t(i) - p_t^\star(i))^2}{p_t(i)^2} \mathbb{E}\left[\ell_t(i)^2 \mathbf{1}\{I_t = i\} \mid \mathcal{F}_t\right] \\ &= \sum_{i=1}^{K} \frac{(p_t(i) - p_t^\star(i))^2}{p_t(i)} \mathbb{E}[\ell_t(i)^2 \mid \mathcal{F}_t].\end{aligned}$$

Since $\ell_t(i) \in [0, 1]$,

$$\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le \sum_{i=1}^{K} \frac{\big(p_t(i) - p_t^\star(i)\big)^2}{p_t(i)}.$$

We first record a consequence of the one-sided coordinate comparison from Lemma 12. Since

$$p_t^\star(i) \le \frac{6}{5} p_t(i),$$

we have

$$|p_t(i) - p_t^\star(i)| \le p_t(i).$$

Indeed, if $p_t(i) \ge p_t^\star(i)$, then

$$0 \le p_t(i) - p_t^\star(i) \le p_t(i),$$

while if $p_t^\star(i) > p_t(i)$, Lemma 12 gives

$$0 < p_t^\star(i) - p_t(i) \le \frac{1}{5} p_t(i) \le p_t(i).$$

For the suboptimal arms it therefore follows that

$$\sum_{i \ne i^\star} \frac{\big(p_t(i) - p_t^\star(i)\big)^2}{p_t(i)} \le \sum_{i \ne i^\star} p_t(i).$$

It remains to control the optimal-arm contribution. Since both p_t and $p_t^\star$ are probability vectors,

$$p_t(i^\star) - p_t^\star(i^\star) = -\sum_{i \ne i^\star} \big(p_t(i) - p_t^\star(i)\big).$$

Hence,

$$|p_t(i^\star) - p_t^\star(i^\star)| \le \sum_{i \ne i^\star} |p_t(i) - p_t^\star(i)| \le \sum_{i \ne i^\star} p_t(i).$$

We now distinguish two cases. If $\sum_{i \ne i^\star} p_t(i) \le 1/2$, then

$$p_t(i^\star) = 1 - \sum_{i \ne i^\star} p_t(i) \ge \frac{1}{2},$$

and therefore

$$\begin{aligned}\frac{\big(p_t(i^\star) - p_t^\star(i^\star)\big)^2}{p_t(i^\star)} &\le \frac{\Big(\sum_{i \ne i^\star} p_t(i)\Big)^2}{1 - \sum_{i \ne i^\star} p_t(i)} \\ &\le \sum_{i \ne i^\star} p_t(i).\end{aligned}$$

Combining this with the contribution of the suboptimal arms gives

$$\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le 2 \sum_{i \ne i^\star} p_t(i).$$

If $\sum_{i \neq i^\star} p_t(i) > 1/2$, the coordinate-wise bound

$$|p_t(i) - p_t^\star(i)| \leq p_t(i)$$

for every $i \in [K]$ gives

$$\begin{aligned}\mathbb{E}[X_t^2 \mid \mathcal{F}_t] &\leq \sum_{i=1}^{K} \frac{p_t(i)^2}{p_t(i)} \\ &= \sum_{i=1}^{K} p_t(i) = 1 < 2 \sum_{i \neq i^\star} p_t(i).\end{aligned}$$

Thus, in both cases,

$$\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \leq 2 \sum_{i \neq i^\star} p_t(i).$$

Consequently,

$$\mathbb{E}[M_t^2 \mid \mathcal{F}_t] \leq 2 \sum_{i \neq i^\star} p_t(i).$$

This proves the claim. □

A.2.1 Detailed proof of equation (2)

Theorem 1 (RESTATED)**.** *Algorithm 1 guarantees that*

$$\mathcal{R}_T \leq \begin{cases} O\left(\sum_{i \neq i^\star} \frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right) & \textit{stochastic setting} \quad (2) \\ O\left(\sqrt{KT} + \sqrt{D \log K} + d_{\max}\sqrt{K}\right) & \textit{adversarial setting} \quad (3) \end{cases}$$

Proof. Define the gap process

$$G_T := \sum_{t=1}^{T} \sum_{i \neq i^\star} p_t(i)\Delta(i).$$

Let $u = e_{i^\star}$. Since the losses are i.i.d. and $i^\star$ is the unique optimal arm, by linearity of expectation,

$$\mathbb{E}[R_T] = \mathbb{E}[G_T].$$

Taking $j = i^\star$ in equation (5) gives

$$\mathbb{E}[G_T] \leq 2\mathbb{E}\left[\lambda_{T+d_{\max}+1} \log K\right] + \frac{131}{5} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \neq i^\star} \sqrt{\frac{p_t(i)}{t}}\right] + 2\sqrt{\theta_2 + 1}\sqrt{K}. \quad (27)$$

By Lemma 4,

$$\mathbb{E}\left[\lambda_{T+d_{\max}+1} \log K\right] \leq \theta_1 \log K + \frac{24}{\log 2} \sum_{i \neq i^\star} \frac{\log T}{\Delta(i)} + \frac{1}{4}\mathbb{E}[G_T] + 4d_{\max}. \quad (28)$$

Hence,

$$2\mathbb{E}\left[\lambda_{T+d_{\max}+1} \log K\right] \leq 2\theta_1 \log K + \frac{48}{\log 2} \sum_{i \neq i^\star} \frac{\log T}{\Delta(i)} + \frac{1}{2}\mathbb{E}[G_T] + 8d_{\max}. \quad (29)$$

It remains to control the Tsallis term. For every $a > 0$, Young's inequality gives

$$\sqrt{\frac{p_t(i)}{t}} \leq a p_t(i)\Delta(i) + \frac{1}{4a\, t\Delta(i)}.$$

Choose

$$a = \frac{5}{1572} = \frac{1}{12(131/5)}.$$

Then

$$\frac{131}{5}a = \frac{1}{12},$$

and

$$\frac{131}{5}\frac{1}{4a} = \frac{51483}{25}.$$

Therefore,

$$\begin{aligned}\frac{131}{5}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\neq i^\star}\sqrt{\frac{p_t(i)}{t}}\right] &\leq \frac{1}{12}\mathbb{E}[G_T] + \frac{51483}{25}\sum_{i\neq i^\star}\frac{1}{\Delta(i)}\sum_{t=1}^{T}\frac{1}{t} \\ &\leq \frac{1}{12}\mathbb{E}[G_T] + \frac{51483}{25}\sum_{i\neq i^\star}\frac{1+\log T}{\Delta(i)}. \end{aligned} \tag{30}$$

Substituting (29) and (30) into (27), we obtain

$$\begin{aligned}\mathbb{E}[G_T] \leq \left(\frac{1}{2}+\frac{1}{12}\right)\mathbb{E}[G_T] + 2\theta_1\log K + 8d_{\max} + 2\sqrt{\theta_2+1}\sqrt{K} \\ + \frac{48}{\log 2}\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)} + \frac{51483}{25}\sum_{i\neq i^\star}\frac{1+\log T}{\Delta(i)}.\end{aligned}$$

Since $\frac{1}{2}+\frac{1}{12}=\frac{7}{12}$, we can move this term to the left-hand side to obtain

$$\begin{aligned}\frac{5}{12}\mathbb{E}[G_T] \leq 2\theta_1\log K + 8d_{\max} + 2\sqrt{\theta_2+1}\sqrt{K} \\ + \frac{48}{\log 2}\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)} + \frac{51483}{25}\sum_{i\neq i^\star}\frac{1+\log T}{\Delta(i)}.\end{aligned}$$

Consequently,

$$\mathbb{E}[G_T] = O\left(\sum_{i\neq i^\star}\frac{1+\log T}{\Delta(i)} + \theta_1\log K + d_{\max} + \sqrt{\theta_2 K}\right).$$

Finally, using

$$\theta_1 = 100 d_{\max}(K-1)^{1/3}, \qquad \theta_2 = 10^4 d_{\max}^2,$$

we have

$$\theta_1\log K = O(d_{\max}\sqrt{K}), \qquad \sqrt{\theta_2 K} = O(d_{\max}\sqrt{K}).$$

Thus,

$$\mathbb{E}[R_T] = \mathbb{E}[G_T] = O\left(\sum_{i\neq i^\star}\frac{1+\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right).$$

This is equivalent to:

$$\mathbb{E}[R_T] = O\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right),$$

which proves the stochastic guarantee. □

A.3 Adversarial setting

A.3.1 Fenchel stability

Lemma 17. *For any x such that $x(i) \geq 0$ we have that*

$$(p_t - p_t^\star)^\top x \leq \sum_{i=1}^{K} \min\{2\gamma_t^{-1} p_t(i)^{3/2}, \lambda_t^{-1} p_t(i)\} x(i) \sum_{s \in P_t} \widehat{\ell}_s(i)\,.$$

Proof. Denote by

$$r_t(x) = \lambda_t x(\log x - 1) - 2\gamma_t \sqrt{x}.$$

and note that $R_t(p) = \sum_{i=1}^{K} r_t(p(i))$.

The first and second derivatives of r_t are:

$$r_t'(x) = -\frac{\gamma_t}{\sqrt{x}} + \lambda_t \log(x)$$
$$r_t''(x) = \frac{2\lambda_t \sqrt{x} + \gamma_t}{2x^{3/2}}$$

In particular, r_t' is strictly increasing on $(0, 1]$.

The KKT conditions imply that there exist multipliers $\mu, \widetilde{\mu} \in \mathbb{R}$ such that, for every $i \in [K]$,

$$r_t'(p_t(i)) = -L_t^{\text{obs}}(i) + \mu, \qquad r_t'(p^\star(i)) = -L_t^\star(i) + \widetilde{\mu}.$$

We first show that

$$\mu \leq \widetilde{\mu}.$$

Suppose, for contradiction, that $\mu > \widetilde{\mu}$. Then, for every $i \in [K]$,

$$\begin{aligned} r_t'(p_t(i)) &= -L_t^{\text{obs}}(i) + \mu \\ &> -L_t^\star(i) + \widetilde{\mu} \\ &= r_t'(p_t^\star(i)), \end{aligned}$$

where the first inequality follows because $L_t^\star(i) \geq L_t^{\text{obs}}$ and $\mu > \widetilde{\mu}$.

Since r_t' is strictly increasing, this implies

$$p_t(i) > p_t^\star(i), \qquad \forall i \in [K].$$

Summing over i contradicts

$$\sum_{i=1}^{K} p_t(i) = \sum_{i=1}^{K} p_t^\star(i) = 1.$$

Thus,

$$\mu \leq \widetilde{\mu}\,.$$

Now, denote by $v_t = (r_t')^{-1}$ the inverse function of r_t'. By Lemmas 18 and 19 we have that v_t is convex. Since $p_t(i) = v_t(-L_t^{\text{obs}}(i) + \mu)$ and $p_t^\star(i) = v_t(-L_t^\star(i) + \widetilde{\mu})$ we have that

$$\begin{aligned} p_t(i) - p_t^\star(i) &= v_t(-L_t^{\text{obs}}(i) + \mu) - v_t(-L_t^\star(i) + \widetilde{\mu}) \\ &\leq v_t'(-L_t^{\text{obs}}(i) + \mu) \left(\mu - \widetilde{\mu} + \sum_{s \in P_t} \widehat{\ell}_s(i)\right) \\ &= \frac{1}{r_t''(v_t(-L_t^{\text{obs}}(i) + \mu))} \left(\mu - \widetilde{\mu} + \sum_{s \in P_t} \widehat{\ell}_s(i)\right) && \text{(Lemma 18)} \\ &= \frac{1}{r_t''(p_t(i))} \left(\mu - \widetilde{\mu} + \sum_{s \in P_t} \widehat{\ell}_s(i)\right) \\ &\leq \frac{1}{r_t''(p_t(i))} \sum_{s \in P_t} \widehat{\ell}_s(i)\,. \end{aligned}$$

Since $\frac{1}{r_t''(p_t(i))} \leq \min\{2\gamma_t^{-1}p_t(i)^{3/2}, \lambda_t^{-1}p_t(i)\}$, we have that for any x such that $x(i) \geq 0$

$$(p_t - p_t^\star)^\top x \leq \sum_{i=1}^{K} \min\{2\gamma_t^{-1}p_t(i)^{3/2}, \lambda_t^{-1}p_t(i)\}x(i) \sum_{s\in P_t} \widehat{\ell}_s(i)\,,$$

which concludes the proof. □

The following is a standard result.

Lemma 18. *Let $D \subseteq \mathbb{R}$ be an interval and let $f\colon D \to \mathbb{R}$ be concave and differentiable with $f'(x) > 0$ for all $x \in D$. Set $I = f(D)$ and let $v\colon I \to D$ be the inverse of f. Suppose that v is differentiable. Then:*

(i) $I \subset \mathbb{R}$ is an interval, and v is strictly increasing on I;

(ii) $v'(y) = \frac{1}{f'(v(y))} > 0$ for all $y \in I$;

(iii) v is convex on I.

Proof. (i) Since $f' > 0$, f is strictly increasing and hence injective; thus $v = f^{-1}\colon I \to D$ is well defined. As f is continuous and D is an interval, $I = f(D)$ is an interval by the intermediate value theorem.

We first prove *(I)*. To see that v is strictly increasing assume that v is decreasing: if $y_1 < y_2$ in I and $v(y_1) \geq v(y_2)$, then applying the increasing map f yields $y_1 \geq y_2$, which is a contradiction.

On to *(ii)*. Since v is the inverse of f, we have $f(v(y)) = y$ for all $y \in I$. Both sides are differentiable on I, so differentiating with the chain rule gives

$$f'(v(y))\, v'(y) = 1 \qquad \text{for all } y \in I.$$

Because $f' > 0$ on D, we may divide by $f'(v(y))$ to obtain

$$v'(y) = \frac{1}{f'(v(y))} > 0.$$

(iii) A differentiable concave function has a nonincreasing derivative, so f' is nonincreasing on D. By (i), v is increasing, hence $f' \circ v$ is nonincreasing on I. Since $f' \circ v$ is strictly positive and $t \mapsto 1/t$ is strictly decreasing on $(0, \infty)$, the function

$$v' = \frac{1}{f' \circ v}$$

is nondecreasing on I. A differentiable function with nondecreasing derivative is convex: for $y_1 < y_2$ in I and $\alpha \in (0,1)$, put $y_\alpha = \alpha y_1 + (1-\alpha)y_2$. By the mean value theorem there exist $\xi_1 \in (y_1, y_\alpha)$ and $\xi_2 \in (y_\alpha, y_2)$ with

$$v(y_\alpha) - v(y_1) = v'(\xi_1)(y_\alpha - y_1), \qquad v(y_2) - v(y_\alpha) = v'(\xi_2)(y_2 - y_\alpha).$$

Since $\xi_1 < \xi_2$, we have $v'(\xi_1) \leq v'(\xi_2)$. Using $y_\alpha - y_1 = (1-\alpha)(y_2 - y_1)$ and $y_2 - y_\alpha = \alpha(y_2 - y_1)$,

$$\alpha\big(v(y_\alpha) - v(y_1)\big) = \alpha(1-\alpha)(y_2-y_1)\, v'(\xi_1) \leq \alpha(1-\alpha)(y_2-y_1)\, v'(\xi_2) = (1-\alpha)\big(v(y_2) - v(y_\alpha)\big),$$

which rearranges to $v(y_\alpha) \leq \alpha v(y_1) + (1-\alpha)v(y_2)$. Hence v is convex on I. □

The following is a standard result.

Lemma 19. *Let $f(x) = -\dfrac{\gamma_t}{\sqrt{x}} + \lambda_t \log x$ for $x > 0$, and let v be the inverse of f. Then v is differentiable.*

Proof. We have $f'(x) = \dfrac{\gamma_t}{2x^{3/2}} + \dfrac{\lambda_t}{x} > 0$ for all $x > 0$, so f is strictly increasing, and hence so is v.

Fix y_0 in the domain of v, let $x_0 = v(y_0)$, and for $y \neq y_0$ let $x = v(y) \neq x_0$. By the mean value theorem there is ξ between x_0 and x with $y - y_0 = f(x) - f(x_0) = f'(\xi)(x - x_0)$, i.e.

$$\frac{v(y) - v(y_0)}{y - y_0} = \frac{1}{f'(\xi)}. \tag{$*$}$$

For $|y - y_0| \leq 1$, monotonicity of v places x, x_0 and ξ in the compact interval $[a, b] := [v(y_0 - 1), v(y_0 + 1)] \subset (0, \infty)$, on which the continuous positive function f' has a minimum $m > 0$. Hence by $(*)$, $|v(y) - v(y_0)| \leq |y - y_0|/m$, so $x \to x_0$ and therefore $\xi \to x_0$ as $y \to y_0$. Since f' is continuous with $f'(x_0) > 0$, the right-hand side of $(*)$ converges to $1/f'(x_0)$. Thus v is differentiable at y_0. □

Lemma 5 (RESTATED). *We have that* $\mathbb{E}[\lambda_{T+d_{\max}+1} \log(K)] \leq \lambda_1 \log(K) + 2\sqrt{\log(K)D}$.

Proof. Let $X_t = (p_t - p_t^\star)^\top \widehat{\ell}_t$. Recall that

$$\lambda_t = \theta_1 + \frac{1}{\log(K)} \max_{s \leq t} \{S_t\},$$

where

$$S_{t+1} = \sum_{s \in \mathcal{O}_{t+1}} X_s \mathbb{1}[\max_{s' \leq s} s' + d_{s'} \leq t]$$

First, we need control over $\lambda_t - \lambda_s$ for s such that $t - s \leq d_{\max}$. For that, we observe that $X_t \leq 1$ since by Lemma 12, $p_s^\star(i) \leq \frac{6}{5} p_s(i)$. Consequently, for s such that $t - s \leq d_{\max}$, we have that

$$\max_{s \leq t+1} \{S_t\} - \max_{s \leq t} \{S_t\} \leq 2d_{\max}.$$

Therefore, for s such that $t - s \leq d_{\max}$, we have that

$$\frac{\lambda_t}{\lambda_s} \leq 1 + \frac{2d_{\max}}{\lambda_s} \leq 2,$$

where the last inequality is due to the definition of λ_s. By Lemma 17 we have that

$$X_t \leq \sum_{i=1}^{K} \lambda_t^{-1} p_t(i) \sum_{s \in P_t} \widehat{\ell}_s(i) \widehat{\ell}_t(i).$$

Let

$$Z_t = \sum_{i=1}^{K} p_t(i) \sum_{s \in P_t} \widehat{\ell}_s(i) \widehat{\ell}_t(i) \mathbb{1}[\max_{s' \leq s} s' + d_{s'} = t].$$

We have that

$$\begin{aligned}
\max_{s \leq t+1} \{S_t\} - \max_{s \leq t} \{S_t\} &= \Big(S_{t+1} - \max_{s \leq t} \{S_t\} \Big)_+ \\
&\leq \Big(S_{t+1} - S_t \Big)_+ \\
&= \Big(\sum_{s=1}^{t} X_s \mathbb{1}[\max_{s' \leq s} s' + d_s = t] \Big)_+ \\
&\leq \sum_{s=1}^{t} \lambda_s^{-1} Z_s .
\end{aligned}$$

Using the identity

$$\lambda_{t+1}^2 - \lambda_t^2 = (\lambda_{t+1} - \lambda_t)(\lambda_{t+1} + \lambda_t),$$

together with $\lambda_t \leq \lambda_{t+1}$, we obtain

$$\begin{aligned}
\lambda_{t+1}^2 - \lambda_t^2 &\leq 2\lambda_{t+1}(\lambda_{t+1} - \lambda_t) \\
&\leq 2\lambda_t \sum_{s=1}^{t} \lambda_s^{-1} Z_s \\
&\leq 4 \sum_{s=1}^{t} Z_s .
\end{aligned}$$

Summing over $t = 1, \dots, T$ we obtain

$$\begin{aligned}
\mathbb{E}[\lambda_{T+d_{\max}+1}^2] - \lambda_1^2 &= \mathbb{E}[\sum_{t=1}^{T+d_{\max}+1} (\lambda_{t+1}^2 - \lambda_t^2)] \\
&\le \frac{4}{\log(K)} \mathbb{E}[\sum_{t=1}^{T+d_{\max}+1} \sum_{s=1}^{t} Z_s] \\
&= \frac{4}{\log(K)} \mathbb{E}[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \sum_{s \in P_t} \widehat{\ell}_s(i) \widehat{\ell}_t(i)] \\
&= \frac{4}{\log(K)} \mathbb{E}[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \sum_{s \in P_t} \ell_s(i) \ell_t(i)] \\
&\qquad \text{(unbiasedness and conditional independence of } \ell_s(i) \text{ and } \ell_t(i) \text{ due to delay)} \\
&\le \frac{4}{\log(K)} \mathbb{E}[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) |P_t|] \qquad \text{(boundedness of } \ell_t(i)) \\
&= \sum_{t=1}^{T} \frac{4}{\log(K)} d_t = \frac{D}{\log(K)} \,.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}[\lambda_{T+d_{\max}+1}] \log(K) &\le \log(K) \sqrt{\mathbb{E}[\lambda_{T+d_{\max}+1}^2]} \qquad \text{(Jensen's inequality)} \\
&\le \lambda_1 \log(K) + 2\sqrt{\log(K) D} \,,
\end{aligned}$$

which completes the proof.

□

A.3.2 Detailed proof of equation (3)

Theorem 1 (RESTATED)**.** *Algorithm 1 guarantees that*

$$\mathcal{R}_T \le \begin{cases} O\left(\sum_{i \neq i^\star} \frac{\log T}{\Delta(i)} + d_{\max}\sqrt{K}\right) & \textit{stochastic setting} \quad (2) \\ O\left(\sqrt{KT} + \sqrt{D \log K} + d_{\max}\sqrt{K}\right) & \textit{adversarial setting} \quad (3) \end{cases}$$

Proof. We start from the shared regret bound in Equation (5). For any comparator $u = e_j$, it gives

$$\mathbb{E}[R_T(u)] \le 2\mathbb{E}[\lambda_{T+d_{\max}+1} \log K] + \frac{131}{5} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \neq j} \sqrt{\frac{p_t(i)}{t}}\right] + 2\sqrt{\theta_2 + 1}\sqrt{K}.$$

By Lemma 5, the adaptive learning-rate term satisfies

$$\mathbb{E}[\lambda_{T+d_{\max}+1} \log K] \le \lambda_1 \log K + \sqrt{D \log K}.$$

Hence,

$$2\mathbb{E}[\lambda_{T+d_{\max}+1} \log K] \le 2\lambda_1 \log K + 2\sqrt{D \log K}.$$

It remains to control the Tsallis stability term. Since $p_t \in \Delta^K$, Cauchy–Schwarz gives

$$\sum_{i \neq j} \sqrt{p_t(i)} \le \sum_{i=1}^{K} \sqrt{p_t(i)} \le \sqrt{K \sum_{i=1}^{K} p_t(i)} = \sqrt{K}.$$

Together with

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T},$$

this implies

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \neq j} \sqrt{\frac{p_t(i)}{t}}\right] \leq 2\sqrt{KT}.$$

Substituting these two estimates into Equation (5), we obtain

$$\mathbb{E}[R_T(u)] \leq \frac{262}{5}\sqrt{KT} + 2\sqrt{D \log K} + 2\lambda_1 \log K + 2\sqrt{\theta_2 + 1}\sqrt{K}.$$

Recall that

$$\lambda_1 = \theta_1 = 100 d_{\max}(K-1)^{1/3}, \qquad \theta_2 = 10^4 d_{\max}^2.$$

Therefore,

$$\lambda_1 \log K = O\left(d_{\max} K^{1/3} \log K\right)$$

and

$$\sqrt{\theta_2 + 1}\sqrt{K} = O\left(d_{\max}\sqrt{K}\right),$$

where we use $d_{\max} \geq 1$. Moreover,

$$K^{1/3} \log K = O(\sqrt{K}),$$

so that

$$\lambda_1 \log K = O\left(d_{\max}\sqrt{K}\right).$$

Consequently,

$$\mathbb{E}[R_T(u)] = O\left(\sqrt{KT} + \sqrt{D \log K} + d_{\max}\sqrt{K}\right).$$

Since $j \in [K]$ was arbitrary, choosing j to be the best arm in hindsight proves

$$\mathbb{E}[R_T] = O\left(\sqrt{KT} + \sqrt{D \log K} + d_{\max}\sqrt{K}\right),$$

which establishes Equation (3).

□

B Details of Section 4 (Experiments)

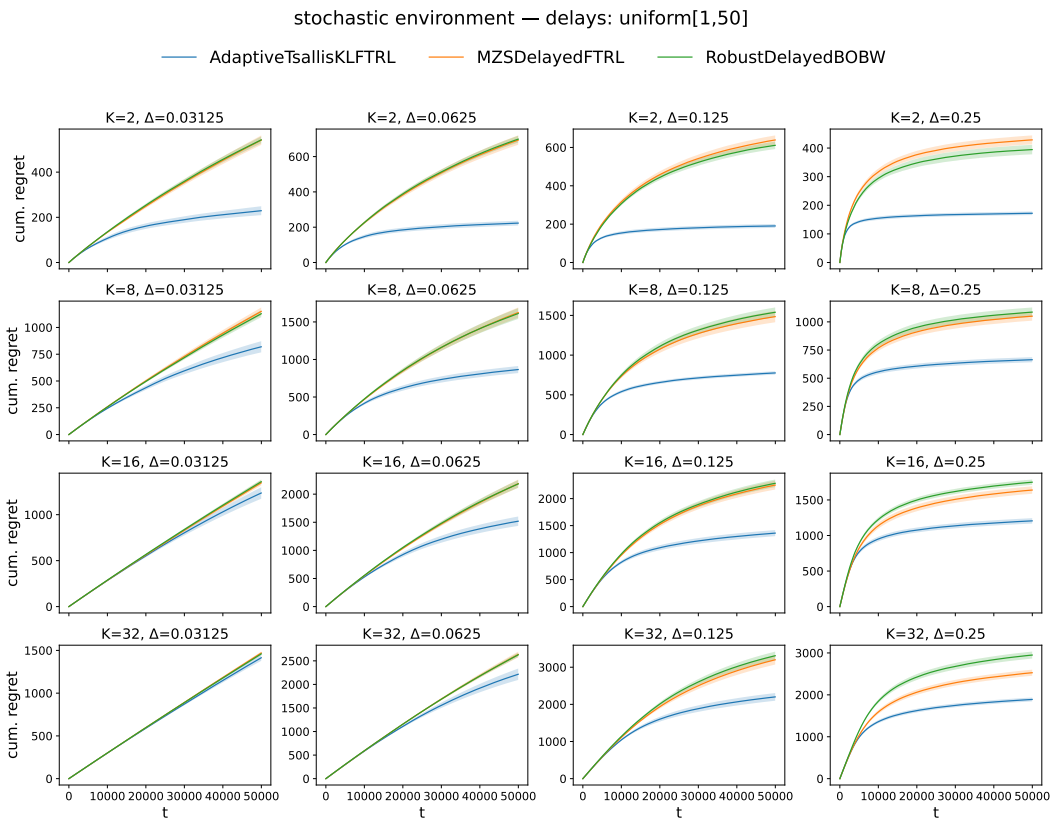


Figure 2: Results for the stochastic environment with Uniform$(1, 50)$ delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.

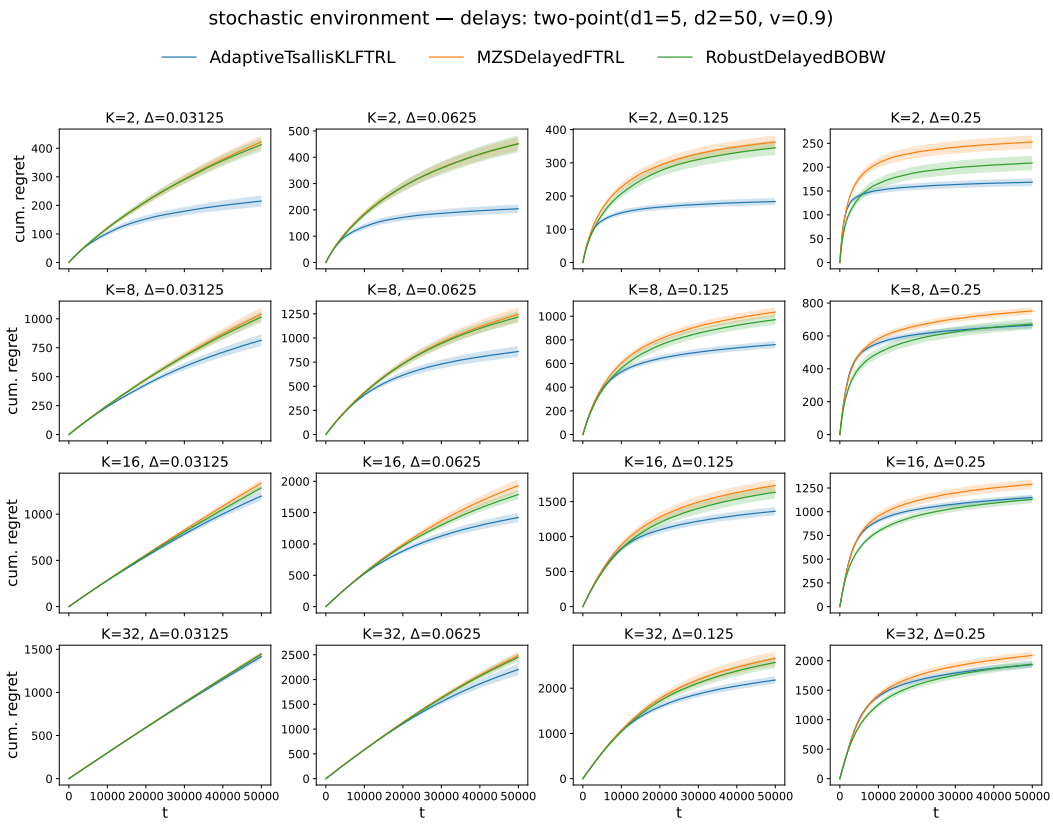


Figure 3: Results for the stochastic environment with two point delays delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.

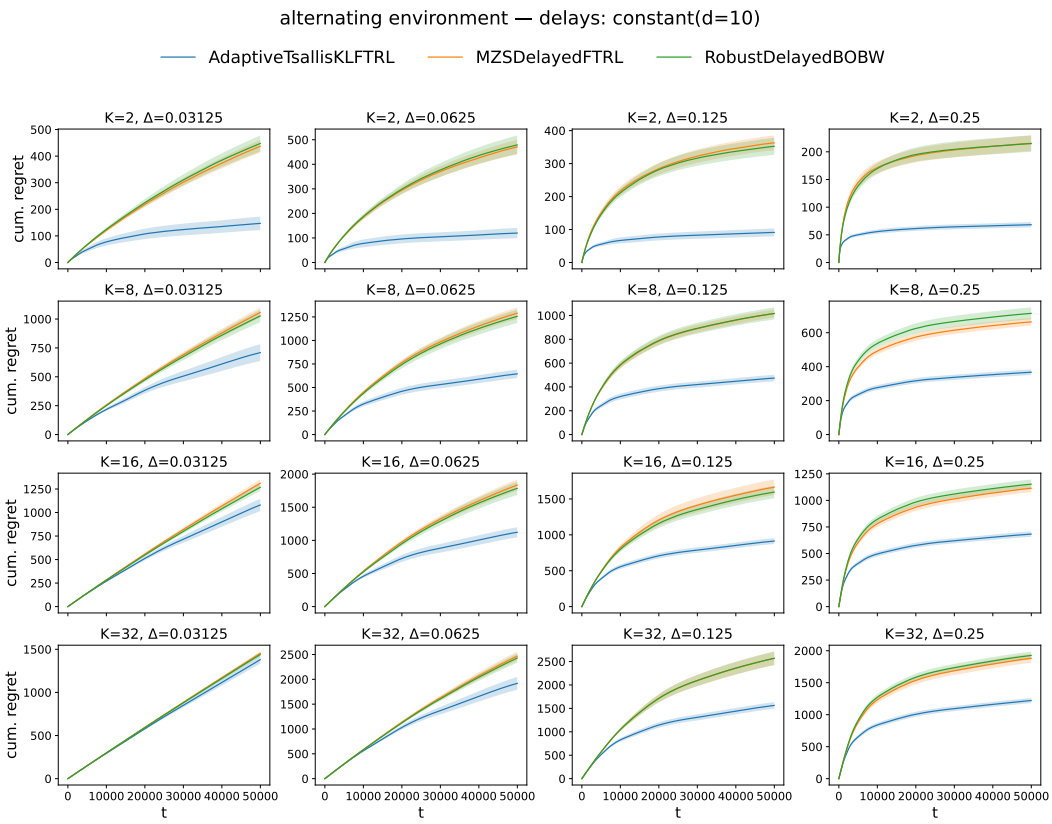


Figure 4: Results for the alternating environment with constant ($d_t = 10$) delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.

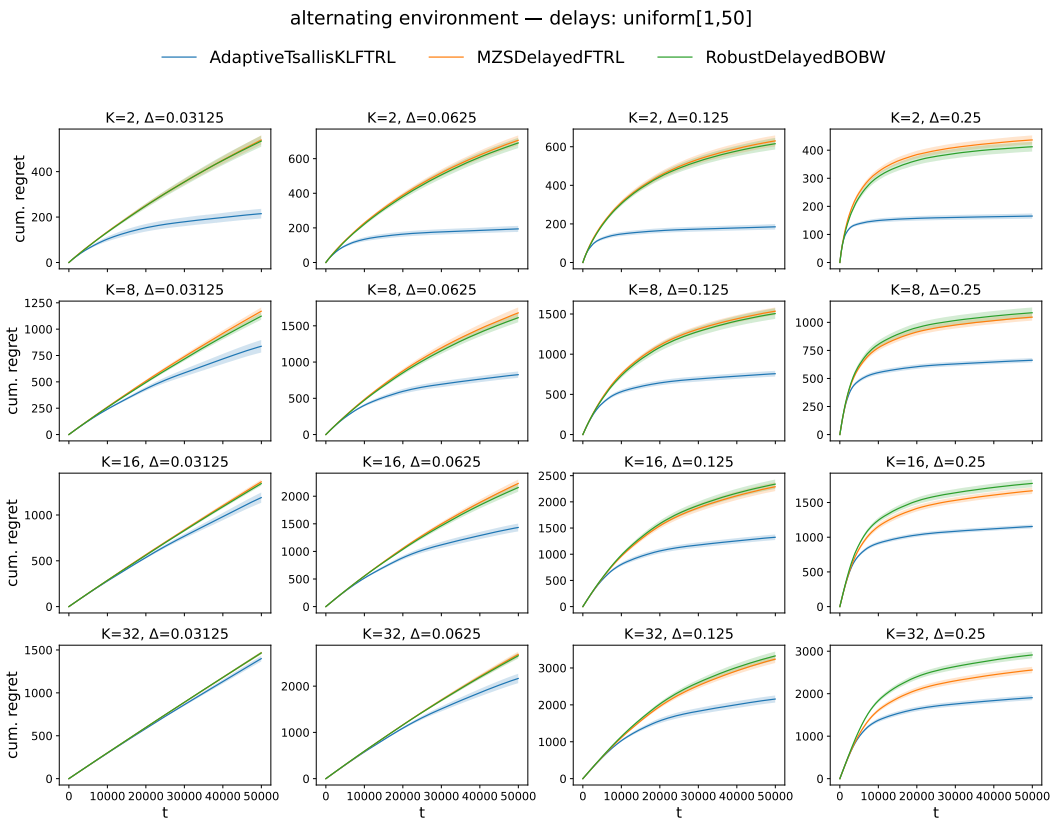


Figure 5: Results for the alternating environment with Uniform$(1, 50)$ delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.

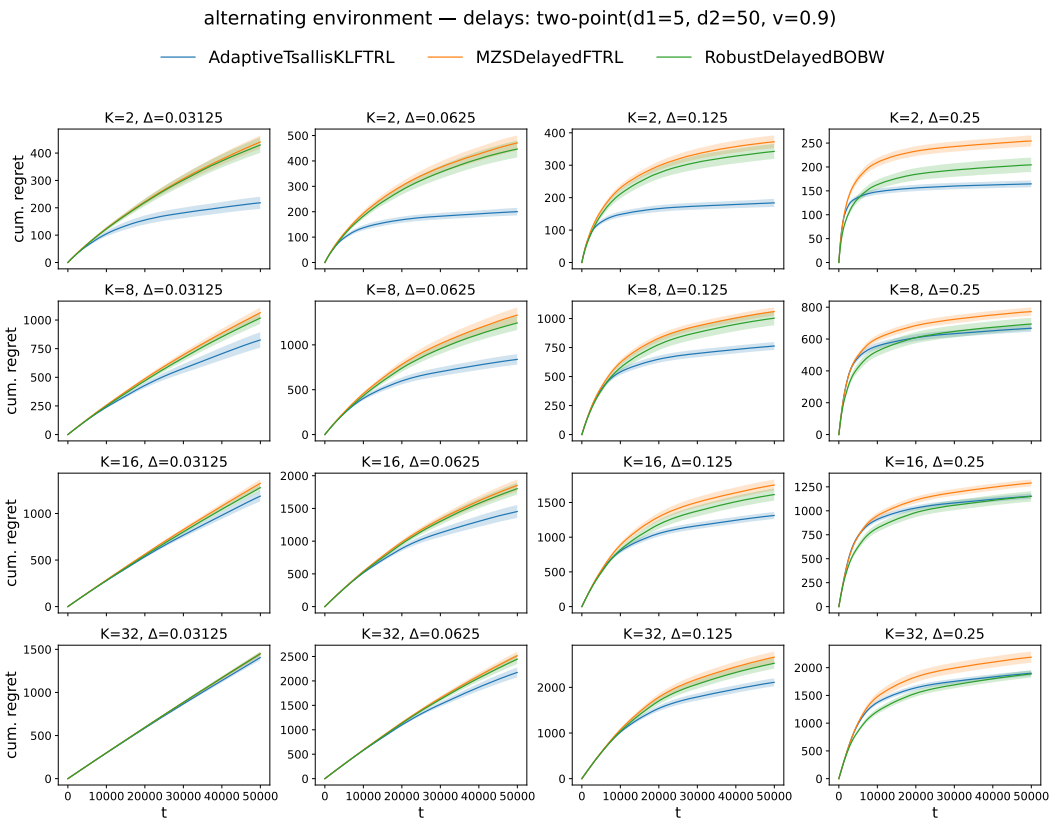


Figure 6: Results for the alternating environment with two point delays. Solid lines are the means after 30 repetitions. The shaded areas are 95% confidence intervals.