No Discounted-Regret Learning in Adversarial Bandits with Delays

Ilai Bistritz* bistritz@stanford.edu
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Zhengyuan Zhou zzhou@stern.nyu.edu
Stern School of Business
New York University
New York, NY 10003, USA

Xi Chen† xchen3@stern.nyu.edu
Stern School of Business
New York University
New York, NY 10003, USA

Nicholas Bambos bambos@stanford.edu
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Jose Blanchet jose.blanchet@stanford.edu
Department of Management Science & Engineering
Stanford University
Stanford, CA 94305, USA

Editor:

Abstract

Consider a player that in each round t out of T rounds chooses an action and observes the incurred cost after a delay of d_t rounds. The cost functions and the delay sequence are chosen by an adversary. We show that even if the players' algorithms lose their "no regret" property due to too large delays, the expected discounted ergodic distribution of play converges to the set of coarse correlated equilibrium (CCE) if the algorithms have "no discounted-regret". For a zero-sum game, we show that no discounted-regret is sufficient for the discounted ergodic average of play to converge to the set of Nash equilibria. We prove that the FKM algorithm with n dimensions achieves a regret of $O\left(nT^{\frac{3}{4}}+\sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ and the EXP3 algorithm with K arms achieves a regret of $O\left(\sqrt{\ln K\left(KT+D\right)}\right)$ even when $D=\sum_{t=1}^{T}d_t$ and T are unknown. These bounds use a novel doubling trick that provably retains the regret bound for when D and T are known. Using these bounds, we show that EXP3 and FKM have no discounted-regret even for $d_t=O\left(t\log t\right)$. Therefore, the CCE of a finite or convex unknown game can be approximated even when only delayed bandit feedback is available via simulation.

Keywords: Online Learning, Adversarial Bandits, Non-cooperative Games, Delays

∗. This is an extended version of Bistritz et al. (2019). For details, see Subsection 1.1 on previous work. This research was supported by the Koret Foundation grant for Smart Cities and Digital Living
†. Xi Chen is supported by NSF via IIS-1845444

1. Introduction

Consider an agent that makes sequential decisions, and each decision incurs some cost. The cost sequences are chosen by an adversary that knows the agent's algorithm. The agent's goal is to minimize this cost over time. The question of **what** the agent learns about the cost functions naturally influences the best performance the agent can guarantee. With full information, after acting at round t, the agent receives the cost function of round t as feedback. With bandit feedback, as we consider here, the agent only receives the cost of her decision. Indeed, receiving only bandit feedback against an adversary results in worse regret bounds than that in the full information case or with a stochastic environment.

Another fundamental question is **when** the agent receives the feedback. The delayed feedback occurs naturally in practice. In most practical learning environments, each agent only gets to learn the cost of her action, but she does not learn it immediately. For example, it takes a while to observe the effect of a decision on a treatment plan. It also requires a certain amount of time before observing the market's response to an advertisement. Online learning with no delays means that the agent always knows how beneficial all the past actions were while making the current decision. However, when the feedback is delayed, decisions must be made before all the feedback from the past choices is received. Receiving delayed bandit feedback against an adversary is certainly worse than receiving mere bandit feedback. Fortunately, we show the commonly used algorithms FKM and EXP3 exhibit surprising robustness to delayed feedback both in terms of the single-agent regret and the emergence of game-theoretic equilibria for the multi-agent case.

Practical environments are non-stationary since they typically consist of other learning agents, and the learning of one agent affects that of the other agents. Moreover, the costs are naturally correlated over time. Hence, guarantees for stochastic environments are not strong enough. Proving regret bounds against an adversary evaluates the robustness of a learning algorithm, regardless of whether an actual malicious adversary exists or not. Following the same reasoning, proving regret bounds with adversarial delays evaluates the robustness of the algorithm to non-stationary delays. Different algorithms can behave differently while facing delays. Since delays could be difficult to quantify and implicit to the decision-maker, a natural question is whether the most widely used online learning algorithms are robust to delays. As such, our goal is not to suggest new algorithms but rather to show the encouraging result that the widely used algorithms are naturally robust to delays, provided with the right tuning.

The two most studied action spaces in the literature are the discrete action set $1, ..., K$ (i.e., arms in multi-armed bandits) and a convex and compact action set $\mathcal{K} \subset \mathbb{R}^n$. For bandit convex optimization with a convex compact action set, the most widely used adversarial bandits learning algorithms is FKM (Flaxman et al., 2005). With no delays, the expected regret of FKM is $O\left(nT^{\frac{3}{4}}\right)$ where n is the dimension of $\mathcal{K}$. For the discrete case, the most popular adversarial bandits learning algorithm is EXP3 (Auer et al., 1995, 2002; Bubeck et al., 2012; Neu et al., 2010). With no delays, the expected regret of EXP3 is $O\left(\sqrt{TK\ln K}\right)$. Having a sublinear regret, the average regret per round of both algorithms goes to zero as $T \to \infty$. This is often considered a criterion for successful learning and is known as the "no-regret property" (Bowling, 2005).

Our first main contribution is to show that with an arbitrary sequence of delays d_t (which can grow unbounded), FKM achieves an expected regret of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} + |\mathcal{M}|\right)$, where $\mathcal{M}$ is the set of rounds that their feedback is not received before round T. Our second main contribution is to show that with an arbitrary sequence of delays d_t, EXP3 achieves an expected regret of $O\left(\sqrt{\left(KT + \sum_{t\notin\mathcal{M}} d_t\right)\ln K} + |\mathcal{M}|\right)$. These expressions make clear which delay sequences will maintain the no-regret property intact and which will lead to linear regret in T.

Surprisingly, these regret bounds reveal inherent robustness to delays that both EXP3 and FKM enjoy. For example, if $d_t \leq K$ for all t, then EXP3 achieves the same $O\left(\sqrt{TK\ln K}\right)$ that exponential weights with $d_t \leq K$ achieves in the full information case (Quanrud and Khashabi, 2015), or that EXP3 achieves with no delays (Auer et al., 1995, 2002). Even more impressive, if $d_t \leq t^{1/4}$ for all t, then FKM achieves the same order of regret it would have achieved **with no delays.**

Like the horizon T, the sum of delays $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$ is not always known to the decision-maker. Hence, T and D might not be available while tuning the algorithm parameters. While the standard doubling trick (Cesa-Bianchi et al., 1997) can be used to deal with an unknown T, it does not help with an unknown D. Our third main contribution is a general novel two-dimensional doubling trick where epochs are indexed by a "delay index" as well as a "time index". The delay index doubles every time the number of missing samples so far doubles, and the time index doubles with the rounds as usual. We show that under general conditions, this novel doubling trick can be applied to an online learning algorithm beyond the case of adversarial bandits or bandit feedback. We apply this result to achieve an expected regret bound of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ for FKM and of $O\left(\sqrt{(KT+D)\ln K}\right)$ for EXP3.

An omnipotent adversary represents the embodiment of the agent's worst fears when learning to optimize her decisions in an unknown environment. In practice, it is more likely that the opponents that the agent faces are online learning agents like herself, which have limited knowledge and power. These agents have interests of their own, that are conflicting but not as malicious as that of an adversary. Non-cooperative games are the natural framework to analyze the outcome of the interaction against other agents instead of against an all-powerful adversary. In the worst-case, these interests are in direct conflict with those of our agent, as in the case of a zero-sum game. Interestingly enough, it turns out that with delayed feedback, the outcome of playing against other agents can be essentially different from playing against an all-powerful adversary.

It is well known that when N agents use a no-regret learning algorithm against each other in a non-cooperative game, the dynamics will result in a coarse correlated equilibrium (CCE) (Hannan, 1957; Hart, 2013). For a zero-sum game, the dynamics result in a Nash equilibrium (NE) (Cai and Daskalakis, 2011). The emergence of a CCE or a NE in a game between no-regret learners provides yet another strong evidence for the importance of these game-theoretic equilibrium concepts. From a practical point of view, the convergence of the expected ergodic distribution to the set of CCE or of the ergodic average to the set of NE makes no-regret algorithms an appealing way to approximate a CCE or a NE when the game matrix is unknown so only simulating the game is possible (see Hellerstein et al. (2019)). When simulating an unknown game, bandit feedback is a more realistic assumption than full information (or gradient feedback).

The convergence to the set of CCE is maintained if the algorithm still enjoys the no-regret property in the presence of delayed feedback. However, for large enough delays (e.g. $d_t = O(t \log t)$), the regret of EXP3 and FKM (or any other algorithm) becomes linear in the horizon T so the no-regret property no longer holds. Our fourth main contribution in this paper is to show that even with delays that cause a linear regret, the discounted ergodic distribution may still converge to a CCE (or to the set of NE for a zero-sum game). This means that computing a CCE or a NE using EXP3 or FKM is still possible even in scenarios where these algorithms do not enjoy a sublinear regret (i.e., no-regret) due to large delays. Since delays are a prominent feature of almost every computational environment, this is an encouraging finding.

1.1 Previous Work

In recent years, online learning with delayed feedback has attracted considerable attention (see Cesa-Bianchi et al. (2019); Agarwal and Duchi (2011); Neu et al. (2010); Mandel et al. (2015); Vernade et al. (2017, 2020)). Most literature deals with the multi-armed case with a discrete set of actions. Fixed delays were considered in Weinberger and Ordentlich (2002) and Zinkevich et al. (2009). Stochastic rewards and stochastic delays have been considered in Pike-Burke et al. (2018). Bandits with adversarial rewards but still stochastic i.i.d. delays were considered in Joulani et al. (2013).

Cesa-Bianchi et al. (2019) considered an interesting case of communications agents that cooperate to solve a common adversarial bandit problem, where the messages between agents may arrive after a bounded delay with a known bound d. Recently, advancements were made for the case of stochastic delays Manegueu et al. (2020); Vernade et al. (2020); Zhou et al. (2019). In contrast, in our scenario, the adversary chooses the delay sequence.

In Quanrud and Khashabi (2015), the case of delayed full information feedback has been considered, where the feedback is the costs of all arms (or the gradient of the cost function). Our goal here is to address the more challenging bandit feedback scenario, motivated by the multi-agent case.

In Cesa-Bianchi et al. (2018), a different adversarial bandits with delayed feedback scenario has been studied, where all the feedback that is received at the same round is summed up and cannot be distinguished, and delays are bounded by d. For both the multi-armed and convex cases, Cesa-Bianchi et al. (2018) designed a wrapper algorithm and proved a regret bound for their delayed feedback scenario as a function of the regret of the algorithm being wrapped for the non-delayed scenario. For EXP3, the resulting regret bound is $O\left(\sqrt{dTK\ln K}\right)$. Compared to their scenario, we consider time-stamped feedback with delays that can be unbounded.

Multi-agent learning and convergence to NE under delays have been considered in Zhou et al. (2017); Héliou et al. (2020) for variationally stable games and monotone games, which both generalize potential games. We study general non-cooperative games and convergence to the set of CCE under delays by generalizing the framework of no-regret learning to no discounted-regret learning. Our analysis applies to both the multi-armed bandit and bandit convex optimization cases. We also study convergence to NE for zero-sum games, which is a well-studied class of games that are not variationally stable or monotone (see Mertikopoulos and Zhou (2016)).

This paper extends a preliminary conference version (Bistritz et al., 2019) that only analyzed the EXP3 algorithm under delays. In this journal version, we also analyze the FKM algorithm for the bandit convex optimization under delays. Additionally, we improve the doubling trick of Bistritz et al. (2019) and show that it can be applied to any online learning algorithm with delayed feedback. Last but not least, we generalize the game-theoretic results of Bistritz et al. (2019) to non-cooperative games and correlated equilibrium.

While preparing this journal version, we became aware that the concurrent work of Thune et al. (2019) published in the same conference as (Bistritz et al., 2019) provides a similar analysis for the single-agent EXP3 case with a constant step size $\eta_t = \eta$. Taking a different approach to deal with unknown $D = \sum_t d_t$ and T, Thune et al. (2019) assume that the delays are available at action time. In this work, we instead provide a novel doubling trick that does not require this assumption and achieves the same $O\left(\sqrt{(TK+D)\ln K}\right)$ that was achieved in Thune et al. (2019). This improves the doubling trick that was proposed in Bistritz et al. (2019) that achieved an expected regret of $O\left(\sqrt{(TK^2+D)\ln K}\right)$. Replacing EXP3 by a novel follow-the-regulated-leader algorithm, Zimmert and Seldin (2019) improved the expected regret to the optimal $O\left(\sqrt{TK+D\ln K}\right)$ even when D is unknown without using a doubling trick.

While this paper was under review, György and Joulani (2020) have achieved $O\left(\sqrt{\ln K\,(TK+D)}\right)$ regret for EXP3 by adaptively tuning the step size. As opposed to our EXP3 bound, their regret bound has been shown to also hold with high probability. Furthermore, assuming a bound on the maximal delay (or that the delay is available at action time) György and Joulani (2020) propose a data-adaptive version of EXP3 which yields a regret that depends on the cumulative cost.

The works Thune et al. (2019); Zimmert and Seldin (2019); György and Joulani (2020); Bistritz et al. (2019) all study the multi-armed bandit problem, while this paper also studies the convex bandit optimization case, using the FKM algorithm under delayed feedback. Moreover, Thune et al. (2019); Zimmert and Seldin (2019); György and Joulani (2020) only studied the single-agent problem while we are mainly motivated by the multi-agent problem, proving convergence to the set of CCE under delayed feedback. Our emphasis on the multi-agent case leads to two technical differences even in our single-agent results. First, we prove regret bounds also against an adaptive adversary that can choose the cost function in response to the players' past actions. This distinction between an oblivious and adaptive adversary is necessary to show convergence to the set of CCE. Additionally, our single-agent results are formulated using the "discounted-regret", which weights the costs of different turns according to the step size sequence. Even for the EXP3 analysis, this formulation leads to several subtleties that did not arise in Thune et al. (2019); Zimmert and Seldin (2019); György and Joulani (2020) (e.g., in Lemma 8 and Lemma 9). Last but not least, this paper provides a novel doubling trick that can deal with an unknown sum of delays in addition to the unknown horizon that the standard doubling trick addresses. Our novel doubling trick can be applied to any online learning algorithm under delayed feedback, beyond the case of adversarial bandits. For example, the preliminary version of this doubling trick was employed in Lancewicki et al. (2020) that proposed novel learning algorithms for delayed feedback in adversarial Markov decision processes.

1.2 Outline

Section 2 formalizes the general problem of learning with delayed bandit feedback and highlights our main results. Section 3 discusses the outcome of the interaction between multiple learners that are each subjected to a possibly different delay sequence. We extend the well-known connection between no-regret learning and CCE to learning with delayed feedback. Surprisingly, even algorithms that have linear regret under delays can still lead to the set of CCE. Section 4 presents our general doubling trick that can be applied to online learning algorithms with delayed feedback, not necessarily against an adversary or with bandit feedback. The results of Section 3 and Section 4 can be applied to a wide variety of online learning algorithms with bandit feedback delay. Section 5 and Section 6 consider the FKM algorithm for adversarial bandit convex optimization and the EXP3 algorithm for adversarial multi-armed bandits, respectively. Section 5 and Section 6 each starts by proving an expected regret bound under delayed bandit feedback for the algorithm in consideration, both for the case of an oblivious adversary and an adaptive adversary. Next, we apply the result on our doubling trick for both FKM and EXP3 to obtain expected regret bounds for the case where T and D are unknown. Then, we show that FKM and EXP3 are no discounted-regret algorithms even with respect to delay sequences for which they both have linear regret in T. This allows us to apply our game-theoretic results for both FKM and EXP3, showing that they are useful to learn a CCE or a NE (in a zero-sum game) in simulated environments where only delayed bandit feedback is available. Finally, Section 7 concludes the paper. Long proofs are postponed to the appendix.

2. Problem Formulation

Consider a player that in each round t from 1 to T has to pick a strategy $\boldsymbol{a}_t \in \mathcal{K}$ from a set $\mathcal{K}$. The cost at round t from playing $\boldsymbol{a}_t$ is $l_t(\boldsymbol{a}_t) \in [0,1]$. We consider two types of adversaries:

1. **Oblivious Adversary**: chooses the cost functions $l_1, ..., l_T$ before the game starts.
2. **Adaptive Adversary**: chooses the cost function l_t after observing $\{a_1, \ldots a_{t-1}\}$, for each t.

With full information and no delays, the player gets to know the function l_t immediately after playing $\boldsymbol{a}_t$. In the bandit delayed feedback scenario, the player only gets to know the value of $l_t(\boldsymbol{a}_t)$ at the beginning of round $t + d_t$. (i.e., after a delay of $d_t \geq 1$ rounds). The adversary (oblivious or adaptive) chooses the delay sequence $\{d_t\}$ before the game starts.

We assume that the feedback includes the timestamp when the cost was generated. This is a mild assumption as the timestamp information is available in many applications of interest. If the delays are bounded by d, Cesa-Bianchi et al. (2018) have shown that EXP3 can still be implemented even with no timestamps, with regret $O(\sqrt{dTK \ln K})$ instead of $O\left(\sqrt{(d+K)T \ln K}\right)$. For our case of unbounded delays, it is not clear if FKM and EXP3 can be implemented without timestamps.

The set of costs (feedback samples) received **and** used at round t is denoted $\mathcal{S}_t$, so $s \in \mathcal{S}_t$ means that the cost of $\boldsymbol{a}_s$ from round s is received and used at round t. Since the game lasts for T rounds, all costs for which $t + d_t > T$ are never received. Of course, the value of d_t does not matter as long as $t + d_t > T$, and these are just samples that the adversary chose to prevent the player from receiving. We name these costs the missing samples and denote their set by $\mathcal{M}$.

Define s_-, s_+ as the steps a moment before and after the algorithm uses the feedback from round s, respectively. These steps are taking place in round t if $s \in \mathcal{S}_t$.

The player wants to have a learning algorithm that uses past observations to make good decisions over time. The performance of the player's algorithm is measured using the regret. The expected regret is the total expected cost over a horizon of T rounds, compared to the total cost that results from playing the best fixed action in all rounds:

Definition 1. Let $\boldsymbol{a}^* = \underset{\boldsymbol{a}\in\mathcal{K}}{\arg\min} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. The expected regret is defined as

$$\mathbb{E}\left\{R(T)\right\} \triangleq \sum_{t=1}^{T} \mathbb{E}\left\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right\} \tag{1}$$

where $\mathbb{E}$ is the expectation over the (possibly) random actions $\boldsymbol{a}_1, ..., \boldsymbol{a}_T$ of the agent.

We analyze two widely applied algorithms for the two central special cases of the scenario above:

1. **Bandit Convex Optimization** - $\mathcal{K} \subset \mathbb{R}^n$ is a compact and convex set and $l_t : \mathcal{K} \to [0,1]$ is convex and Lipschitz continuous with parameter L. With no delays, the FKM algorithm, also known as "gradient descent without the gradient" (Flaxman et al., 2005), achieves an expected regret of $O\left(nT^{\frac{3}{4}}\right)$ for this problem.
2. **Multi-armed Bandit** - $\mathcal{K} = \{1, ..., K\}$, $l_t : \{1, ..., K\} \to [0,1]$. With no delays, the EXP3 algorithm (Auer et al., 2002) achieves an expected regret of $O\left(\sqrt{TK\ln K}\right)$ for this problem.

2.1 Results and Contribution

Our main results for the single-agent case are summarized and compared to the literature in Table 1. They are based on the regret bounds proven in Theorem 5 for FKM and Theorem 7 for EXP3 for an unknown T and D.

The regret bound $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$, for $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$ reveals a remarkable robustness of FKM to delayed feedback. Consider the sequence of delays $d_t = t^{\frac{1}{4}}$. Then, the order of magnitude of the regret is still $O\left(nT^{\frac{3}{4}}\right)$, exactly as that of FKM without delays (Flaxman et al., 2005). Even with delays as large as $d_t = t^{\frac{4}{5}}$, the expected regret is $O\left(nT^{\frac{14}{15}}\right)$, so the no-regret property still holds and the learning can be considered successful.

Similarly, the regret bound $O\left(\sqrt{(TK+D)\ln K}\right)$ reveals a significant robustness of EXP3 for delayed feedback. This follows since the T term is factored by K while the delay term D is not. Consider bounded delays of the form $d_t = K$. Then, the order of magnitude of the regret as a function of T and K is $O\left(\sqrt{TK\ln K}\right)$, exactly as that of EXP3 without delays (Auer et al., 1995, 2002). For comparison, consider the full information case where at each round the costs of all arms are received. Assume that the player uses the exponential weights algorithm, which is the equivalent of EXP3 for the full information case. For the same delay sequence $d_t = K$, exponential weights achieves a regret bound of $O\left(\sqrt{TK\ln K}\right)$ (Quanrud and Khashabi, 2015), $\sqrt{K}$ times worse than the $O\left(\sqrt{T\ln K}\right)$ that exponential weights with no delays achieves.

	Convex Optimization		K Experts	
	OGD (Gradient Feedback)	FKM (Bandit Feedback)	Exponential Weights (Full Information)	EXP3 (Bandit Feedback)
No-delay	$O\left(\sqrt{T}\right)$ Zinkevich (2003)	$O\left(nT^{3/4}\right)$ Flaxman et al. (2005)	$O\left(\sqrt{T\ln K}\right)$	$O\left(\sqrt{TK\ln K}\right)$ Auer et al. (1995)
Adversarial Delays	$O\left(\sqrt{D}\right)$ Quanrud and Khashabi (2015)	$O\left(nT^{\frac{3}{4}}+\sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ Theorem 5	$O\left(\sqrt{D\ln K}\right)$ Quanrud and Khashabi (2015)	$O\left(\sqrt{(TK+D)\ln K}\right)$ Theorem 7 Thune et al. (2019); György and Joulani (2020)

Table 1: Expected regret with an adversary (assuming all feedback is received before T, for the ease of comparison of results). For shorthand, we use $D=\sum_{t=1}^{T}d_t$.

Both partial bandit feedback and delays are obstacles that hurt the performance of the learning of the agent, as reflected in the expected regret. Surprisingly, even when the adversary has control over both of these obstacles, the degradation in the regret is mild. Intuitively, with bandit feedback, the effect of delay is much weaker than with full information since less information is delayed. This is an encouraging finding since practical systems typically have both bandit feedback and delays.

Our results for non-cooperative games under delays are summarized in Table 2. They are based on the sufficient conditions for no discounted-regret for FKM (Lemma 4) and EXP3 (Lemma 6). Surprisingly, the delays do not have to be bounded for the convergence to the set of CCE to hold (or to the set of NE for a zero-sum game), and they can even increase as fast as $d_t=O\left(t\log t\right)$. Moreover, the feedback of the players does not need to be synchronized, and they may be subjected to different delay sequences. If $\frac{d_t}{t}\to 0$ as $t\to\infty$ the convergence to the set of CCE follows from the sublinear regret of FKM and EXP3. This is no longer the case for $d_t=t$ or $d_t=t\log t$, where the regret of FKM, EXP3, or any other algorithm is $O\left(T\right)$, so the learning against the adversary fails. Our results show that against other agents the situation is more optimistic, as the discounted ergodic average can still converge to the set of CCE (see Proposition 1 and Proposition 2). To achieve that, the decision-maker needs to choose a time-varying step size η_t, as can be seen in Table 2. In fact, one can go up to $d_t=t\log t\log\left(\log t\right)$ and continue iteratively in this manner, as long as $\sum_{t=1}^{\infty}\frac{1}{d_t}=\infty$. For larger delays, it is not possible to converge to the set of CCE or NE using our approach. Where the utility functions of the game are unknown, no-regret algorithms are appealing since they can be used to approximate the CCE of a non-cooperative game (or the NE of a zero-sum game) by simply simulating it. Our results show that computing a CCE or a NE in this manner is still possible even with large delays that lead to linear (trivial) regret bounds.

3. Non-cooperative Games with Delayed Bandit Feedback

One of the main reasons why adversarial regret bounds are needed is that practical environments consist of multiple interacting agents, leading to non-stationary reward processes. In this section, we study a non-cooperative game where each player only receives delayed bandit feedback, given some arbitrary sequence of delay that can be different for different players.

It is well known that without delays, players that use an online algorithm with sublinear regret (i.e., no-regret) will converge to the set of CCE in the empirical distribution sense (Hannan, 1957; Hart, 2013), and to the set of NE for a zero-sum game (Cai and Daskalakis, 2011). With large

	$d_t \le t^{\frac{1}{4}}$	$d_t \le t^{\frac{3}{4}}$	$d_t \le t$	$d_t \le t \log t$
Parameters for no discounted-regret: FKM	$\eta_t = \frac{1}{t^{\frac{5}{8}} \log(t+1)}$ $\delta = T^{-\frac{3}{16}}$	$\eta_t = \frac{1}{t^{\frac{7}{8}} \log(t+1)}$ $\delta = T^{-\frac{1}{16}}$	$\eta_t = \frac{1}{t \log(t+1)}$ $\delta = (\log \log T)^{-\frac{1}{3}}$	$\eta_t = \frac{1}{t \log(t+1) \log \log(t+1)}$ $\delta = (\log \log \log T)^{-\frac{1}{3}}$
Distance from CCE: FKM	$O\left(\frac{\log T}{T^{\frac{3}{16}}}\right)$	$O\left(\frac{\log T}{T^{\frac{1}{16}}}\right)$	$O\left(\frac{1}{(\log \log T)^{\frac{1}{3}}}\right)$	$O\left(\frac{1}{(\log \log \log T)^{\frac{1}{3}}}\right)$
FKM Regret	$O\left(nT^{\frac{3}{4}}\right)$	$O\left(nT^{\frac{11}{12}}\right)$	$O(T)$	$O(T)$
Parameters for no discounted-regret: EXP3	$\eta_t = \frac{1}{t^{\frac{5}{8}} \log(t+1)}$	$\eta_t = \frac{1}{t^{\frac{7}{8}} \log(t+1)}$	$\eta_t = \frac{1}{t \log(t+1)}$	$\eta_t = \frac{1}{t \log(t+1) \log \log(t+1)}$
Distance from CCE: EXP3	$O\left(\frac{\log T}{T^{\frac{3}{8}}}\right)$	$O\left(\frac{\log T}{T^{\frac{1}{8}}}\right)$	$O\left(\frac{1}{\log \log T}\right)$	$O\left(\frac{1}{\log \log \log T}\right)$
EXP3 Regret	$O\left(T^{\frac{5}{8}}\sqrt{\ln K}\right)$	$O\left(T^{\frac{7}{8}}\sqrt{\ln K}\right)$	$O(T)$	$O(T)$

Table 2: Conditions for no discounted-regret for different delay sequences, along with the corresponding single agent expected regret bounds. For shorthand, we use $D = \sum_{t=1}^{T} d_t$.

enough delays, the regret becomes linear in T. With linear regret, there is no guarantee that the dynamics between the players converge to the set of CCE or NE in any sense. Surprisingly, we show that a CCE (or a NE for a zero-sum game) can still emerge even with linear regret that results from too large delays. Our single-agent regret bounds for FKM and EPX3 provide sufficient conditions under which CCE or a NE can be approximated this way for a convex or finite game, respectively.

Our key observation is that with delayed feedback, it is not the regret that matters for the game dynamics but rather what we call here the discounted-regret. The discounted-regret weights the costs in different rounds according to the step size sequence and therefore coincides with the regret when $\eta_t = \eta, \forall t$. We define the "no discounted-regret" to replace the traditional no-regret concept:

Definition 2. Let $\{l_t\}$ be a sequence of convex cost functions such that $l_t(\boldsymbol{a}) \in [0,1]$. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. We say that an algorithm that produces the random sequence of play $\{\boldsymbol{a}_t\}$ using the step size sequence $\{\eta_t\}$ has no discounted-regret with respect to $\{d_t\}$ if

$$\lim_{T \to \infty} \mathbb{E}\left\{\frac{\sum_{t=1}^{T} \eta_t \left(l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right)}{\sum_{t=1}^{T} \eta_t}\right\} = 0 \tag{2}$$

where the expectation is with respect to the random $\{\boldsymbol{a}_t\}$ generated by the algorithm.

When taking the limit $T \to \infty$, it is important to emphasize that we do not change the infinite sequence of delays $\{d_t\}$, but only reveal more elements in this sequence. In other words, we are looking at the same game but over a longer time horizon. Therefore, while $d_t = \frac{T}{2}$ makes sense for a constant T, it is misleading when taking $T \to \infty$ since it represents a delay that occurred at time t but changes with the limit, so the limit is no longer of the "same game".

The no discounted-regret property is only non-trivial when $\sum_{t=1}^{\infty} \eta_t = \infty$. If $\sum_{t=1}^{\infty} \eta_t < \infty$, then the feedback from the last $\frac{T}{2}$ rounds (for example) can be discarded and the no discounted-regret property will still trivially hold. When $\sum_{t=1}^{\infty} \eta_t = \infty$, all rounds are always important. There is no point after which all feedback can be discarded without causing linear discounted-regret.

3.1 Coarse Correlated Equilibrium for N-player Games

In this subsection we consider a non-cooperative game defined as follows:

Definition 3. In our N-players non-cooperative game, the action set $\mathcal{A}$ of each player is either finite or some convex and compact set $\mathcal{A} \subset \mathbb{R}^n$. The utility function of each player $u_n : \mathcal{A}^N \to [0,1]$ is assumed to be continuous.

The coarse correlated equilibrium (CCE), is a well-established equilibrium concept for learning in games (Hannan, 1957; Ashlagi et al., 2008; Hart, 2013). Our convergence argument will use the notion of an ε-CCE:

Definition 4. Assume that $\mathcal{A} = \{1, \dots, K\}$ or that $\mathcal{A}$ is convex and compact. The set of all ε-CCE points is the set of distributions over $\mathcal{A}^N$ such that:

$$\mathcal{C}_\varepsilon = \left\{ \rho \in \Delta\left(\mathcal{A}^N\right) \middle| \mathbb{E}^{\boldsymbol{a}^* \sim \rho} \left\{ u_n\left(a_n^*, \boldsymbol{a}_{-n}^*\right) \right\} \geq \max_{a_n \in \mathcal{A}} \mathbb{E}^{\boldsymbol{a}^* \sim \rho} \left\{ u_n\left(a_n, \boldsymbol{a}_{-n}^*\right) \right\} - \varepsilon \,, \forall n \right\} \tag{3}$$

and the set of CCE points is $\mathcal{C}_0$ with $\varepsilon = 0$.

We note that in the game of Definition 3 a CCE always exists so $\mathcal{C}_0$ is non-empty (Hart, 2013).

The entity that converges to the set of CCE $\mathcal{C}_0$ in our non-cooperative game scenario is the discounted ergodic distribution of the actions $\boldsymbol{a}_t$. For the special case of $\eta_t = \frac{1}{T}$ for all t, the discounted ergodic distribution of $\boldsymbol{a}_t$ is simply its ergodic distribution of the sequence $\boldsymbol{a}_t$.

Definition 5. For a step sizes sequence $\{\eta_t\}$ and horizon T, the ergodic distribution of a sequence of strategies $\{\boldsymbol{a}_t\}_t$ at strategy profile $\boldsymbol{a}$ is defined as:

$$\rho_T(\boldsymbol{a}) \triangleq \frac{\sum_{t=1}^T \eta_t \mathbb{1}_{\{\boldsymbol{a}_t = \boldsymbol{a}\}}}{\sum_{t=1}^T \eta_t}. \tag{4}$$

Note that $\mathbb{E}\left\{\mathbb{1}_{\{\boldsymbol{a}_t = \boldsymbol{a}\}}\right\} = \mathbb{E}\left\{f_{\boldsymbol{a}_t}(\boldsymbol{a})\right\}$ where $f_{\boldsymbol{a}_t}(\boldsymbol{a})$ is the probability density function of $\boldsymbol{a}_t$ given the information the algorithm has at round t. Hence, we can track $\frac{\sum_{t=1}^T \eta_t f_{\boldsymbol{a}_t}(\boldsymbol{a})}{\sum_{t=1}^T \eta_t}$ instead, which exploits more information.

Then, the following theorem establishes the convergence to the set of CCE $\mathcal{C}_0$.

Theorem 1. *Let N players play a non-cooperative game where the utility function of each player $u_n : \mathcal{A}^N \to [0,1]$ is continuous. Assume that $\mathcal{A}$ is finite or that $\mathcal{A}$ is convex and compact. Let $\{\eta_t\}_{t=1}^\infty$ be the step size sequence for all players. If each player n runs a no discounted-regret algorithm with respect to its delay sequence $\{d_t^n\}_t$ then $\mathbb{E}\{\rho_T\}$ converges to the set of CCE $\mathcal{C}_0$ as $T \to \infty$.*

Proof See Appendix. ■

The expectation $\mathbb{E}\{\rho_T\}$ is taken with respect to the random actions. By definition, the set of ε-CCE is convex, so the average of multiple ε-CCE is also an ε-CCE. Hence, the implication of Theorem 1 is that to approximate a CCE using T samples, one can run $\sqrt{T}$ independent simulations of the game and then average the resulting $\left\{\rho_{\sqrt{T}}^{(i)}\right\}_{i=1}^{\sqrt{T}}$. From the strong law of large numbers, this estimation converges as $T \to \infty$ with probability 1 to $\mathcal{C}_0$ since $\rho_T(\boldsymbol{a})$ is bounded and $\left\{\rho_{\sqrt{T}}^{(i)}\right\}_{i=1}^{\sqrt{T}}$ are independent.

The general result of Theorem 1 implies stronger results for special classes of games where the set of CCE has an interesting structure. For example, a polymatrix game is a finite action set game where each player plays a separate zero-sum game against each of her neighbors on a given graph. For polymatrix games, for which zero-sum game is a special case, it was shown in Cai and Daskalakis (2011) that the marginal distributions of the CCE are a Nash equilibrium. However, this result holds only for our multi-armed bandit case since it assumes discrete action sets. Our next section establishes that two no discounted-regret algorithms in a zero-sum game lead to Nash equilibrium **both** for the multi-armed bandit case and the bandit convex optimization case.

3.2 Nash Equilibrium for Two-Player Zero-Sum Games

In this subsection, we focus on two-player zero-sum games. In a two-player zero-sum game, when the row player plays $\boldsymbol{y}$ and the column player plays $\boldsymbol{z}$, the first pays a cost of $u(\boldsymbol{y},\boldsymbol{z})$ and the second gains a reward of $u(\boldsymbol{y},\boldsymbol{z})$, for a cost function u defined as follows:

Definition 6. Let $U:\mathcal{K}\times\mathcal{K}\to[0,1]$ be the cost function of the zero-sum game, restricted to pure actions. Then:

1. When $\mathcal{K}=\{1,\ldots,K\}$ we define $\mathcal{A}=\Delta^K$ (K-dimensional simplex) and for every $\boldsymbol{y},\boldsymbol{z}\in\mathcal{A}$

$$u(\boldsymbol{y},\boldsymbol{z})\triangleq\sum_{i=1}^{K}\sum_{j=1}^{K}y^{(i)}z^{(j)}U(i,j). \tag{5}$$

2. When $\mathcal{K}\subset\mathbb{R}^n$ is convex and compact we define $\mathcal{A}=\mathcal{K}$ and $u(\boldsymbol{y},\boldsymbol{z})=U(\boldsymbol{y},\boldsymbol{z})$. We assume that $U(\boldsymbol{y},\boldsymbol{z})$ is continuous, convex in $\boldsymbol{y}$ and concave in $\boldsymbol{z}$.

For zero-sum games, we show that no discounted-regret algorithms lead to the set of Nash equilibria, which is a more exclusive solution concept than CCE. In fact, if a Nash equilibrium (NE) exists, it is always a CCE (Hart, 2013). A NE is an action profile such that no player wants to switch an action given that the other players keep their actions. For our convergence argument, we define the set of all approximate (pure) NE of a zero-sum game:

Definition 7. The set of all ε-NE points of a two-player zero-sum game is

$$\mathcal{N}_\varepsilon=\left\{(\boldsymbol{y}^*,\boldsymbol{z}^*)\in\mathcal{A}\times\mathcal{A}\;\middle|\;u(\boldsymbol{y}^*,\boldsymbol{z}^*)\le\min_{\boldsymbol{y}\in\mathcal{A}}u(\boldsymbol{y},\boldsymbol{z}^*)+\varepsilon,u(\boldsymbol{y}^*,\boldsymbol{z}^*)\ge\max_{\boldsymbol{z}\in\mathcal{A}}u(\boldsymbol{y}^*,\boldsymbol{z})-\varepsilon\right\} \tag{6}$$

and the set of NE points is $\mathcal{N}_0$ with $\varepsilon=0$.

For the game in Definition 6, a pure NE always exists (Nikaidô et al., 1955) so $\mathcal{N}_0$ is non-empty.

It was shown in Bailey and Piliouras (2018) that for the no-delay case, the last iterate $\boldsymbol{a}_t$ does not converge in general to a NE and even moves away from it. Instead, it is the ergodic average strategy that converges to the set of NE $\mathcal{N}_0$. With delayed feedback, the entity that converges to $\mathcal{N}_0$ in our zero-sum game scenario is the discounted ergodic average of the strategies $\{\boldsymbol{a}_t\}_t$. For the special case of $\eta_t=\frac{1}{T}$ for all t, the discounted ergodic average of $\boldsymbol{a}_t$ is just its ergodic average.

Definition 8. For a step sizes sequence $\{\eta_t\}$ and horizon T, the discounted ergodic average of a sequence of strategies $\{\boldsymbol{a}_t\}_t$ is defined as

$$\bar{\boldsymbol{a}}_T\triangleq\frac{\sum_{t=1}^{T}\eta_t\boldsymbol{a}_t}{\sum_{t=1}^{T}\eta_t}. \tag{7}$$

Then, the following theorem establishes the convergence to the set of NE $\mathcal{N}_0$.

Theorem 2. *Let two players play a zero-sum game with the cost function $u(\boldsymbol{y},\boldsymbol{z}):\mathcal{A}\times\mathcal{A}\to[0,1]$. Assume that $u(\boldsymbol{y},\boldsymbol{z})$ is convex in $\boldsymbol{y}$ and concave in $\boldsymbol{z}$ and is continuous. Let $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$ be the actions of the row and column players at round t, and let $\bar{\boldsymbol{y}}_T$ and $\bar{\boldsymbol{z}}_T$ be their discounted ergodic averages. Let $\{\eta_t\}_{t=1}^{\infty}$ be the step size sequence of both players. Let $\{d_t^r\}$ and $\{d_t^c\}$ be the delay sequence of the row player and the column player. If both players use a no discounted regret algorithm with respect to $\{d_t^r\},\{d_t^c\}$ then , as $T\to\infty$:*

1. *$(\bar{\boldsymbol{y}}_T,\bar{\boldsymbol{z}}_T)$ converges in L^1 to the set of Nash equilibria $\mathcal{N}_0$ of the zero-sum game:*
2. *$U(\bar{\boldsymbol{y}}_T,\bar{\boldsymbol{z}}_T)$ converges in L^1 to the value of the game $\min_{\boldsymbol{y}}\max_{\boldsymbol{z}}U(\boldsymbol{y},\boldsymbol{z})=\max_{\boldsymbol{z}}\min_{\boldsymbol{y}}U(\boldsymbol{y},\boldsymbol{z})$.*

Proof See Appendix. ■

4. Doubling Trick for Online Learning with Delays

Online learning under delayed feedback introduces another key parameter, which is the sum of delays $D = \sum_{t=1}^{T} d_t$. The sum of delays appears in the expected regret bound of many online algorithms and is required to tune their step size and other parameters. If D or a tight upper bound for it is not known, then an adaptive algorithm is needed.

With no delays, the standard doubling trick (see Cesa-Bianchi et al. (1997)) can be used if T is unknown. However, the same doubling trick does not work with delayed feedback. We now present a novel doubling trick for the delayed feedback case, where T and D are unknown. Compared to the doubling trick presented in Bistritz et al. (2019), our enhanced doubling trick is two-dimensional, as each epoch is indexed by a delay index as well as a time index.

We divide the time horizon into super-epochs, indexed by ν. A super-epoch $\mathcal{E}_\nu$ is a set of consecutive rounds that all use the same parameters $\mathcal{P}_\nu$ (e.g., a step size η_ν), as defined in Algorithm 1. Let ν_t be the index of the super-epoch that contains round t. Let m_t be the number of missing feedback samples at round t. A missing feedback sample at round t is a sample from $\tau \leq t$ such that $\nu_t = \nu_\tau$ that was not received before the beginning of round t. We increase the delay index w every time when $\sum_{\tau=1}^{t} m_\tau$, that tracks D, doubles. We increase the time index h every time when the number of rounds t doubles. Define the (w, h) epoch as

$$\mathcal{T}_{w,h} = \left\{ t \,|\, 2^{w-1} \leq \sum_{\tau=1}^{t} m_\tau < 2^w,\, 2^{h-1} \leq t < 2^h \right\} \tag{8}$$

which is the set of consecutive rounds where the sum of delays is within a given interval and the time index is within another given interval. During the $(w, h) \in \mathcal{E}_\nu$ epoch, the algorithm equipped with our doubling trick uses the parameters $\mathcal{P}_\nu$ (e.g., $\mathcal{P}_\nu = \{\eta_\nu, \delta_\nu\}$ for FKM). We emphasize that different epochs (w_1, h_1) and (w_2, h_2) in the same super-epoch use the same set of parameters. As shown in Fig. 1, this can only be the case if $h_1 = h_2$ or $w_1 = w_2$. Feedback samples originated in a different super-epoch are discarded once received, and are no longer counted in $\sum_{\tau=1}^{t} m_\tau$ after their super-epoch has ended. The resulting algorithm is detailed in Algorithm 1.

Fig. 1 illustrates our doubling trick that assumes a regret bound of the form $O\left(T^a D^b + T^c + D^d\right)$ is available if T and D are known. We can see that the (w, h) is split into three regions with three different types of super-epochs. In the upper one (in blue) D^d dominates the regret, in the middle one (in pink) $T^a D^b$ dominates the regret and in the lower one (in orange) T^c dominates the regret. Each blue, pink, or orange box is a super-epoch on which we apply our regret bound separately. The grey arrows represent the actual path that (w, h) went through.

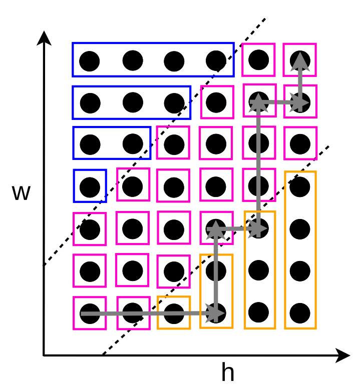

Figure 1: 2D Doubling Trick. Each box is a different super-epoch, and each dot is a different epoch.

Algorithm 1 Adaptive Wrapper for Unknown T and D

Initialization: Set $w=0,h=0,\nu=0$. Choose an algorithm $\text{Alg}\left(\mathcal{P}_{\nu}\right)$ and initialize $\boldsymbol{p}_1,\mathcal{P}_0$. Let $\mathcal{P}^*\left(T,D\right)$ be the parameters that are used when D,T are known, and let $R\left(T,D,\mathcal{P}^*\left(T,D\right)\right)=k_1D^d+k_2T^c+k_3T^aD^b$ be the resulting regret bound. Divide the (w,h) space into super-epochs:

- For each h, the super-epoch $\mathcal{E}=\mathcal{S}_h$ is the set of (w,h) such that $2k_2 2^{ch}\geq k_1 2^{dw}+k_3 2^{ah}2^{bw}$.
- For each w, the super-epoch $\mathcal{E}=\mathcal{S}_w$ is the set of (w,h) such that $2k_1 2^{dw}\geq k_2 2^{ch}+k_3 2^{ah}2^{bw}$.
- All other epochs $(w,h)\notin\mathcal{S}_w\cup\mathcal{S}_h$ are each a separate super-epoch $\mathcal{E}=\{(w,h)\}$.

For $t=1,...,T$ **do**

1. Pick an action $\boldsymbol{a}_t\in\mathcal{K}$ at random according to the distribution $\boldsymbol{p}_t$.
2. Let $\tilde{\mathcal{C}}_t$ be the set of delayed costs $l_s\left(\boldsymbol{a}_s\right)$ received at round t such that $\nu_s=\nu_t$. Update the number of missing samples so far
$$\left(t-\min_{\tau\in\{t'\mid\nu_t=\nu_{t'}\}}\tau+1\right)-\sum_{\tau\in\{t'\mid\nu_t=\nu_{t'}\}}\left|\tilde{\mathcal{C}}_\tau\right|. \tag{9}$$
3. If $\sum_{\tau=1}^{t}m_\tau\geq 2^w$, then update $w\leftarrow w+1$. If $t\geq 2^h$, then update $h\leftarrow h+1$.
4. If $(w,h)\notin\mathcal{E}_\nu$ then start a new super-epoch with $\mathcal{P}_{\nu+1}\leftarrow\mathcal{P}^*\left(2^{w_\nu},2^{h_\nu}\right)$ where w_ν,h_ν are the maximal w,h indices in super-epoch $\mathcal{E}_\nu$, initialize $\text{Alg}\left(\mathcal{P}_{\nu+1}\right)$, and set $\nu\leftarrow\nu+1$.
5. Using only the samples in $\tilde{\mathcal{C}}_t$, update the distribution $\boldsymbol{p}_t$ according to $\text{Alg}\left(\mathcal{P}_\nu\right)$.

End

The next Lemma proves that our doubling trick tracks D, similar in spirit to how the standard doubling trick tracks T up to a factor of 2. It also bounds the largest delay and time indices possible.

Lemma 1. *Let H be the last time index and let W be the last delay index. Define $\mathcal{S}_w,\mathcal{S}_h$ as in Algorithm 1. Let $\mathcal{T}_w$ be the set of all rounds in $\mathcal{S}_w$ and $\tau_w\triangleq\max_{t\in\mathcal{T}_w}t$. Let W_h be the maximal w such that $(w,h)\in\mathcal{S}_h$. Let $\mathcal{T}_h$ be the set of all rounds in $\mathcal{S}_h$ and $\tau_h\triangleq\max_{t\in\mathcal{T}_h}t$. Algorithm 1 maintains:*

1. *For every w and h, $\sum_{t\in\mathcal{T}_w}\min\left\{d_t,\tau_w-t+1\right\}\leq 2^{w-1}$ and $\sum_{t\in\mathcal{T}_h}\min\left\{d_t,\tau_h-t+1\right\}\leq 2^{W_h}$.*
2. *$W\leq\log_2\sum_{t=1}^{T}\min\left\{d_t,T-t+1\right\}+1$ and $H\leq\log_2\left(T+2\right)-1$.*

Proof Let $\mathcal{M}_{\mathcal{S}_w}$ be the set of feedback samples for costs in $\mathcal{S}_w$ that are not received within $\mathcal{S}_w$. Every round $t\in\mathcal{T}_w$ such that $t\notin\mathcal{M}_{\mathcal{S}_w}$ contributes exactly d_t to $\sum_{t\in\mathcal{T}_w}m_t$, since the t-th feedback is missing for d_t rounds in $\mathcal{T}_w$. Every round $t\in\mathcal{M}_{\mathcal{S}_w}$ contributes $\tau_w-t+1\leq d_t$ to $\sum_{t\in\mathcal{T}_w}m_t$ before it stops being counted. Therefore

$$\sum_{t\in\mathcal{T}_w}\min\left\{d_t,\tau_w-t+1\right\}\leq\sum_{t\in\mathcal{T}_w}m_t\underset{(a)}{\leq}2^{w-1} \tag{10}$$

where (a) follows since if $\sum_{t\in\mathcal{T}_w}m_t>2^{w-1}$ then $\sum_{\tau=1}^{t}m_\tau\geq 2^{w-1}+2^{w-1}=2^w$ and the w should have already increased to $w+1$. Applying the same argument on $\mathcal{T}_h$, we obtain that we must have

$$\sum_{t\in\mathcal{T}_h}\min\left\{d_t,\tau_h-t+1\right\}\leq\sum_{t\in\mathcal{T}_h}m_t\underset{(a)}{\leq}2^{W_h} \tag{11}$$

where (a) follows since if $\sum_{t\in\mathcal{T}_h} m_t > 2^{W_h}$ then $\sum_{\tau=1}^{t} m_\tau \geq 2^{W_h}$ so w must have increased to W_h+1.

For the second part of the lemma, $2^{H+1}-2 = \sum_{h=1}^{H} 2^h \leq T$ so $H \leq \log_2(T+2)-1$, and

$$\sum_{t=1}^{T} \min\{d_t, T-t+1\} \underset{(a)}{\geq} \sum_{t=1}^{T} m_t \geq 2^{W-1} \tag{12}$$

where (a) uses that every sample is counted in $\sum_{t=1}^{T} m_t$ for at most d_t or $T-t+1$ rounds. ■

Now we can prove our main result of this section. The first assumption is merely the regret bound that one can obtain if T and D are known. The second assumption says that holding the parameters $\mathcal{P}$ fixed, the regret bound is non-decreasing with T and D. As we see in Section 5 and Section 6, these basic assumptions hold for FKM and EXP3.

Theorem 3. *Let $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$. Let $R(T, D, \mathcal{P}(T, D))$ be an upper bound on the expected regret of an online learning algorithm that uses the parameters $\mathcal{P}(T, D)$. Assume:*

1. *There exists a sequence $\mathcal{P}^*(T, D)$ and constants $k_1, k_2, k_3 \geq 0$ and $0 \leq a, b, c, d \leq 1$ such that for all T, D:*
$$R(T, D, \mathcal{P}^*(T, D)) \leq k_1 D^d + k_2 T^c + k_3 T^a D^b. \tag{13}$$
2. *For a fixed $\mathcal{P}^*$, $R(T, D, \mathcal{P}^*)$ is non-decreasing with T and D.*

Then if Algorithm 1 is used to track $\mathcal{P}^*(T, D)$, it achieves a total expected regret of

$$R(T, D) = O\left(k_1 D^d + k_2 T^c + k_3 T^a D^b\right). \tag{14}$$

Proof We apply the regret bound for each of the super-epoch types $\mathcal{S}_w, \mathcal{S}_h$ and $\{(w, h)\}$ defined in Algorithm 1 as follows (also illustrated Fig. 1):

Type 1 ($\mathcal{S}_h$): Let W_h be the largest w such that $(w, h) \in \mathcal{S}_h$. Let T_h be the length of $\mathcal{S}_h$ and let $D_h \triangleq \sum_{t\in\mathcal{T}_h} \min\{d_t, \tau_h - t + 1\} \leq 2^{W_h}$, using Lemma 1. By applying the regret bound on $\mathcal{S}_h$:

$$R_{\mathcal{S}_h} \underset{(a)}{\leq} R\left(T_h, D_h, \mathcal{P}^*\left(2^h, 2^{W_h}\right)\right) \underset{(b)}{\leq} R\left(2^h, 2^{W_h}, \mathcal{P}^*\left(2^h, 2^{W_h}\right)\right) \leq k_1 2^{W_h d} + k_2 2^{hc} + k_3 2^{ha} 2^{W_h b} \leq 3k_2 2^{hc} \tag{15}$$

where (a) follows since Algorithm 1 uses $\mathcal{P}^*\left(2^h, 2^{W_h}\right)$, and (b) from condition 2 of the Theorem.

Type 2 ($\mathcal{S}_w$): Let H_w be the largest h such that $(w, h) \in \mathcal{S}_w$. Let T_w be the length of $\mathcal{S}_w$ and let $D_w \triangleq \sum_{t\in\mathcal{T}_w} \min\{d_t, \tau_w - t + 1\} \leq 2^{w-1}$, using Lemma 1. By applying the regret bound on $\mathcal{S}_w$:

$$R_{\mathcal{S}_w} \leq R\left(T_w, D_w, \mathcal{P}^*\left(2^{H_w}, 2^w\right)\right) \leq R\left(2^{H_w}, 2^w, \mathcal{P}^*\left(2^{H_w}, 2^w\right)\right) \leq k_1 2^{wd} + k_2 2^{H_w c} + k_3 2^{H_w a} 2^{wb} \leq 3k_1 2^{wd}. \tag{16}$$

Type 3 ($\{(w, h)\}$): For this case, we must have $2k_3 2^{ha} 2^{wb} \geq k_1 2^{wd} + k_2 2^{ch}$, so

$$R_{w,h} \leq R\left(T_{w,h}, D_{w,h}, \mathcal{P}^*\left(2^h, 2^w\right)\right) \leq R\left(2^h, 2^w, \mathcal{P}^*\left(2^h, 2^w\right)\right) \leq k_1 2^{wd} + k_2 2^{hc} + k_3 2^{ha} 2^{wb} \leq 3k_3 2^{ha} 2^{wb}. \tag{17}$$

Then, for the constant $C_0 = 3\max\left\{\frac{2^{2d}}{2^d-1}, \frac{2^{2c}}{2^c-1}, \frac{2\cdot 2^{2a} 2^{2b}}{(2^a-1)(2^b-1)}\right\}$, the total regret is bounded by

$$\begin{aligned}\mathbb{E}\{R(T)\} &= \sum_{h=1}^{H} R_{\mathcal{S}_h} + \sum_{w=1}^{W} R_{\mathcal{S}_w} + \sum_{(w,h)\notin\mathcal{S}_w\cup\mathcal{S}_h} R_{w,h} \\ &\underset{(a)}{\leq} 3k_1 \sum_{w=1}^{W} 2^{dw} + 3k_2 \sum_{h=1}^{H} 2^{ch} + 3k_3 \sum_{h=1}^{H} 2^{ah} \sum_{w=1}^{W} 2^{bw} \underset{(b)}{\leq} C_0\left(k_1 D^d + k_2 T^c + k_3 T^a D^b\right)\end{aligned} \tag{18}$$

where (a) uses the bounds in (15),(16),(17) even for epochs that do not occur (that trivially have zero regret). Inequality (b) uses part 2 of Lemma 1 to upper bound H and W. ■

5. The FKM Algorithm for Adversarial Bandit Convex Optimization with Delayed Feedback

In bandit convex optimization, the action $\boldsymbol{a}_t$ is chosen from a convex and compact set $\mathcal{K} \subset \mathbb{R}^n$ with diameter $|\mathcal{K}| \triangleq \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}} \|\boldsymbol{x}-\boldsymbol{y}\|$. Without the loss of generality, we assume that $\mathcal{K}$ contains the unit ball centered at the zero vector. The cost functions $l_t(\boldsymbol{a}_t) \in [0,1]$ are convex for all t and Lipschitz continuous with parameter L. The player has no access to the gradient of l_t, and only receives the value of $l_t(\boldsymbol{a}_t)$ at round $t+d_t-1$. In the FKM algorithm, the player uses the cost value to estimate the gradient. The idea is to use a perturbation $\boldsymbol{u}_t$, drawn uniformly at random on the n-dimensional unit sphere $\mathbb{S}_1$. Then, instead of playing the unperturbed action $\boldsymbol{x}_t$, the player plays $\boldsymbol{a}_t = \boldsymbol{x}_t + \delta\boldsymbol{u}_t$ for a sampling radius $\delta > 0$. Let $\mathcal{K}_\delta = \left\{\boldsymbol{x} \mid \frac{1}{1-\delta}\boldsymbol{x} \in \mathcal{K}\right\}$. To ensure that $\boldsymbol{x}_t + \delta\boldsymbol{u}_t \in \mathcal{K}$, we maintain $\boldsymbol{x}_t \in \mathcal{K}_\delta$ by projecting into $\mathcal{K}_\delta$ after each gradient step. Define the following filtration

$$\mathcal{F}_t = \sigma\left(\{\boldsymbol{x}_\tau \mid \tau \le t\}\right) \tag{19}$$

which is generated from all the past unperturbed actions. With a slight abuse of notation, we use $\mathcal{F}_{s_-}$ to denote the filtration induced from all the $\boldsymbol{x}_q$ such that $q \in \mathcal{S}_r$ for $r < t$ and all $q \in \mathcal{S}_t$ that were already used before the algorithm uses the feedback from time s, but including $\boldsymbol{x}_s$ itself.

The purpose of the action perturbation is to allow for an estimator for the gradient in $\boldsymbol{a}_t$ with a bias that is controlled by δ:

Lemma 2 (Flaxman et al. (2005, Lemma 2.1))**.** *Let $\delta > 0$ and define $\hat{l}(\boldsymbol{x}) \triangleq \mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{l(\boldsymbol{x}+\delta\boldsymbol{u})\}$ where $\mathbb{S}_1$ is the unit sphere. Let $\boldsymbol{g} = \frac{n}{\delta} l(\boldsymbol{x}+\delta\boldsymbol{u})\boldsymbol{u}$. Then $\mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{\boldsymbol{g}\} = \nabla\hat{l}(\boldsymbol{x})$.*

The next Lemma is the main result of this section, used to prove both Theorem 4 and Lemma 4.

Lemma 3. *Let $\{\eta_t\}$ be a non-increasing step size sequence. Let δ be the sampling radius. For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{K}} \sum_{t=1}^{T} \eta_t l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Define the set $\mathcal{M}$ of all samples that are not received before round T. Then*

1. *For an oblivious adversary:*

$$\sum_{t=1}^{T} \eta_t\left(\mathbb{E}\{l_t(\boldsymbol{a}_t)\} - l_t(\boldsymbol{a}^*)\right) \le \sum_{t\in\mathcal{M}} \eta_t + \frac{|\mathcal{K}|^2}{2} + (3+|\mathcal{K}|)L\delta \sum_{t\notin\mathcal{M}} \eta_t + \frac{1}{2}\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}} \eta_t^2 + 2L\frac{n}{\delta}\sum_{t\notin\mathcal{M}} \eta_t^2 d_t. \tag{20}$$

2. *For an adaptive adversary:*

$$\sum_{t=1}^{T} \eta_t \mathbb{E}\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\} \le$$
$$\sum_{t\in\mathcal{M}} \eta_t + \frac{|\mathcal{K}|^2}{2} + (3+|\mathcal{K}|)L\delta\sum_{t\notin\mathcal{M}} \eta_t + |\mathcal{K}|\sqrt{2\sum_{t\notin\mathcal{M}} \eta_t^2\left(\frac{n^2}{\delta^2} + L^2\right)} + \frac{1}{2}\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}} \eta_t^2(1+4d_t). \tag{21}$$

Proof See Appendix. ■

The following theorem establishes the expected regret bound for FKM with delays. It is proved by optimizing over a constant step size η and sampling radius δ in Lemma 3.

Algorithm 2 FKM with delays

Initialization: Let $\{\eta_t\}$ be positive non-increasing sequences. Let $\delta < 1$. Set $\boldsymbol{x}_1 = 0$.
For $t = 1, ..., T$ **do**

1. Draw $\boldsymbol{u}_t \in \mathbb{S}_1$ uniformly at random.
2. Play $\boldsymbol{a}_t = \boldsymbol{x}_t + \delta \boldsymbol{u}_t$.
3. Obtain a set of delayed costs $l_s(\boldsymbol{a}_s)$ for all $s \in \mathcal{S}_t$ and compute $\boldsymbol{g}_s = \frac{n}{\delta} l_s(\boldsymbol{a}_s) \boldsymbol{u}_s$.
4. Let $s_{\min} = \min\limits_{s \in \mathcal{S}_t} s$ and $s_{\max} = \max\limits_{s \in \mathcal{S}_t} s$. Set $\boldsymbol{x}_{s_{\min}^-} = \boldsymbol{x}_t$. For every $s \in \mathcal{S}_t$, update

$$\boldsymbol{x}_{s_+} = \prod_{\mathcal{K}_\delta} \left(\boldsymbol{x}_{s_-} - \eta_s \boldsymbol{g}_s\right) \tag{22}$$

where $\mathcal{K}_\delta = \left\{\boldsymbol{x} \mid \frac{1}{1-\delta}\boldsymbol{x} \in \mathcal{K}\right\}$, and then set $\boldsymbol{x}_{t+1} = \boldsymbol{x}_{s_{\max}^+}$.

End

Theorem 4. *Let $\eta > 0$ and $0 < \delta < 1$. For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min\limits_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Define the set $\mathcal{M}$ of all samples that are not received before round T. Then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \sum_{t=1}^{T} \mathbb{E}\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\} \leq$$
$$|\mathcal{M}| + \left((3 + |\mathcal{K}|)\delta L + \frac{1}{2}\eta\frac{n^2}{\delta^2}\right)(T - |\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta} + 2Ln\frac{\eta}{\delta}\sum_{t \notin \mathcal{M}} d_t. \tag{23}$$

Furthermore, for

$$\eta = |\mathcal{K}| \min\left\{\frac{1}{n}T^{-\frac{3}{4}}, \frac{1}{\sqrt{n}}T^{-\frac{1}{3}}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{-\frac{1}{3}}\right\} \text{ and } \delta = \max\left\{T^{-\frac{1}{4}}, T^{-\frac{2}{3}}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}}\right\} \tag{24}$$

we obtain

$$\mathbb{E}\{R(T)\} = O\left(nT^{\frac{3}{4}} + \sqrt{n}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}} + |\mathcal{M}|\right). \tag{25}$$

Proof First note that if $|\mathcal{M}| = \Theta(T)$ then $\mathbb{E}\{R(T)\} = O(T)$ and otherwise $T - |\mathcal{M}| = \Theta(T)$. To obtain (23), substitute $\eta_t = \eta$ Lemma 3 for an oblivious adversary and divide both sides by η. We have $\Theta(\delta T) = \Theta\left(\frac{\eta}{\delta^2}T\right) = \Theta\left(\frac{1}{\eta}\right)$ for $\eta = \frac{|\mathcal{K}|}{n}T^{-\frac{3}{4}}$ and $\delta = T^{-\frac{1}{4}}$, hence

$$\min_{\delta,\eta}\left((3 + |\mathcal{K}|)\delta L(T - |\mathcal{M}|) + \frac{1}{2}\eta\frac{n^2}{\delta^2}(T - |\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta}\right) = O\left(nT^{\frac{3}{4}}\right). \tag{26}$$

Since $\Theta(\delta T) = \Theta\left(\frac{\eta}{\delta}\sum_{t \notin \mathcal{M}} d_t\right) = \Theta\left(\frac{1}{\eta}\right)$ for $\eta = \frac{1}{\sqrt{n}}\left(\frac{1}{T\sum_{t \notin \mathcal{M}} d_t}\right)^{\frac{1}{3}}$ and $\delta = T^{-\frac{2}{3}}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}}$, then

$$\min_{\delta,\eta}\left((3 + |\mathcal{K}|)\delta L(T - |\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta} + 2L\eta\frac{n}{\delta}\sum_{t \notin \mathcal{M}} d_t\right) = O\left(\sqrt{n}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}}\right). \tag{27}$$

Therefore (23) cannot have a better T dependence than in (25) for any η, δ. For the choice in (24) we have

$$n^2 \frac{\eta}{\delta^2} T = \frac{|\mathcal{K}| \min\left\{ nT^{\frac{1}{4}}, n^{\frac{3}{2}} T^{\frac{2}{3}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{-\frac{1}{3}} \right\}}{\max\left\{ T^{-\frac{1}{2}}, T^{-\frac{4}{3}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{2}{3}} \right\}} \underset{(a)}{\leq} O\left(nT^{\frac{3}{4}} \right) \tag{28}$$

where (a) follows since $\frac{\min\{a,b\}}{\max\{c,d\}} \leq \frac{a}{c}$. For the choice in (24) we also have

$$\frac{\eta}{\delta} n \sum_{t \notin \mathcal{M}} d_t = \frac{|\mathcal{K}| \min\left\{ T^{-\frac{3}{4}} \sum_{t \notin \mathcal{M}} d_t, \sqrt{n} T^{-\frac{1}{3}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{2}{3}} \right\}}{\max\left\{ T^{-\frac{1}{4}}, T^{-\frac{2}{3}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} \right\}} \underset{(a)}{\leq} O\left(\sqrt{n} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} T^{\frac{1}{3}} \right) \tag{29}$$

where (a) follows since $\frac{\min\{a,b\}}{\max\{c,d\}} \leq \frac{b}{d}$. Therefore, for the choice in (24)

$$\sum_{t=1}^{T} \left(\mathbb{E}\left\{ l_t\left(\boldsymbol{a}_t \right) \right\} - l_t\left(\boldsymbol{a}^* \right) \right) \leq |\mathcal{M}| + \left((3 + |\mathcal{K}|) L + \frac{|\mathcal{K}|^2}{2} \right) \max\left\{ nT^{\frac{3}{4}}, \sqrt{n} T^{\frac{1}{3}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} \right\}$$
$$+ O\left(nT^{\frac{3}{4}} \right) + O\left(\sqrt{n} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} T^{\frac{1}{3}} \right) = O\left(nT^{\frac{3}{4}} + \sqrt{n} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} T^{\frac{1}{3}} + |\mathcal{M}| \right). \tag{30}$$

■

The next Corollary shows that the bound of Theorem 4 is slightly tighter than $O\left(nT^{\frac{3}{4}} + \sqrt{n} D^{\frac{1}{3}} T^{\frac{1}{3}} \right)$, where one sums $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$ instead of $\sum_{t \notin \mathcal{M}} d_t$ in our regret bound (25), depending on the pattern of the missing samples.

Corollary 1. *Choose the fixed η and δ according to (24). For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Let $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$. Then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \sum_{t=1}^{T} \left(\mathbb{E}\left\{ l_t\left(\boldsymbol{a}_t \right) \right\} - l_t\left(\boldsymbol{a}^* \right) \right) = O\left(nT^{\frac{3}{4}} + \sqrt{n} D^{\frac{1}{3}} T^{\frac{1}{3}} \right). \tag{31}$$

Proof The $m \triangleq |\mathcal{M}|$ missing samples contribute at least $\frac{m(m+1)}{2} \geq \frac{m^2}{2}$ to $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$. This follows since the best case is when the feedback of round T is delayed by one and arrives after T, the feedback of round $T - 1$ now has to be delayed by at least 2 to arrive after T and so on, m times. Hence, if $D \leq T^{\frac{5}{4}}$ then we must have that $m \leq \sqrt{2} T^{\frac{5}{8}}$ and therefore

$$(T - m)^{\frac{3}{4}} + m \leq T^{\frac{3}{4}} + \sqrt{2} T^{\frac{5}{8}} \leq 3T^{\frac{3}{4}}. \tag{32}$$

On the other hand, we always have

$$T^{\frac{1}{3}} D^{\frac{1}{3}} \geq \left(T \sum_{t \notin \mathcal{M}} d_t + T \frac{m^2}{2} \right)^{\frac{1}{3}} \underset{(a)}{\geq} \frac{1}{2} \left(2T \sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} + \frac{1}{2} \left(Tm^2 \right)^{\frac{1}{3}} \geq \frac{T^{\frac{1}{3}}}{2^{\frac{2}{3}}} \left(\sum_{t \notin \mathcal{M}} d_t \right)^{\frac{1}{3}} + \frac{m}{2} \tag{33}$$

where (a) follows from the concavity of $f(x) = x^{\frac{1}{3}}$. We conclude that in any case ($D \leq T^{\frac{5}{4}}$ or not), the regret in (31) is greater than that in (25). ■

5.1 Agnostic FKM

If the horizon T and sum of delays D are unknown, then we can apply Algorithm 1 to wrap FKM. The next Theorem is an immediate application Theorem 3 on the FKM regret bound for this agnostic case. The resulting bound retains the same order of magnitude as the bound of Corollary 1 even though D and T are unknown. The only difference with the bound of Theorem 4 arises because the doubling trick discards samples that cross super-epochs. Hence, the bound below uses $D=\sum_{t=1}^{T}\min\{d_t,T-t+1\}$ instead of $\sum_{t\notin\mathcal{M}}d_t$ and $|\mathcal{M}|$.

Theorem 5. *For every t, let $l_t:\mathcal{K}\to[0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let $D=\sum_{t=1}^{T}\min\{d_t,T-t+1\}$. If the player uses Algorithm 1 to wrap FKM (Algorithm 2) with $\eta_{w,h}=|\mathcal{K}|\min\left\{\frac{1}{n}2^{-\frac{3h}{4}},\frac{1}{\sqrt{n}}2^{-\frac{h+w}{3}}\right\}$ and $\delta_{w,h}=\delta_0\max\left\{2^{-\frac{h}{4}},2^{\frac{w-2h}{3}}\right\}$ for $\delta_0<1$, then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\}=O\left(nT^{\frac{3}{4}}+\sqrt{n}D^{\frac{1}{3}}T^{\frac{1}{3}}\right). \tag{34}$$

Proof Consider the regret bound in (23) after bounding $|\mathcal{M}|\le\sqrt{2D}$ (as in Corollary 1), $T-|\mathcal{M}|\le T$ and $\sum_{t\notin\mathcal{M}}d_t\le D$. For the choice in (24), this bound yields the bound of Corollary 1 which is of the form $nT^{\frac{3}{4}}+\sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}$, so it matches Assumption 1 in Theorem 3 with $a=b=\frac{1}{3}, c=\frac{3}{4}$ and $d=0$. This bound also satisfies Assumption 2 since it is increasing with T and D for any η,δ. ■

5.2 No Discounted-Regret Property for FKM

In this subsection, we provide conditions for FKM to have no discounted regret with respect to the delay sequence. As discussed in Section 3, $\sum_{t=1}^{\infty}\eta_t=\infty$ is necessary for the no discounted-regret property to be non-trivial. All other conditions of Lemma 4 are as tight as the bound of Lemma 3.

Note that one can choose $\eta_t=\frac{1}{t\log t\log\log t\log\log\log t}$ and $\delta_T=O\left((\log\log(\log T))^{-\frac{1}{3}}\right)$ to guarantee no discounted-regret for all sequences such that $d_t=O(t\log t\log\log t)$. This boundary can only be slightly improved by adding $\log(\log(\ldots\log(T)))$ iteratively in this manner as long as $\sum_{t=1}^{T}\frac{1}{d_t}=\infty$. Hence, no knowledge of the actual sequence d_t is required to tune η_t and δ_T. However, if a tighter bound on the rate of growth of d_t is available then one can improve the convergence rate to the set of CCE by picking a more slowly decaying η_t than $\eta_t=\frac{1}{t\log t\log\log t\log\log\log t}$. In general, for a given T, FKM gives an ε-CCE with $\varepsilon=O\left(\max\left\{\frac{1}{\delta_T^2\sum_{t=1}^{T}\eta_t},\delta_T\right\}\right)$, as given in (36).

Lemma 4. *FKM with step size sequence $\{\eta_t\}$ and sampling radius δ_T has no discounted-regret (even with an adaptive adversary) with respect to the sequence of delays $\{d_t\}$ if the following three conditions hold:*

1. $\sum_{t=1}^{\infty}\eta_t=\infty$.
2. $\lim_{t\to\infty}\eta_t d_t<\infty$ *and* $\sum_{t=1}^{\infty}d_t\eta_t^2<\infty$.
3. $\lim_{T\to\infty}\delta_T=0$ *and* $\lim_{T\to\infty}\delta_T^2\sum_{t=1}^{T}\eta_t=\infty$.

Proof Let $\mathcal{M}=\{t\,|\,t+d_t>T,\,t\in[1,T]\}$. Define $t^*(T)=\min_{t\in\mathcal{M}}t$, and note that $t^*(T)\to\infty$ as $T\to\infty$ since $t+d_t\ge t$, and $f(t)=t$ is increasing. Since η_t is non-increasing then

$$\sum_{t\in\mathcal{M}}\eta_t\le|\mathcal{M}|\,\eta_{t^*(T)}\le(T-t^*(T)+1)\,\eta_{t^*(T)}\le d_{t^*(T)}\eta_{t^*(T)}. \tag{35}$$

Let $A = (3 + |\mathcal{K}|)\, L$. Therefore

$$\lim_{T\to\infty} \mathbb{E}\left\{ \frac{\sum_{t=1}^{T} \eta_t \left(l_t\left(\boldsymbol{a}_t\right) - l_t\left(\boldsymbol{a}^*\right)\right)}{\sum_{t=1}^{T} \eta_t} \right\}$$
$$\underset{(a)}{\leq} \lim_{T\to\infty} \frac{d_{t^*(T)}\eta_{t^*(T)} + \frac{|\mathcal{K}|^2}{2} + A\delta_T \sum_{t=1}^{T}\eta_t + \frac{2n|\mathcal{K}|}{\delta_T}\sqrt{\sum_{t=1}^{T}\eta_t^2} + \frac{1}{2}\frac{n^2}{\delta_T^2}\sum_{t=1}^{T}\eta_t^2\left(1+4d_t\right)}{\sum_{t=1}^{T}\eta_t} \underset{(b)}{=} 0 \quad (36)$$

where (a) is Lemma 3 and (35), and (b) follows since $\lim_{t\to\infty} d_t\eta_t < \infty$, $\sum_{t=1}^{\infty}\eta_t^2 d_t < \infty$, $\sum_{t=1}^{\infty}\eta_t = \infty$, $\lim_{T\to\infty}\delta_T = 0$ and $\lim_{T\to\infty}\delta_T^2\sum_{t=1}^{T}\eta_t = \infty$. ■

The following Proposition shows that FKM can have no discounted-regret even when it has linear regret in T. As a result, FKM can lead to the set of CCE in a non-cooperative game or to a NE in a zero-sum despite not having no-regret guarantees.

Proposition 1. *There exist a delay sequence $\{d_t\}$ and Lipschitz continuous convex functions $\{l_1, ..., l_T\}$ on $[0,1]$ such that for a large enough T*

$$\mathbb{E}\left\{R\left(T\right)\right\} = \sum_{t=1}^{T}\mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right)\right\} - \sum_{t=1}^{T} l_t\left(\boldsymbol{a}^*\right) \geq \frac{T}{2\left(|\mathcal{K}|+1\right)^2} \quad (37)$$

but still the step sizes $\{\eta_t\}$ and sampling radius δ_T for Algorithm 2 (FKM) can be chosen such that it has no discounted-regret with respect to $\{d_t\}$.

Proof Let $d_t = t$ and choose $\delta_T = \frac{1}{L}\left(\log\log T\right)^{-\frac{1}{3}}, \eta_t = \frac{1}{t\log(t+1)}$ for all $t > 0$, for which $\sum_{t=1}^{T}\eta_t = O\left(\log\log T\right), \sum_{t=1}^{\infty} d_t\eta_t^2 < \infty$ and also $\delta_T \to 0$ and $\delta_T^2\sum_{t=1}^{T}\eta_t \to \infty$ as $T\to\infty$. Hence, by Lemma 4 we obtain that FKM has no discounted-regret with respect to d_t. However, the feedback for the last $\frac{T}{2}$ rounds is never received. Therefore, the unperturbed action $\boldsymbol{x}_t$ is $\boldsymbol{x}_{\frac{T}{2}}$ for all $t \geq \frac{T}{2}$. Consider the sequence of costs $l_t\left(\boldsymbol{a}\right) = \boldsymbol{0}$ for all $t \leq \frac{T}{2}$ and $l_t\left(\boldsymbol{a}\right) = \frac{\left\|\boldsymbol{a} - \frac{1}{\sqrt{n}}\boldsymbol{1}\right\|^2}{\left(|\mathcal{K}|+1\right)^2}$ for all $t > \frac{T}{2}$ where $\boldsymbol{1} \in \mathbb{R}^n$ is a vector of ones. Starting from $\boldsymbol{x}_0 = \boldsymbol{0}$ and computing $\boldsymbol{g}_t = \boldsymbol{0}$ for all $t \leq \frac{T}{2}$ we have $\boldsymbol{x}_{\frac{T}{2}} = \boldsymbol{0}$. Then, from the Lipschitz continuity of l_t we obtain for all $t > \frac{T}{2}$, for large enough T,

$$\mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right)\right\} = \mathbb{E}\left\{l_t\left(\boldsymbol{x}_{\frac{T}{2}} + \delta_T\boldsymbol{u}_t\right)\right\} \geq \mathbb{E}\left\{l_t\left(\boldsymbol{0}\right)\right\} - \delta_T L = \frac{1}{\left(|\mathcal{K}|+1\right)^2} - \delta_T L \geq \frac{1}{2\left(|\mathcal{K}|+1\right)^2} \quad (38)$$

which means that this sequence yields an expected regret of at least $\frac{T}{2(|\mathcal{K}|+1)^2}$. ■

6. The EXP3 Algorithm for Adversarial Multi-armed Bandits with Delayed Feedback

Consider a player that at each round t has to pick one out of K arms. Let a_t be the arm the player chooses at round t. The cost at round t from arm i is $l_t^{(i)} \in [0,1]$, and let $\boldsymbol{l}_t = \left(l_t^{(1)}, ..., l_t^{(K)}\right)$ be the cost vector. At round t, the EXP3 algorithm, detailed in Algorithm 3, chooses an arm at random using a distribution that depends on the history of the game. The variant when $\gamma_t \neq 0$, as we use against an adaptive adversary, is known as EXP3-IX (see Neu (2015)). We denote the vector of probabilities of the player for choosing arms at round t by $\boldsymbol{p}_t \in \Delta^K$, where Δ^K denotes the K-simplex. This is also known as the mixed strategy of the player. We also define the following filtration

$$\mathcal{F}_t = \sigma\left(\{\boldsymbol{a}_s \,|\, s + d_s \leq t\} \cup \{\boldsymbol{l}_s \,|\, s \leq t\}\right) \tag{39}$$

which is generated from all the actions for which the feedback was received up to round t and all previous cost functions. Note that the mixed strategy $\boldsymbol{p}_t$ is a $\mathcal{F}_t$-measurable random variable since it uses information that was received up to time t. With a slight abuse of notation, we use $\mathcal{F}_{s_-}$ to denote the filtration induced from all actions for which the feedback has been received up to time s_- and all the cost functions of the rounds before t.

The next Lemma is the main result of this section, used to prove both Theorem 6 and Lemma 6.

Lemma 5. *Let $\{\eta_t\}$ be a non-increasing step size sequence. Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Define the set $\mathcal{M}^*$ of all samples that are not received before round T or that were delayed by $d_t \geq \frac{1}{e^2 \eta_t} - 1$. Then using EXP3 (Algorithm 3) guarantees*

1. *With an oblivious adversary and $\gamma_t = 0$ for all t:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} \eta_t l_t^{(a_t)} - \min_i \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} l_t^{(i)}\right\} \leq \ln K + 4K \sum_{t=1}^{T} \eta_t^2 + 4 \sum_{t \notin \mathcal{M}^*} \eta_t^2 d_t + \sum_{t \in \mathcal{M}^*} \eta_t. \tag{40}$$

2. *With an adaptive adversary and $\gamma_t = \eta_t$ for all t:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} \eta_t l_t^{(a_t)} - \min_i \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} l_t^{(i)}\right\} \leq 2 + 2\ln K + 5K \sum_{t=1}^{T} \eta_t^2 + 30K \sum_{t \notin \mathcal{M}^*} \eta_t^2 d_t + \sum_{t \in \mathcal{M}^*} \eta_t. \tag{41}$$

The following theorem establishes the expected regret bound for EXP3 with delays. It is proved by optimizing over a constant step size η in Lemma 3.

Theorem 6. *Define the set $\mathcal{M}$ of all samples that are not received before round T. Let us choose the fixed step size $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\ln K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$. Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \mathbb{E}\left\{\sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{p}_t \rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} = O\left(\sqrt{\ln K \left(KT + \sum_{t \notin \mathcal{M}} d_t\right)} + |\mathcal{M}|\right). \tag{44}$$

Proof We choose $\eta_t = \eta$ in (40) of Lemma 5 and define $\mathcal{D} = \left\{t \,|\, d_t \geq \frac{1}{e^2 \eta} - 1 \text{ and } t + d_t \leq T\right\}$ so $\mathcal{M}^* = \mathcal{M} \cup \mathcal{D}$ in the statement of Lemma 5. Now divide both sides by η:

$$\mathbb{E}\left\{\sum_{t=1}^{T} l_t^{(a_t)} - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} \leq \frac{\ln K}{\eta} + 4\eta KT + 4\eta \sum_{t \notin \mathcal{M}^*} d_t + |\mathcal{M}| + |\mathcal{D}|. \tag{45}$$

Algorithm 3 EXP3 with delays

Initialization: Let $\{\eta_t\}$ and $\{\gamma_t\}$ be a positive non-increasing sequences such that $\eta_1 < \frac{e^{-2}}{2}$ and set $\tilde{L}_1^{(i)} = 0$ and $p_1^{(i)} = \frac{1}{K}$ for $i = 1, ..., K$.

For $t = 1, ..., T$ **do**

1. Choose an arm a_t at random according to the distribution $\boldsymbol{p}_t$.
2. Collect in $\mathcal{S}_t$ all the rounds s for which $l_s^{(a_s)}$ arrived at round t after a delay of $d_s \leq \frac{1}{e^2 \eta_s} - 1$.
3. Update the weights of arm a_s for all $s \in \mathcal{S}_t$ using
$$\tilde{L}_t^{(a_s)} = \tilde{L}_{t-1}^{(a_s)} + \eta_s \frac{l_s^{(a_s)}}{p_s^{(a_s)} + \gamma_s}. \tag{42}$$
4. Update the mixed strategy
$$p_{t+1}^{(i)} = \frac{e^{-\tilde{L}_t^{(i)}}}{\sum_{j=1}^{n} e^{-\tilde{L}_t^{(i)}}}. \tag{43}$$

End

Then, choosing $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\ln K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$ yields (44). Note that for this choice $|\mathcal{D}| \leq \frac{\sum_{t \notin \mathcal{M}} d_t}{e^{-2}\eta^{-1} - 1} \leq \sqrt{\left(\sum_{t \notin \mathcal{M}} d_t\right) \ln K}$, since $\sum_{t \notin \mathcal{M}} d_t \geq \left(\frac{1}{e^2 \eta} - 1\right) |\mathcal{D}|$ (discarded samples in $\mathcal{D}$ are not missing). ■

Similar to the bandit convex optimization case, the bound of Theorem 6 is tighter than $O\left(\sqrt{\ln K \left(KT + D\right)}\right)$ for $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$, as the next Corollary shows.

Corollary 2. *Let* $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\ln K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$. *Let* $\left\{l_t^{(i)}\right\}$ *be a cost sequence such that* $l_t^{(i)} \in [0, 1]$ *for every* t, i. *Let* $\{d_t\}$ *be a delay sequence such that the cost from round* t *is received at round* $t + d_t$. *Let* $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$. *Then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \mathbb{E}\left\{\sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{p}_t \rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} = O\left(\sqrt{\ln K \left(KT + D\right)}\right). \tag{46}$$

Proof The $m = |\mathcal{M}|$ missing samples contribute at least $\frac{m(m+1)}{2}$ to D (as in Corollary 2), so

$$\sqrt{\ln K \left(KT + D\right)} \geq \sqrt{\ln K \left(KT + \sum_{t \notin \mathcal{M}} d_t + \frac{m(m+1)}{2}\right)}$$
$$\underset{(a)}{\geq} \frac{1}{2}\sqrt{2 \ln K \left(KT + \sum_{t \notin \mathcal{M}} d_t\right)} + \frac{1}{2}\sqrt{\ln K m (m+1)} \geq O\left(\sqrt{\ln K \left(KT + \sum_{t \notin \mathcal{M}} d_t\right)}\right) + \frac{|\mathcal{M}|}{2} \tag{47}$$

where (a) follows from the concavity of $f(x) = \sqrt{x}$. ■

6.1 Agnostic EXP3

The step size $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\ln K}{KT+\sum_{t\notin\mathcal{M}} d_t}}$ used in Algorithm 3 requires knowing the horizon T and the sum of delays D. When these parameters are unknown, we can apply Algorithm 1 to wrap EXP3. The next Theorem is an immediate application Theorem 3 on the EXP3 regret bound for this agnostic case. The resulting bound retains the same order of magnitude as the bound of Corollary 2 even though D and T are unknown. The only difference with the bound of Theorem 6 arises because the doubling trick discards samples that cross super-epochs. Hence, the bound below uses $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$ instead of $\sum_{t\notin\mathcal{M}} d_t$ and $|\mathcal{M}|$.

Theorem 7. *Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$. If the player uses Algorithm 1 to wrap EXP3 (Algorithm 3) with step size $\eta_{w,h} = \frac{1}{2}e^{-2}\sqrt{\frac{\ln K}{\max\{K2^h, 2^w\}}}$ for epoch (w,h). Then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \mathbb{E}\left\{\sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{p}_t\rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} = O\left(\sqrt{\ln K\,(KT+D)}\right) \tag{48}$$

Proof Consider the regret bound in (45) after bounding $|\mathcal{M}| \le \sqrt{2D}$, (as in the proof of Corollary 2), $T - |\mathcal{M}| \le T$, $\sum_{t\notin\mathcal{M}} d_t \le D$ and $|\mathcal{D}| \le \sqrt{D\ln K}$ (as in the proof of Theorem 6). For $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\ln K}{KT+D}}$, this bound yields that of Corollary 2 which is of the form $\sqrt{TK\ln K} + \sqrt{D\ln K}$, so it matches Assumption 1 in Theorem 3 with $a = b = 0$ and $c = d = \frac{1}{2}$. This bound also satisfies Assumption 2 since it is increasing with T and D for any η. ■

6.2 No Discounted-Regret Property for EXP3

In this subsection, we provide conditions for EXP3 to have no discounted regret with respect to the delay sequence. As discussed in Section 3, $\sum_{t=1}^{\infty}\eta_t = \infty$, is necessary for the no discounted-regret property to be non-trivial. All other conditions of Lemma 6 are as tight as the bound of Lemma 5. Note that one can choose $\eta_t = \frac{1}{t\log t \log\log t \log\log\log t}$ to guarantee no discounted-regret for all sequences such that $d_t = (t\log t\log\log t)$. This boundary can only be slightly improved by adding $\log(\log(\ldots\log(T)))$ iteratively in this manner as long as $\sum_{t=1}^{\infty}\frac{1}{d_t} = \infty$. Hence, no knowledge of the actual sequence d_t is required to tune η_t and δ_T. However, if a tighter bound on the rate of growth of d_t is available then one can improve the convergence rate to the set of CCE by picking a more slowly decaying η_t than $\eta_t = \frac{1}{t\log t\log\log t\log\log\log t}$. In general, for a given T, EXP3 gives an ε-CCE with $\varepsilon = O\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)$, as given in (50).

Lemma 6. *EXP3 with step size sequence $\{\eta_t\}$ has no discounted-regret with respect to the sequence of delays $\{d_t\}$ if the following two conditions hold:*

1. $\sum_{t=1}^{\infty}\eta_t = \infty$.
2. $\lim\limits_{t\to\infty}\eta_t d_t < \infty$ *and* $\sum_{t=1}^{\infty} d_t\eta_t^2 < \infty$.

Proof Define the set of missing samples $\mathcal{M} = \{t \mid t + d_t > T\}$ and the set of discarded samples $\mathcal{D}_T = \left\{t \mid d_t\eta_t > e^{-2} - \eta_t\right\}$. Define $t^*(T) = \min\limits_{t\in\mathcal{M}} t$, and note that $t^*(T) \to \infty$ as $T \to \infty$ since $t + d_t \ge t$, and $f(t) = t$ is increasing. Since η_t is non-increasing then

$$\sum_{t\in\mathcal{M}}\eta_t \le |\mathcal{M}|\,\eta_{t^*(T)} \le (T - t^*(T) + 1)\,\eta_{t^*(T)} \le d_{t^*(T)}\eta_{t^*(T)}. \tag{49}$$

Given $\sum_{t=1}^{\infty} d_t \eta_t^2 < \infty$ and $\sum_{t=1}^{\infty} \eta_t = \infty$ we must have $\lim_{t\to\infty} \eta_t d_t = 0$ if this limit exists, so $\lim_{T\to\infty} |\mathcal{D}_T| < \infty$. Therefore for the optimal arm i^*

$$\lim_{T\to\infty} \frac{\mathbb{E}\left\{\sum_{t=1}^{T} \eta_t \left(l_t^{(a_t)} - l_t^{(i^*)}\right)\right\}}{\sum_{t=1}^{T} \eta_t} \underset{(a)}{\leq} \lim_{T\to\infty} \frac{\eta_1 |\mathcal{D}_T| + d_{t^*(T)}\eta_{t^*(T)} + 4\ln K + 5K\sum_{t=1}^{T} \eta_t^2 (1+6d_t)}{\sum_{t=1}^{T} \eta_t} \underset{(b)}{=} 0 \tag{50}$$

where (a) is Lemma 5 and (49), and (b) uses $d_t\eta_t \to 0$ as $t \to \infty$, $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} d_t \eta_t^2 < \infty$. ■

The following Proposition shows that EXP3 can have no discounted-regret even when it has linear regret in T. As a result, EXP3 can lead to the set of CCE in a non-cooperative game or to the set of NE in a zero-sum game despite not having no-regret guarantees.

Proposition 2. *There exist a delay sequence $\{d_t\}$ and a cost sequence $\left\{l_t^{(1)}, ..., l_t^{(K)}\right\}_t$ with $0 \leq l_t^{(i)} \leq 1$ for all t and i, such that*

$$\mathbb{E}\{R(T)\} = \mathbb{E}\left\{\sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{p}_t \rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} \geq \left(1 - \frac{1}{K}\right)\frac{T}{2} \tag{51}$$

but still the step sizes $\{\eta_t\}$ for Algorithm 3 (EXP3) can be chosen such that it has no discounted-regret with respect to $\{d_t\}$.

Proof Let $d_t = t$ and $\eta_t = \frac{1}{t\log(t+1)}$ for all t, for which $d_t\eta_t^2 = \frac{1}{t\log^2(t+1)}$ so $\sum_{t=1}^{\infty} \eta_t = \infty$, $\sum_{t=1}^{\infty} d_t\eta_t^2 < \infty$ and $\lim_{t\to\infty} \eta_t d_t = 0$. Hence, by Lemma 6 we obtain that EXP3 has no discounted regret with respect to d_t. However, the feedback for the last $\frac{T}{2}$ rounds is never received. Therefore, the strategy $\boldsymbol{a}_t$ stays constant for all $t \geq \frac{T}{2}$. Then the cost sequence $l_t^{(i)} = 0$ for all i and all $t \leq \frac{T}{2}$ and $l_t^{(1)} = 0$, $l_t^{(j)} = 1$ for all $j > 1$ and all $t > \frac{T}{2}$ yields an expected regret of exactly $\left(1 - \frac{1}{K}\right)\frac{T}{2}$. ■

7. Conclusions

We studied the regret of online learning algorithms with adversarial bandit feedback and an arbitrary sequence of delays $\{d_t\}$. We showed that FKM achieves an expected regret of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ and EXP3 achieves an expected regret of $O\left(\sqrt{\ln K\,(KT+D)}\right)$ where $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$. These bounds hold even if D and T are unknown thanks to a novel doubling trick. Our doubling trick can be applied to any online learning algorithm with delays, and we showed that under mild conditions it retains the same order of magnitude dependence on D, T of the regret bound for the case when D, T are known. Our single-agent bounds show that the widely applied FKM and EXP3 are inherently robust to delays, which is a valuable performance guarantee.

Our single-agent results in this paper focus on FKM and EXP3 since they are the most widely used algorithms for multi-armed bandits and bandit convex optimization, respectively. Therefore it is crucial to understand how they perform under delays, which are prevalent in practical systems. However, this leaves open the question of what are the best algorithms for delayed bandit feedback.

For multi-armed bandits the lower bound is $O\left(\sqrt{KT + D\ln K}\right)$, which is achieved by the algorithm of Zimmert and Seldin (2019). EXP3 achieves this lower bound up to the $\ln K$ factor, which is negligible if the average delay is larger than $O\left(\frac{K}{\ln K}\right)$, and has lower computational complexity.

For bandit convex optimization, much less is understood. Recently, a bandit convex optimization algorithm has been proposed in Bubeck and Eldan (2016), that achieves regret $O\left(n^{9.5}\sqrt{T}\log^{7.5} T\right)$.

However, their proposed algorithm suffers from a few drawbacks. First, the n dependence is a high-degree polynomial, which is much worse than the linear dependence of FKM. It was conjectured in Bubeck et al. (2017) (but is still open) that a variant of their algorithm may achieve an expected regret of $O\left(n^{1.5}\sqrt{T}\right)$. Second, the algorithm proposed in Bubeck and Eldan (2016), as well as its conjectured variant, still has a T dependent complexity per round. It is an open question how robust their algorithm is to delays. As the comparison between our EXP3 and FKM results suggests, better T dependence does not imply better D dependence. This is a challenging question since their algorithm requires an increasing step-size sequence, which conflicts with the tuning that is required to optimize against delays or to guarantee no discounted-regret for unbounded delay sequences.

For the multi-agent case, we introduced the notion of the discounted regret, where the regret is discounted by the step-size sequence. When learning with delayed feedback, algorithms have no discounted regret even under significant unbounded delay sequences (e.g., $d_t = O(t \log t)$) for which their regret is linear in T. Based on that, we were able to show that the discounted ergodic distribution of play converges to a coarse correlated equilibrium for a general non-cooperative game. For a zero-sum game, the discounted ergodic average of play converges to the set of Nash equilibria.

Our encouraging game-theoretic results hold for any online learning algorithm that has no discounted-regret with respect to the delay sequence. This is a much broader family than just FKM and EXP3. Studying this family of algorithms and their limitations is a fascinating research avenue. Is no discounted-regret necessary to learn a CCE or a NE with delays?

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, 2011.

Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. On the value of correlation. *Journal of Artificial Intelligence Research*, 33:575–613, 2008.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995.

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

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 2018.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2019.

Michael Bowling. Convergence and no-regret in multiagent learning. In *Advances in neural information processing systems*, 2005.

Sébastien Bubeck and Ronen Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85. ACM, 2017.

Yang Cai and Constantinos Daskalakis. On minmax theorems for multiplayer games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 217–234, 2011.

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773, 2018.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

András György and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. *arXiv preprint arXiv:2010.06022*, 2020.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Sergiu Hart. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*, volume 4. World Scientific, 2013.

Amélie Héliou, Panayotis Mertikopoulos, and Zhengyuan Zhou. Gradient-free online learning in games with delayed rewards. *arXiv preprint arXiv:2006.10911*, 2020.

Lisa Hellerstein, Thomas Lidbetter, and Daniel Pirutinsky. Solving zero-sum games using best-response oracles with applications to search games. *Operations Research*, 2019.

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

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. *arXiv preprint arXiv:2012.14843*, 2020.

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

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

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

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, pages 1–43, 2016.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, 2010.

Hukukane Nikaidô, Kazuo Isoda, et al. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(Suppl. 1):807–815, 1955.

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

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in neural information processing systems*, 2015.

George F Simmons. *Introduction to topology and modern analysis*, volume 44. Tokyo, 1963.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208, 2007.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, 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*, 2020.

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

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W Glynn, and Claire Tomlin. Countering feedback delays in multi-agent learning. In *Advances in Neural Information Processing Systems*, 2017.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. *arXiv preprint arXiv:1910.06054*, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003.

Martin Zinkevich, John Langford, and Alex J Smola. Slow learners are fast. In *Advances in neural information processing systems*, 2009.

8. Proof of Theorem 1

We start by showing that $\rho_T(\boldsymbol{a}) = \frac{\sum_{t=1}^{T} \eta_t \mathbb{1}_{\{\boldsymbol{a}_t = \boldsymbol{a}\}}}{\sum_{t=1}^{T} \eta_t}$ converges in probability to an ε-CCE of the game as $T \to \infty$, for every $\varepsilon > 0$. For each n, define the cost function $l_{n,t}(\boldsymbol{a}_n) = 1 - u_n(\boldsymbol{a}_n, \boldsymbol{a}_{-n,t})$. Let $\varepsilon > 0$. Since each player n has no discounted regret with respect to d_t^n, then there exists a $T_0 > 0$ such that for all $T > T_0$, we have for every n and every strategy $\boldsymbol{a}_n \in \mathcal{A}$:

$$\mathbb{E}\left\{\mathbb{E}^{\boldsymbol{a}^* \sim \rho_T}\left\{u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n}^*\right) - u_n\left(\boldsymbol{a}^*\right)\right\}\right\} \underset{(a)}{=} \mathbb{E}\left\{\frac{\sum_{t=1}^{T} \eta_t \left(u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n,t}\right) - u_n\left(\boldsymbol{a}_t\right)\right)}{\sum_{t=1}^{T} \eta_t}\right\}$$

$$= \mathbb{E}\left\{\frac{\sum_{t=1}^{T} \eta_t \left(l_{n,t}\left(\boldsymbol{a}_{n,t}\right) - l_{n,t}\left(\boldsymbol{a}_n\right)\right)}{\sum_{t=1}^{T} \eta_t}\right\} \underset{(b)}{\leq} \varepsilon \quad (52)$$

where (a) uses the definition of ρ_T and (b) follows since player n has no discounted-regret. Now pick $\boldsymbol{a}_n = \underset{\boldsymbol{a}'_n \in \mathcal{A}}{\arg\max} \mathbb{E}^{\boldsymbol{a}^* \sim \rho_T} \left\{ u_n \left(\boldsymbol{a}'_n, \boldsymbol{a}^*_{-n} \right) \right\}$ in (52). Since $\mathbb{E}^{\boldsymbol{a}^* \sim \rho_T} \left\{ u_n \left(\boldsymbol{a}_n, \boldsymbol{a}^*_{-n} \right) - u_n \left(\boldsymbol{a}^* \right) \right\}$ is linear in ρ_T, then

$$\mathbb{E} \left\{ \mathbb{E}^{\boldsymbol{a}^* \sim \rho_T} \left\{ u_n \left(\boldsymbol{a}_n, \boldsymbol{a}^*_{-n} \right) - u_n \left(\boldsymbol{a}^* \right) \right\} \right\} = \mathbb{E}^{\boldsymbol{a}^* \sim \mathbb{E}\{\rho_T\}} \left\{ u_n \left(\boldsymbol{a}_n, \boldsymbol{a}^*_{-n} \right) - u_n \left(\boldsymbol{a}^* \right) \right\} \tag{53}$$

so we conclude that by definition $\mathbb{E}\{\rho_T\}$ is an ε-CCE. Let $\Delta > 0$. From Lemma 7, we know that there exists an $\varepsilon_\Delta > 0$ such that for all $\varepsilon \leq \varepsilon_\Delta$ we have $\min_{\rho^* \in \mathcal{C}_0} \|\rho_\varepsilon - \rho^*\| \leq \Delta$ for all $\rho_\varepsilon \in \mathcal{C}_\varepsilon$. Let $\delta > 0$ and let $\varepsilon = \varepsilon_\Delta \delta > 0$. Then from (52) we know that there exists a large enough T_1 such that for all $T > T_1$ we have $\mathbb{E}\{\rho_T\} \in \mathcal{C}_{\varepsilon_\Delta}$, which implies that $\min_{\rho^* \in \mathcal{C}_0} \|\mathbb{E}\{\rho_T\} - \rho^*\| \leq \Delta$. Therefore, $\mathbb{E}\{\rho_T\}$ converges to the set of CCE as $T \to \infty$.

9. Proof of Theorem 2

We start by showing that $(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T)$ converges in L^1 to the set of ε-NE of the game as $T \to \infty$, for every $\varepsilon > 0$. Define the ergodic average of the value of the game by

$$\overline{u}_T = \frac{\sum_{t=1}^T \eta_t u \left(\boldsymbol{y}_t, \boldsymbol{z}_t \right)}{\sum_{t=1}^T \eta_t}. \tag{54}$$

Define the row cost function $l_{r,t}(\boldsymbol{y}) = u(\boldsymbol{y}, \boldsymbol{z}_t)$. Since the row player has no discounted-regret with respect to d_t^r, then there exists a $T_1 > 0$ such that for all $T > T_1$, and every $\boldsymbol{y} \in \mathcal{A}$:

$$\begin{aligned} \mathbb{E}\left\{ \overline{u}_T - u(\boldsymbol{y}, \bar{\boldsymbol{z}}_T) \right\} \underset{(a)}{\leq} \mathbb{E} &\left\{ \frac{\sum_{t=1}^T \eta_t \left(u\left(\boldsymbol{y}_t, \boldsymbol{z}_t\right) - u\left(\boldsymbol{y}, \boldsymbol{z}_t\right) \right)}{\sum_{t=1}^T \eta_t} \right\} \\ &= \mathbb{E}\left\{ \frac{\sum_{t=1}^T \eta_t \left(l_{r,t}\left(\boldsymbol{y}_t\right) - l_{r,t}\left(\boldsymbol{y}\right) \right)}{\sum_{t=1}^T \eta_t} \right\} \underset{(b)}{\leq} \frac{\varepsilon}{2} \end{aligned} \tag{55}$$

where (a) uses the concavity of $u(\boldsymbol{y}, \bar{\boldsymbol{z}}_T)$ in $\bar{\boldsymbol{z}}_T$ and (b) uses the no discounted-regret of the algorithm.

Define the column cost function $l_{c,t}(\boldsymbol{z}) = 1 - u(\boldsymbol{y}_t, \boldsymbol{z})$. Since the column player has no discounted-regret with respect to d_t^c, then there exists a $T_2 > 0$ such that for all $T > T_2$ and for every $\boldsymbol{z} \in \mathcal{A}$:

$$\begin{aligned} \mathbb{E}\left\{ u(\bar{\boldsymbol{y}}_T, \boldsymbol{z}) - \overline{u}_T \right\} \underset{(a)}{\leq} \mathbb{E} &\left\{ \frac{\sum_{t=1}^T \eta_t \left(u\left(\boldsymbol{y}_t, \boldsymbol{z}\right) - u\left(\boldsymbol{y}_t, \boldsymbol{z}_t\right) \right)}{\sum_{t=1}^T \eta_t} \right\} \\ &= \mathbb{E}\left\{ \frac{\sum_{t=1}^T \eta_t \left(l_{c,t}\left(\boldsymbol{z}_t\right) - l_{c,t}\left(\boldsymbol{z}\right) \right)}{\sum_{t=1}^T \eta_t} \right\} \underset{(b)}{\leq} \frac{\varepsilon}{2} \end{aligned} \tag{56}$$

where (a) uses the convexity of $u(\bar{\boldsymbol{y}}_T, \boldsymbol{z})$ in $\bar{\boldsymbol{y}}_T$ and (b) uses the no discounted-regret of the algorithm.

Now define the best-response to $\bar{\boldsymbol{z}}_T$ as $\boldsymbol{y}_T^b = \underset{\boldsymbol{y}'}{\arg\min}\, u(\boldsymbol{y}', \bar{\boldsymbol{z}}_T)$ and the best-response to $\bar{\boldsymbol{y}}_T$ as $\boldsymbol{z}_T^b = \underset{\boldsymbol{z}'}{\arg\max}\, u(\bar{\boldsymbol{y}}_T, \boldsymbol{z}')$. By choosing $\boldsymbol{y} = \boldsymbol{y}_T^b$, $\boldsymbol{z} = \bar{\boldsymbol{z}}_T$ in (55) and (56) and adding them together we conclude that for all $T > \max\{T_1, T_2\}$

$$\mathbb{E}\left\{ \left| u(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \min_{\boldsymbol{y}'} u(\boldsymbol{y}', \bar{\boldsymbol{z}}_T) \right| \right\} \underset{(a)}{=} \mathbb{E}\left\{ \overline{u}_T - u\left(\boldsymbol{y}_T^b, \bar{\boldsymbol{z}}_T\right) \right\} + \mathbb{E}\left\{ u(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \overline{u}_T \right\} \leq \varepsilon \tag{57}$$

where (a) follows since $u(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) \geq u\left(\boldsymbol{y}_T^b, \bar{\boldsymbol{z}}_T\right)$. By choosing instead $\boldsymbol{y} = \bar{\boldsymbol{y}}_T$, $\boldsymbol{z} = \boldsymbol{z}_T^b$ in (55) and (56) and adding them together we conclude that for all $T > \max\{T_1, T_2\}$

$$\mathbb{E}\left\{ \left| u(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \max_{\boldsymbol{z}'} u(\bar{\boldsymbol{y}}_T, \boldsymbol{z}') \right| \right\} \underset{(a)}{=} \mathbb{E}\left\{ \overline{u}_T - u(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) \right\} + \mathbb{E}\left\{ u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}_T^b\right) - \overline{u}_T \right\} \leq \varepsilon \tag{58}$$

where (a) follows since $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \leq u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}_T^b\right)$.

Let $\Delta > 0$. From Lemma 7, we know that there exists an $\varepsilon_\Delta > 0$ such that for all $\varepsilon \leq \varepsilon_\Delta$ we have $\min\limits_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\right\| \leq \Delta$ for all $\boldsymbol{x}_\varepsilon \in \mathcal{N}_\varepsilon$. Let $\delta > 0$ and let $\varepsilon = \frac{\varepsilon_\Delta \delta}{2} > 0$. Then from (57),(58) we know that there exists a large enough T_3 we have for all $T > T_3$, using Markov inequality:

$$\mathbb{P}\left(\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)\right| \geq \varepsilon_\Delta\right) \leq \frac{\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)\right|\right\}}{\varepsilon_\Delta} = \frac{\delta}{2} \tag{59}$$

and

$$\mathbb{P}\left(\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)\right| \geq \varepsilon_\Delta\right) \leq \frac{\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)\right|\right\}}{\varepsilon_\Delta} = \frac{\delta}{2}. \tag{60}$$

Hence by the union bound over (59) and (59):

$$\mathbb{P}\left(\min_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \boldsymbol{x}^*\right\| \leq \Delta\right) \geq \mathbb{P}\left(\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \in \mathcal{N}_{\varepsilon_\Delta}\right) \geq 1 - \delta \tag{61}$$

so $\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right)$ converges in probability to the set of NE. Since $\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \boldsymbol{x}^*$ is bounded, it implies that $\mathbb{E}\left\{\min\limits_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \boldsymbol{x}^*\right\|\right\} \to 0$ as $T \to \infty$. Since u is continuous, $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \to v$ in probability as $T \to \infty$ where v is the value of the game. Since u is bounded, $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \to v$ in L^1 as $T \to \infty$.

10. The Set of Approximate Equilibria Approaches the Set of Equilibria

The following lemma shows that for a given game, the sets of ε-NE and ε-CCE converge to the sets of NE and CCE, respectively, when $\varepsilon \to 0$. It allows us to convert convergence to $\mathcal{N}_\varepsilon$ and $\mathcal{C}_\varepsilon$ for each $\varepsilon > 0$ to convergence to $\mathcal{N}_0$ and $\mathcal{C}_0$.

Lemma 7. *Let* $d_{\mathcal{N}}\left(\varepsilon\right) = \max\limits_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_\varepsilon} \min\limits_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\right\|$ *and* $d_{\mathcal{C}}\left(\varepsilon\right) = \max\limits_{\rho_\varepsilon \in \mathcal{C}_\varepsilon} \min\limits_{\rho^* \in \mathcal{C}_0} \left\|\rho_\varepsilon - \rho^*\right\|$. *Then* $d_{\mathcal{N}}\left(\varepsilon\right) \to 0$ *and* $d_{\mathcal{C}}\left(\varepsilon\right) \to 0$ *as* $\varepsilon \to 0$.

Proof Let $A_\varepsilon\left(\boldsymbol{y}\right) = \left\{\boldsymbol{x} \in \mathcal{A}^N \mid u_n\left(\boldsymbol{x}\right) \geq u_n\left(\boldsymbol{y}_n, \boldsymbol{x}_{-n}\right) - \varepsilon, \forall n\right\}$. This is a compact set since $A_\varepsilon\left(\boldsymbol{y}\right) = \bigcap_n f_{n,\boldsymbol{y}}^{-1}\left(\left[-\varepsilon, u_{n,\max}\right]\right)$ for the continuous $f_{n,\boldsymbol{y}}\left(\boldsymbol{x}\right) = u_n\left(\boldsymbol{x}\right) - u_n\left(\boldsymbol{y}_n, \boldsymbol{x}_{-n}\right)$ and $u_{n,\max} = \max\limits_{\boldsymbol{x} \in \mathcal{A}^N} f_{n,\boldsymbol{y}}\left(\boldsymbol{x}\right)$. Now note that $\mathcal{N}_\varepsilon = \bigcap_{\boldsymbol{y} \in \mathcal{A}^N} A_\varepsilon\left(\boldsymbol{y}\right)$, since $\mathcal{N}_\varepsilon$ only includes strategy profiles where no deviation gives any player more than ε gain over $\boldsymbol{x} \in \bigcap_{\boldsymbol{y} \in \mathcal{A}^N} A_\varepsilon\left(\boldsymbol{y}\right)$ (if exists). Hence $\mathcal{N}_\varepsilon$ is compact for all $\varepsilon \geq 0$.

Since $\mathcal{N}_{\frac{1}{n}}$ and $\mathcal{N}_0$ are compact, we can define the sequence $\left\{\tilde{\boldsymbol{x}}_n\right\}$ such that

$$\tilde{\boldsymbol{x}}_n \in \arg\max_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_{\frac{1}{n}}} \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\right\|. \tag{62}$$

Since $\mathcal{A}^N$ is compact then the infinite sequence $\tilde{\boldsymbol{x}}_n$ has a subsequence $\tilde{\boldsymbol{x}}_{n_k}$ that converges to a point $\tilde{\boldsymbol{x}} \in \mathcal{A}^N$ (i.e., Bolzano–Weierstrass Theorem, see Simmons (1963)). Since we must have $\tilde{\boldsymbol{x}} \in \bigcap_{n=1}^\infty \mathcal{N}_{\frac{1}{n}}$ then $\max\limits_{\boldsymbol{y}_n} u_n\left(\boldsymbol{y}_n, \tilde{\boldsymbol{x}}_{-n}\right) - u_n\left(\tilde{\boldsymbol{x}}\right) \leq \frac{1}{n}$ for all n, so $\max\limits_{\boldsymbol{y}_n} u_n\left(\boldsymbol{y}_n, \tilde{\boldsymbol{x}}_{-n}\right) \leq u_n\left(\tilde{\boldsymbol{x}}\right)$ and $\tilde{\boldsymbol{x}} \in \mathcal{N}_0$. Since $\mathcal{N}_{\varepsilon_1} \subseteq \mathcal{N}_{\varepsilon_2}$ if $\varepsilon_2 \geq \varepsilon_1$ then $d_{\mathcal{N}}\left(\varepsilon\right)$ is non-increasing. Additionally, $d_{\mathcal{N}}\left(\varepsilon\right)$ is bounded since $\mathcal{A}^N$ is bounded. Hence, $\lim\limits_{n \to \infty} d_{\mathcal{N}}\left(\frac{1}{n}\right)$ exists. However, since $\tilde{\boldsymbol{x}}_{n_k} \in \mathcal{N}_{\frac{1}{n_k}}$ and $\tilde{\boldsymbol{x}} \in \mathcal{N}_0$ then

$$d_{\mathcal{N}}\left(\frac{1}{n_k}\right) = \max_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_{\frac{1}{n_k}}} \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\right\| = \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \left\|\tilde{\boldsymbol{x}}_{n_k} - \boldsymbol{x}^*\right\| \leq \left\|\tilde{\boldsymbol{x}}_{n_k} - \tilde{\boldsymbol{x}}\right\| \tag{63}$$

so $\lim\limits_{k \to \infty} d_{\mathcal{N}}\left(\frac{1}{n_k}\right) = 0$ since $\tilde{\boldsymbol{x}}_{n_k} \to \tilde{\boldsymbol{x}}$. Hence, we must have $\lim\limits_{n \to \infty} d_{\mathcal{N}}\left(\frac{1}{n}\right) = 0$ and $\lim\limits_{\varepsilon \to 0} d_{\mathcal{N}}\left(\varepsilon\right) = 0$.

Let $\mathcal{P}\left(\mathcal{A}^N\right)$ be the set of all Borel probability measures over $\mathcal{A}^N$, equipped with the weak-* topology Simmons (1963). By replacing the zero constant of the half-space with ε in the proof of Stoltz and Lugosi (2007, Theorem 9), we conclude that $\mathcal{C}_\varepsilon$ is compact for all $\varepsilon \geq 0$. Note that a CE is CCE when all of the departure functions are constant and therefore continuous. Define $\tilde{\rho}_n \in \arg\max_{\rho_\varepsilon \in \mathcal{C}_{\frac{1}{n}}} \min_{\rho^* \in \mathcal{C}_0} \|\rho_\varepsilon - \rho^*\|$. Then, Prokhorov's Theorem, given as Proposition 8 in Stoltz and Lugosi (2007), states that there exists a subsequence $\tilde{\rho}_{n_k}$ that converges to a point $\tilde{\rho} \in \mathcal{P}\left(\mathcal{A}^{\mathcal{N}}\right)$. Since we must have $\tilde{\rho} \in \bigcap_{n=1}^{\infty} \mathcal{C}_{\frac{1}{n}}$ then $\mathbb{E}^{\boldsymbol{x}^* \sim \tilde{\rho}} \left\{ \max_{\boldsymbol{y}_n} u_n\left(\boldsymbol{y}_n, \boldsymbol{x}^*_{-n}\right) - u_n\left(\boldsymbol{x}^*\right) \right\} = 0$ so $\tilde{\rho} \in \mathcal{C}_0$. Following the same argument as in (63) we conclude that $\lim_{\varepsilon \to 0} d_{\mathcal{C}}(\varepsilon) = 0$. ■

11. Upper Bound on the Effect of the Delays on the Regret

Lemma 8. *Let $\{\eta_t\}$ be a non-increasing step size sequence. Let d_t be the delay of the cost of the action at round t. Let $\mathcal{S}_t$ be the set of feedback samples received and used at round t. Define the set $\mathcal{M}^*$ of all samples that are not received and used before round T. Then*

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \eta_q + \sum_{q \in \mathcal{S}_t, q < s} \eta_q \right) \leq 2 \sum_{t \notin \mathcal{M}^*} \eta_t^2 d_t. \tag{64}$$

Proof The quantity $Q_{s,t} \triangleq \sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \eta_q + \sum_{q \in \mathcal{S}_t, q<s} \eta_q$ is a discounted count of the number of feedback samples received and used between round s and round t, before the feedback from round s was used. We want to upper bound $\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q_{s,t}$ for all possible delay sequences $\{d_t\}$. We do so by (over) counting the contribution of each feedback from the T feedback samples, in the different $Q_{s,t}$ "buckets". There are two possible cases of contributing feedback samples, so we write $Q_{s,t} = Q^1_{s,t} + Q^2_{s,t}$. The first type is a feedback from $q \geq s$ that is received and used before the feedback from round s is used. There are a maximum of d_s feedback samples of this type. Each feedback sample can contribute $\eta_q \leq \eta_s$ with $q \geq s$ (since η_t is non-increasing) to $Q^1_{s,t}$ for $s \in \mathcal{S}_t$. We over count them by giving each $Q^1_{s,t}$ term all of its d_s possible samples of this type. So

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q^1_{s,t} \leq \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 d_s = \sum_{t \notin \mathcal{M}^*} \eta_s^2 d_s. \tag{65}$$

The second type is a feedback sample from $q < s$ that is received and used before s is used. The samples from round q can contribute to a maximum of d_q different $Q^2_{s,t}$ terms, all with $s \geq q$. This follows simply because the feedback sample of q is not received before $q + d_q$. Let Γ_q be the set of rounds s such that the sample from round q contributes to $Q^2_{s,t}$. Then

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q^2_{s,t} \underset{(a)}{=} \sum_{q \notin \mathcal{M}^*} \sum_{s \in \Gamma_q} \eta_s \eta_q \underset{(b)}{\leq} \sum_{q \notin \mathcal{M}^*} \eta_q^2 \left|\Gamma_q\right| \leq \sum_{q \notin \mathcal{M}^*} \eta_q^2 d_q \tag{66}$$

where (a) follows since only rounds q that their feedback is received sometime before round T are counted in $Q^2_{s,t}$ for some s, t. Inequality (b) uses $\eta_s \leq \eta_q$ since η_t is non-increasing and $s \geq q$ for all $s \in \Gamma_q$. Summing (65) and (66) we obtain (64). ■

12. Proof of Lemma 3: Discounted-Regret of FKM with Delays

Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} \sum_{t=1}^{T} \eta_t l_t(\boldsymbol{a})$ and note that $\boldsymbol{a}^*$ is random for an adaptive adversary. We have

$$\sum_{t=1}^{T} \eta_t \mathbb{E}\left\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right\} = \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(l_s(\boldsymbol{a}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \mathbb{E}\left\{\sum_{t\in\mathcal{M}} \eta_t\left(l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right)\right\}$$
$$\underset{(a)}{\leq} \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(l_s(\boldsymbol{a}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \sum_{t\in\mathcal{M}} \eta_t$$
$$\underset{(b)}{\leq} \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \sum_{t\in\mathcal{M}} \eta_t + \delta L \sum_{t\notin\mathcal{M}} \eta_t \quad (67)$$

where (a) uses $l_t(\boldsymbol{a}) \in [0,1]$ and (b) uses $|l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}_s)| \leq L\|\boldsymbol{x}_s - \boldsymbol{a}_s\| \leq \delta L$.

Define s_-, s_+ as the step a moment before and a moment after the algorithm uses the feedback from round s, which updates the strategy from $\boldsymbol{a}_{s_-}$ to $\boldsymbol{a}_{s_+}$. Both s_- and s_+ are taking place in round t if $s \in \mathcal{S}_t$. Define the projection $\boldsymbol{a}_\delta^* = \prod_{\mathcal{K}_\delta}(\boldsymbol{a}^*)$. We bound the first term in (67) as follows

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}^*)\right) \underset{(a)}{\leq} L|\mathcal{K}|\delta \sum_{t\notin\mathcal{M}} \eta_t + \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}_\delta^*)\right)$$
$$\underset{(b)}{\leq} (2+|\mathcal{K}|)L\delta \sum_{t\notin\mathcal{M}} \eta_t + \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s\left(\hat{l}_s(\boldsymbol{x}_s) - \hat{l}_s(\boldsymbol{a}_\delta^*)\right) \quad (68)$$

where (a) follows from the Lipschitz continuity of l_s since $\left\|\boldsymbol{a}^* - \prod_{\mathcal{K}_\delta}(\boldsymbol{a}^*)\right\| \leq \|\boldsymbol{a}^* - (1-\delta)\boldsymbol{a}^*\| \leq \delta|\mathcal{K}|$. Inequality (b) uses the Lipschitz continuity again as follows:

$$\hat{l}_s(\boldsymbol{x}) - l_s(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{u}\in\mathbb{B}}\left\{l_s(\boldsymbol{x}+\delta\boldsymbol{u}) - l_s(\boldsymbol{x})\right\} \leq \delta L \mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\left\{\|\boldsymbol{u}\|\right\} = \delta L \quad (69)$$

which we apply on $l_s(\boldsymbol{x}_s)$ and on $l_s(\boldsymbol{a}_\delta^*)$. Now recall that $g_t = \frac{n}{\delta} l_t(\boldsymbol{x}_t + \delta\boldsymbol{u}_t)\boldsymbol{u}_t$ where $\boldsymbol{u}_t$ is on the unit sphere $\mathbb{S}_1$, and define

$$h_t(\boldsymbol{x}) = \hat{l}_t(\boldsymbol{x}) + \left(\boldsymbol{g}_t - \nabla\hat{l}_t(\boldsymbol{x}_t)\right)^T \boldsymbol{x} \quad (70)$$

for which $\nabla h_t(\boldsymbol{x}_t) = \boldsymbol{g}_t$ and $\mathbb{E}\{h_t(\boldsymbol{x}_t)\} = \mathbb{E}\left\{\hat{l}_t(\boldsymbol{x}_t)\right\}$ (see Lemma 2). Next note that

$$\left\|\boldsymbol{x}_{s_+} - \boldsymbol{a}_\delta^*\right\|^2 = \left\|\prod_{\mathcal{K}_\delta}\left(\boldsymbol{x}_{s_-} - \eta_s\nabla h_s(\boldsymbol{x}_s)\right) - \boldsymbol{a}_\delta^*\right\|^2 \underset{(a)}{\leq} \left\|\left(\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right) - \eta_s\nabla h_s(\boldsymbol{x}_s)\right\|^2$$
$$= \left\|\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right\|^2 - 2\eta_s\left\langle\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*, \nabla h_s(\boldsymbol{x}_s)\right\rangle + \eta_s^2\|\nabla h_s(\boldsymbol{x}_s)\|^2 = \left\|\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right\|^2 + \eta_s^2\|\nabla h_s(\boldsymbol{x}_s)\|^2$$
$$- 2\eta_s\left\langle\boldsymbol{x}_s - \boldsymbol{a}_\delta^*, \nabla h_s(\boldsymbol{x}_s)\right\rangle - 2\eta_s\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right) + \sum_{q\in\mathcal{S}_t, q<s}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right), \nabla h_s(\boldsymbol{x}_s)\right\rangle$$
$$\underset{(b)}{\leq} \left\|\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right\|^2 - 2\eta_s\left(h_s(\boldsymbol{x}_s) - h_s(\boldsymbol{a}_\delta^*)\right) + \eta_s^2\|\boldsymbol{g}_s\|^2$$
$$- 2\eta_s\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right) + \sum_{q\in\mathcal{S}_t, q<s}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right), \nabla h_s(\boldsymbol{x}_s)\right\rangle \quad (71)$$

where (a) follows since $\prod_{\mathcal{K}_\delta}$ is the projection of $\boldsymbol{x}_{s_-} - \eta_s\nabla h_s(\boldsymbol{x}_s)$ onto the convex $\mathcal{K}_\delta$. Inequality (b) uses the convexity and differentiability of h_s on $\mathcal{K}_\delta$, so $h_s(\boldsymbol{a}_\delta^*) \geq h_s(\boldsymbol{x}_s) + \langle\boldsymbol{a}_\delta^* - \boldsymbol{x}_s, \nabla h_s(\boldsymbol{x}_s)\rangle$.

12.1 Adaptive Adversary

First note that for any $\boldsymbol{x} \in \mathcal{K}$

$$\left|\sum_{t \notin \mathcal{M}} \eta_t \left(\hat{l}_t(\boldsymbol{x}) - h_t(\boldsymbol{x})\right)\right| = \left|\sum_{t \notin \mathcal{M}} \eta_t \left\langle \boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t), \boldsymbol{x}\right\rangle\right| \\ \leq \|\boldsymbol{x}\| \left\|\sum_{t \notin \mathcal{M}} \eta_t \left(\boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t)\right)\right\| \leq |\mathcal{K}| \left\|\sum_{t \notin \mathcal{M}} \eta_t \left(\boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t)\right)\right\|. \quad (72)$$

The expectation of the last term can be bounded as follows

$$\mathbb{E}^2\left\{\left\|\sum_{t \notin \mathcal{M}} \eta_t \left(\boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t)\right)\right\|\right\} \leq \mathbb{E}\left\{\left\|\sum_{t \notin \mathcal{M}} \eta_t \left(\boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t)\right)\right\|^2\right\} \\ = \sum_{t \notin \mathcal{M}} \eta_s^2 \mathbb{E}\left\{\left\|\boldsymbol{g}_t - \nabla \hat{l}_t(\boldsymbol{x}_t)\right\|^2\right\} + \sum_{t_1=1}^{T} \sum_{t_1 \neq t_2} \eta_{t_1}\eta_{t_2} \mathbb{E}\left\{\left\langle \boldsymbol{g}_{t_1} - \nabla \hat{l}_{t_1}(\boldsymbol{x}_{t_1}), \boldsymbol{g}_{t_2} - \nabla \hat{l}_{t_2}(\boldsymbol{x}_{t_2})\right\rangle\right\} \underset{(a)}{\leq} \\ 2\sum_{t \notin \mathcal{M}} \eta_t^2 \mathbb{E}\left\{\|\boldsymbol{g}_t\|^2 + \left\|\nabla \hat{l}_t(\boldsymbol{x}_t)\right\|^2\right\} \leq 2\sum_{t \notin \mathcal{M}} \eta_t^2 \left(\frac{n^2}{\delta^2} + L^2\right) \quad (73)$$

where (a) follows since for $t_1 < t_2$ we get from Lemma 2 that

$$\left\langle \boldsymbol{g}_{t_1} - \nabla \hat{l}_{t_1}(\boldsymbol{x}_{t_1}), \mathbb{E}\left\{\boldsymbol{g}_{t_2} - \nabla \hat{l}_{t_2}(\boldsymbol{x}_{t_2}) \mid \sigma\left(\{\boldsymbol{x}_{t_2}, \boldsymbol{x}_{t_1}, \boldsymbol{u}_{t_1}\}\right)\right\}\right\rangle = 0. \quad (74)$$

Then

$$\mathbb{E}\left\{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\hat{l}_s(\boldsymbol{x}_s) - \hat{l}_s(\boldsymbol{a}_\delta^*)\right)\right\} \underset{(a)}{\leq} \\ \mathbb{E}\left\{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(h_s(\boldsymbol{x}_s) - h_s(\boldsymbol{a}_\delta^*)\right)\right\} + |\mathcal{K}| \sqrt{2\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 \left(\frac{n^2}{\delta^2} + L^2\right)} \\ \underset{(b)}{\leq} \frac{1}{2}\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \mathbb{E}\left\{\left\|\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right\|^2 - \left\|\boldsymbol{x}_{s_+} - \boldsymbol{a}_\delta^*\right\|^2\right\} + \frac{1}{2}\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 \mathbb{E}\left\{\|\boldsymbol{g}_s\|^2\right\} + |\mathcal{K}| \sqrt{2\sum_{t \notin \mathcal{M}} \eta_t^2 \left(\frac{n^2}{\delta^2} + L^2\right)} \\ -\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \mathbb{E}\left\{\left\langle \sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right) + \sum_{q \in \mathcal{S}_t, q < s} \left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right), \mathbb{E}\left\{\nabla h_s(\boldsymbol{x}_s) \mid \mathcal{F}_{s_-}\right\}\right\rangle\right\} \\ \underset{(c)}{\leq} \frac{|\mathcal{K}|^2}{2} + \frac{n^2}{2\delta^2} \sum_{t \notin \mathcal{M}} \eta_t^2 + |\mathcal{K}| \sqrt{2\sum_{t \notin \mathcal{M}} \eta_t^2 \left(\frac{n^2}{\delta^2} + L^2\right)} \\ + \frac{n}{\delta}\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\} + \sum_{q \in \mathcal{S}_t, q < s} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\}\right) \quad (75)$$

where (a) uses (72) and (73) on $\boldsymbol{x} = \boldsymbol{a}_\delta^*$ and $\mathbb{E}\{h_s(\boldsymbol{x}_s)\} = \mathbb{E}\left\{\hat{l}_s(\boldsymbol{x}_s)\right\}$, (b) uses (71) and (c) uses the telescopic sum with $\|\boldsymbol{x}_0 - \boldsymbol{a}_\delta^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{a}_\delta^*\|^2 \leq |\mathcal{K}|^2$, that $\boldsymbol{g}_s = \frac{n}{\delta} l_s(\boldsymbol{x}_s + \delta \boldsymbol{u}_s)\boldsymbol{u}_s$, and also Cauchy-Schwarz with Lemma 2:

$$\left\|\mathbb{E}\left\{\nabla h_s(\boldsymbol{x}_s) \mid \mathcal{F}_{s_-}\right\}\right\| \leq \mathbb{E}\left\{\|\boldsymbol{g}_s\| \mid \mathcal{F}_{s_-}\right\} \leq \frac{n}{\delta}. \quad (76)$$

Finally we bound the last term by bounding

$$\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|_2 = \left\|\prod_{\mathcal{K}_\delta}\left(\boldsymbol{x}_{q_-} - \eta_q \boldsymbol{g}_q\right) - \boldsymbol{x}_{q_-}\right\|_2 \underset{(a)}{\leq} \eta_q \left\|\boldsymbol{g}_q\right\|_2 \leq n\frac{\eta_q}{\delta} \tag{77}$$

where (a) follows since $\prod_{\mathcal{K}_\delta}$ is the projection of $\boldsymbol{x}_{q_-} - \eta_q \boldsymbol{g}_q$ onto the convex $\mathcal{K}_\delta$. We obtain

$$\frac{n}{\delta}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\} + \sum_{q\in\mathcal{S}_t, q<s}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\}\right)$$
$$\leq \frac{n^2}{\delta^2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q + \sum_{q\in\mathcal{S}_t, q<s}\eta_q\right) \underset{(a)}{\leq} 2\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}}\eta_t^2 d_t \tag{78}$$

where (a) uses Lemma 8. We conclude by applying (78) on (75) and adding (68) and (67).

12.2 Oblivious Adversary

For an oblivious adversary, l_q for $q > s$ is not a random variable that depends on $\boldsymbol{a}_s$, so by Lemma 2

$$\left\|\mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right) \mid \mathcal{F}_{s_-}\right\}\right\| \underset{(a)}{=} \left\|\mathbb{E}\left\{\frac{n}{\delta} l_s\left(\boldsymbol{x}_s + \delta \boldsymbol{u}_s\right)\boldsymbol{u}_s \mid \mathcal{F}_{s_-}\right\}\right\| = \left\|\nabla \hat{l}_s\left(\boldsymbol{x}_s\right)\right\| \underset{(b)}{\leq} L \tag{79}$$

where (a) uses (70) and (b) follows since $\hat{l}_s$ is differentiable and Lipschitz continuous with parameter L. Then

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{x}_s\right) - \hat{l}_s\left(\boldsymbol{a}_\delta^*\right)\right\} \underset{(a)}{=} \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\mathbb{E}\left\{h_s\left(\boldsymbol{x}_s\right) - h_s\left(\boldsymbol{a}_\delta^*\right)\right\}$$
$$\underset{(b)}{\leq} \frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\mathbb{E}\left\{\left\|\boldsymbol{x}_{s_-} - \boldsymbol{a}_\delta^*\right\|^2 - \left\|\boldsymbol{x}_{s_+} - \boldsymbol{a}_\delta^*\right\|^2\right\} + \frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\mathbb{E}\left\{\left\|\boldsymbol{g}_s\right\|^2\right\} -$$
$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\mathbb{E}\left\{\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right) + \sum_{q\in\mathcal{S}_t, q<s}\left(\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right), \mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right) \mid \mathcal{F}_{s_-}\right\}\right\rangle\right\}$$
$$\underset{(c)}{\leq} \frac{\left|\mathcal{K}\right|^2}{2} + \frac{n^2}{2\delta^2}\sum_{t\notin\mathcal{M}}\eta_t^2 + L\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\} + \sum_{q\in\mathcal{S}_t, q<s}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\}\right) \tag{80}$$

where (a) uses $\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{x}_s\right)\right\} = \mathbb{E}\left\{h_s\left(\boldsymbol{x}_s\right)\right\}$ and $\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{a}_\delta^*\right)\right\} = \mathbb{E}\left\{h_s\left(\boldsymbol{a}_\delta^*\right)\right\}$, (b) uses (71) and (c) uses the telescopic sum with $\left\|\boldsymbol{x}_0 - \boldsymbol{a}_\delta^*\right\|^2 - \left\|\boldsymbol{x}_T - \boldsymbol{a}_\delta^*\right\|^2 \leq \left|\mathcal{K}\right|^2$, $\left\|\boldsymbol{g}_s\right\| \leq \frac{n}{\delta}$, and Cauchy-Schwarz with (79). Finally, we use (77) to bound the last term in (80):

$$L\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\} + \sum_{q\in\mathcal{S}_t, q<s}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+} - \boldsymbol{x}_{q_-}\right\|\right\}\right)$$
$$\leq \frac{Ln}{\delta}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q + \sum_{q\in\mathcal{S}_t, q<s}\eta_q\right) \underset{(a)}{\leq} 2L\frac{n}{\delta}\sum_{t\notin\mathcal{M}}\eta_t^2 d_t \tag{81}$$

where (a) uses Lemma 8. We conclude by applying (81) on (80) and adding (68) and (67).

13. Proof of Lemma 6: Discounted-Regret of EXP3 with Delays

For each i^*, and for the set of missing or discarded samples $\mathcal{M}^*$

$$\sum_{t=1}^{T}\eta_t\left(l_t^{(a_t)}-l_t^{(i^*)}\right)\underset{(a)}{\leq}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(l_s^{(a_s)}-l_s^{(i^*)}\right)+\sum_{t\in\mathcal{M}^*}\eta_t \tag{82}$$

where (a) uses that $0\leq l_t^{(i)}\leq 1$ for every i and t. Let s_-,s_+ as the step a moment before and a moment after the algorithm uses the feedback from round s, which updates the mixed strategy from $\boldsymbol{p}_{s_-}$ to $\boldsymbol{p}_{s_+}$. Both s_- and s_+ are part of round t if $s\in\mathcal{S}_t$. Let s_T be the last feedback to be updated.

We begin by the standard EXP3 analysis for arbitrary $\tilde{\boldsymbol{l}}_s$ (Lattimore and Szepesvári, 2020), but with careful consideration to both the interleaved arrivals and the discount factor η_s.

Define $\Phi(t)=-\ln\left(\sum_{i=1}^{K}e^{-\tilde{L}_t^{(i)}}\right)$ and $\tilde{\boldsymbol{l}}_t=\left(0,...,\frac{l_t^{(a_t)}}{p_t^{(a_t)}+\gamma_t},...,0\right)$. Then

$$\begin{aligned}\Phi(s_+)-\Phi(s_-)&=-\ln\left(\frac{\sum_{i=1}^{K}e^{-\tilde{L}_{s_-}^{(i)}}e^{-\eta_s\tilde{l}_s^{(i)}}}{\sum_{j=1}^{K}e^{-\tilde{L}_{s_-}^{(j)}}}\right)=-\ln\left(\sum_{i=1}^{K}p_{s_-}^{(i)}e^{-\eta_s\tilde{l}_s^{(i)}}\right)\\&\underset{(a)}{\geq}-\ln\left(\sum_{i=1}^{K}p_{s_-}^{(i)}\left(1-\eta_s\tilde{l}_s^{(i)}+\frac{1}{2}\eta_s^2\left(\tilde{l}_s^{(i)}\right)^2\right)\right)=-\ln\left(1-\sum_{i=1}^{K}p_{s_-}^{(i)}\left(\eta_s\tilde{l}_s^{(i)}-\frac{1}{2}\eta_s^2\left(\tilde{l}_s^{(i)}\right)^2\right)\right)\\&\underset{(b)}{\geq}\eta_s\sum_{i=1}^{K}p_{s_-}^{(i)}\tilde{l}_s^{(i)}-\frac{\eta_s^2}{2}\sum_{i=1}^{K}p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2\end{aligned} \tag{83}$$

where (a) is $e^{-x}\leq 1-x+\frac{1}{2}x^2$ and (b) is $\ln(1-x)\leq -x$. Hence, iterating (83) over s yields

$$\Phi\left(s_T^+\right)-\Phi(1)=\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\Phi(s_+)-\Phi(s_-)\right)\geq\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\sum_{i=1}^{K}p_{s_-}^{(i)}\tilde{l}_s^{(i)}-\frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2. \tag{84}$$

Now we upper bound $\Phi\left(s_T^+\right)-\Phi(1)$. We have for every i

$$\Phi\left(s_T^+\right)-\Phi(1)=-\ln\left(\sum_{j=1}^{K}e^{-\tilde{L}_{s_T^+}^{(j)}}\right)+\ln K\underset{(a)}{\leq}\tilde{L}_{s_T^+}^{(i)}+\ln K=\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\tilde{l}_s^{(i)}+\ln K \tag{85}$$

where (a) omits positive terms from $\sum_{i=1}^{K}e^{-\tilde{L}_{s_T^+}^{(i)}}$. We conclude from (84) and (85) that

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\sum_{i=1}^{K}p_{s_-}^{(i)}\tilde{l}_s^{(i)}-\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\tilde{l}_s^{(i)}\leq\ln K+\frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2. \tag{86}$$

Now observe that

$$\begin{aligned}\sum_{i=1}^{K}p_{s_-}^{(i)}\tilde{l}_s^{(i)}&=\sum_{i=1}^{K}p_{s_-}^{(i)}\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}=\sum_{i=1}^{K}\left(p_s^{(i)}+\gamma_s\right)\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}-\sum_{i=1}^{K}\left(p_s^{(i)}+\gamma_s-p_{s_-}^{(i)}\right)\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\\&=l_s^{(a_s)}-\gamma_s\sum_{i=1}^{K}\tilde{l}_s^{(i)}+\sum_{i=1}^{K}\left(p_{s_-}^{(i)}-p_s^{(i)}\right)\tilde{l}_s^{(i)}.\end{aligned} \tag{87}$$

Using $\frac{p_{s_-}^{(i)}}{p_s^{(i)}} \leq e^2$ from Lemma 9, we obtain

$$\sum_{i=1}^{K} p_{s_-}^{(i)} \left(\tilde{l}_s^{(i)}\right)^2 = \sum_{i=1}^{K} p_{s_-}^{(i)} \left(\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)} + \gamma_s}\right)^2 \leq e^2 \sum_{i=1}^{K} \frac{p_s^{(i)}}{p_s^{(i)} + \gamma_s} \frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)} + \gamma_s} \leq e^2 \sum_{i=1}^{K} \tilde{l}_s^{(i)}. \tag{88}$$

Taking the discounted sum of (87) over $t \notin \mathcal{M}^*$ and subtracting $\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \tilde{l}_s^{(i)}$ from both sides:

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s l_s^{(a_s)} - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \gamma_s \eta_s \sum_{i=1}^{K} \tilde{l}_s^{(i)} + \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \sum_{i=1}^{K} \left(p_{s_-}^{(i)} - p_s^{(i)}\right) \tilde{l}_s^{(i)} - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \tilde{l}_s^{(i)} \\
&\quad = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \sum_{i=1}^{K} p_{s_-}^{(i)} \tilde{l}_s^{(i)} - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \tilde{l}_s^{(i)} \underset{(a)}{\leq} \ln K + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 \sum_{i=1}^{K} p_{s_-}^{(i)} \left(\tilde{l}_s^{(i)}\right)^2 \\
&\quad \underset{(b)}{\leq} \ln K + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} e^2 \eta_s^2 \sum_{i=1}^{K} \tilde{l}_s^{(i)}
\end{aligned} \tag{89}$$

where (a) uses (86) and (b) uses (88). Rearranging (89) and subtracting $\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s l_s^{(i)}$ from both sides gives

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(l_s^{(a_s)} - l_s^{(i)}\right) \\
&\leq \underbrace{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \sum_{i=1}^{K} \left(p_s^{(i)} - p_{s_-}^{(i)}\right) \tilde{l}_s^{(i)}}_{A} + \ln K + \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right) + \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \left(\gamma_s \eta_s + \frac{e^2}{2} \eta_s^2\right) \sum_{i=1}^{K} \tilde{l}_s^{(i)}.
\end{aligned} \tag{90}$$

13.1 Adaptive Adversary with $\gamma_t = \eta_t$

13.1.1 Taking the Expectation

Define $W_1^{(i)} = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right) - \ln K$. From Lemma 11 with $\delta \leftarrow \frac{\delta}{K}$ and $\alpha_s^{(i)} = \eta_s$ and $\alpha_s^{(j)} = 0$ for all $j \neq i$ we get by using the union bound that

$$\mathbb{P}\left(\max_i W_1^{(i)} \geq \ln \frac{1}{\delta}\right) \leq K \mathbb{P}\left(\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right) \geq \ln \frac{K}{\delta}\right) \leq \delta \tag{91}$$

so by substituting $x = \ln \frac{1}{\delta}$ so $dx = -\frac{d\delta}{\delta}$

$$\mathbb{E}\left\{\max_i W_1^{(i)}\right\} \leq \int_0^\infty \mathbb{P}\left(\max_i W_1^{(i)} \geq x\right) dx = \int_0^1 \frac{1}{\delta} \mathbb{P}\left(\max_i W_1^{(i)} \geq \ln \frac{1}{\delta}\right) d\delta \leq 1. \tag{92}$$

Define $W_2 = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \left(\gamma_s \eta_s + \frac{e^2}{2} \eta_s^2\right) \sum_{i=1}^{K} \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right)$. From Lemma 11 with $\alpha_s^{(i)} = \gamma_s \eta_s + \frac{e^2}{2} \eta_s^2 \leq 2\eta_s$ for all i we get $\mathbb{P}\left(W_2 \geq \ln \frac{1}{\delta}\right) \leq \delta$, so using (92) on W_2 we obtain $\mathbb{E}\left\{W_2\right\} \leq 1$.

From $\mathbb{E}\left\{\max_i W_1^{(i)}\right\} \leq 1$ and $\mathbb{E}\left\{W_2\right\} \leq 1$ we conclude that

$$\mathbb{E}\left\{\max_i \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\tilde{l}_s^{(i)}-l_s^{(i)}\right)+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right\}$$
$$\leq \ln K+2+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}l_s^{(i)} \leq \ln K+2+K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right). \quad (93)$$

13.1.2 The Effect of Delays

Next we bound the A term in (90), which quantifies the effect of the delays on the regret. Define $h_i\left(\tilde{\boldsymbol{L}}_{q_-}\right) \triangleq p_{q_-}^{(i)} = \frac{e^{-\tilde{L}_{q_-}^{(i)}}}{\sum_{j=1}^{K}e^{-\tilde{L}_{q_-}^{(j)}}}$, so $p_{q^+}^{(i)} = h_i\left(\tilde{\boldsymbol{L}}_{q_-}+\eta_q\tilde{\boldsymbol{l}}_q\right)$. Using Lemma 10 with $\boldsymbol{x} = \tilde{\boldsymbol{L}}_{q_-}$ and $\Delta = \eta_q\tilde{\boldsymbol{l}}_q$ (so $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{p}_{q_-}$) yields

$$\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1 \leq 2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}\tilde{l}_q^{(i)} = 2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}\frac{l_q^{(i)}1_{\{a_q=i\}}}{p_q^{(i)}+\gamma_q}$$
$$\underset{(a)}{\leq} 2e^2\eta_q\sum_{i=1}^{K}p_q^{(i)}\frac{l_q^{(i)}1_{\{a_q=i\}}}{p_q^{(i)}+\gamma_q} \leq 2e^2\eta_q\sum_{i=1}^{K}l_q^{(i)}1_{\{a_q=i\}} = 2e^2\eta_q l_q^{(a_q)} \leq 2e^2\eta_q \quad (94)$$

where (a) uses $\frac{p_{q_-}^{(i)}}{p_q^{(i)}} \leq e^2$ from Lemma 9. Hence

$$\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle = \sum_{i=1}^{K}\left(p_{q_-}^{(i)}-p_{q_+}^{(i)}\right)\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} \underset{(a)}{\leq} 2e^2\eta_q\sum_{i=1}^{K}\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}. \quad (95)$$

where (a) follows since $p_{q_-}^{(i)}-p_{q_+}^{(i)} \leq \left|p_{q_-}^{(i)}-p_{q_+}^{(i)}\right| \leq \left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1 \leq 2e^2\eta_q$. Using (95) we can write

$$\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle \mid \mathcal{F}_s\right\} \leq 2e^2\eta_q\sum_{i=1}^{K}\mathbb{E}\left\{\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\mid\mathcal{F}_s\right\} \underset{(a)}{=} 2e^2\eta_q\sum_{i=1}^{K}\frac{p_s^{(i)}}{p_s^{(i)}+\gamma_s}l_s^{(i)} \leq 2e^2\eta_q K. \quad (96)$$

where (a) uses that $l_s^{(i)}, p_s^{(i)} \in \mathcal{F}_s$ and that $a_s = i$ with probability $p_s^{(i)}$. Hence

$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_s-\boldsymbol{p}_{s_-}\right\rangle\right\} = \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_t-\boldsymbol{p}_{s_-}\right\rangle+\sum_{r=s}^{t-1}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_r-\boldsymbol{p}_{r+1}\right\rangle\right)\right\}$$
$$= \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right)\right\}$$
$$= \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}\right)$$
$$\underset{(a)}{\leq} 2e^2K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\eta_q+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q\right) \underset{(b)}{\leq} 4e^2K\sum_{t\notin\mathcal{M}^*}\eta_t^2 d_t \quad (97)$$

where (a) uses (96) and (b) uses Lemma 8.

13.1.3 Concluding the Proof

We obtain

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T}\eta_t\left(l_t^{(a_t)}-l_t^{(i^*)}\right)\right\} &= \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(l_s^{(a_s)}-l_s^{(i^*)}\right)\right\}+\sum_{t\in\mathcal{M}^*}\eta_t \\
&\underset{(a)}{\leq} \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\sum_{i=1}^{K}\left(p_s^{(i)}-p_{s_-}^{(i)}\right)\tilde{l}_s^{(i)}\right\}+\ln K \\
&+\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\eta_s\left(\tilde{l}_s^{(i)}-l_s^{(i^*)}\right)+\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right)\right\}+\sum_{t\in\mathcal{M}^*}\eta_t \\
&\underset{(b)}{\leq} 2+\left(1+\frac{e^2}{2}\right)K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2+2\ln K+4e^2K\sum_{t\notin\mathcal{M}^*}\eta_t^2 d_t+\sum_{t\in\mathcal{M}^*}\eta_t \qquad (98)
\end{aligned}$$

where (a) is (90) and (b) is (93) and (97) for $\gamma_s=\eta_s$.

13.2 Oblivious Adversary with $\gamma_t = 0$

13.2.1 Taking the Expectation

With an oblivious adversary $\mathbb{E}\left\{\tilde{l}_s^{(i)}\right\} = l_s^{(i)}\mathbb{E}\left\{\frac{1_{\{a_s=i\}}}{p_s^{(i)}}\right\} = l_s^{(i)}$ for each s and i, since $l_s^{(i)}$ is not random. Then

$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\tilde{l}_s^{(i)}-l_s^{(i)}\right)+\frac{e^2}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right\}=\frac{e^2}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}l_s^{(i)}\leq 4K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2. \qquad (99)$$

13.2.2 The Effect of Delays

Using Lemma 10 with $\boldsymbol{x}=\tilde{\boldsymbol{L}}_{q_-}$ and $\Delta=\eta_q\tilde{\boldsymbol{l}}_q$, so $\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{p}_{q_-}$ yields

$$\begin{aligned}
\mathbb{E}\left\{\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1 \mid \mathcal{F}_{q_-}\right\} &\leq 2\eta_q\mathbb{E}\left\{\sum_{i=1}^{K}p_{q_-}^{(i)}\tilde{l}_q^{(i)}\mid\mathcal{F}_{q_-}\right\} \\
&\underset{(a)}{=} 2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}\mathbb{E}\left\{\tilde{l}_q^{(i)}\mid\mathcal{F}_{q_-}\right\}\underset{(b)}{=}2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}l_q^{(i)}\leq 2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}=2\eta_q \qquad (100)
\end{aligned}$$

where (a) uses $p_{q_-}^{(i)}\in\mathcal{F}_{q_-}$ and (b) uses $p_q^{(i)}\in\mathcal{F}_{q_-}$ (since $q<q_-$) together with the fact that $\tilde{l}_q^{(i)}$ is $\frac{l_q^{(i)}}{p_q^{(i)}}$ with probability $p_q^{(i)}$ and zero otherwise. Note that a_q given $\mathcal{F}_q$ is independent of $\mathcal{F}_{q_-}$ since by definition the feedback from a_q was not received until round q_-. This is unique to the oblivious adversary case when $\boldsymbol{l}_r$ for $r>s$ is not a random variable that depends on a_q, which the adversary observes already at the end of round q. Then for every $q\in\mathcal{S}_r$ for $r<t$ or $q\in\mathcal{S}_t$ such that $q<s$ we have

$$\begin{aligned}
\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)}-p_{q_+}^{(i)}\right)\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}}\mid\mathcal{F}_{s_-}\right\}\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)}-p_{q_+}^{(i)}\right)\mathbb{E}\left\{\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}}\mid\mathcal{F}_{s_-}\right\}\right\}=\mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)}-p_{q_+}^{(i)}\right)l_s^{(i)}\right\} \\
&= \mathbb{E}\left\{\left\langle\boldsymbol{l}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}\underset{(a)}{\leq}\mathbb{E}\left\{\|\boldsymbol{l}_s\|_\infty\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1\right\}\underset{(b)}{\leq}\mathbb{E}\left\{\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1\right\}\underset{(c)}{\leq}2\eta_q \qquad (101)
\end{aligned}$$

where (a) is Hölder's inequality, (b) uses $l_t^{(i)} \leq 1$ and (c) uses (100) and the tower rule. Therefore

$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_s-\boldsymbol{p}_{s_-}\right\rangle\right\}=\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_t-\boldsymbol{p}_{s_-}\right\rangle+\sum_{r=s}^{t-1}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_r-\boldsymbol{p}_{r+1}\right\rangle\right)\right\}$$
$$=\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}\right)$$
$$\underset{(a)}{\leq}2\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\eta_q+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q\right)\underset{(b)}{\leq}4\sum_{t\notin\mathcal{M}^*}\eta_t^2 d_t \quad (102)$$

where (a) follows from (101) and (b) follows from Lemma 8.

13.2.3 Concluding the Proof

Combining (82),(90), (99) and (102) yields

$$\mathbb{E}\left\{\sum_{t=1}^{T}\eta_t l_t^{(a_t)}-\min_i\sum_{t=1}^{T}\eta_t\sum_{i=1}^{K}l_t^{(i)}\right\}\leq\ln K+4K\sum_{t=1}^{T}\eta_t^2+4\sum_{t\notin\mathcal{M}^*}\eta_t^2 d_t+\sum_{t\in\mathcal{M}^*}\eta_t. \quad (103)$$

13.3 EXP3 Auxiliary Lemmas

The following lemma generalizes Lemma 2 from Cesa-Bianchi et al. (2019) to a sequence of delays $\{d_t\}$ and a sequence of step sizes $\{\eta_t\}$.

Lemma 9. *Let $\{\eta_t\}$ be a positive non-increasing step size sequence such that $\eta_t \leq \frac{1}{2}e^{-2}$ for all t. Let $\mathcal{D}=\left\{t\,|\,d_t\geq\frac{1}{e^2\eta_t}-1\right\}$. Then for every s,t such that $s\in\mathcal{S}_t$ (so $s\notin\mathcal{D}$) Algorithm 3 maintains for all $i=1,...,K$ both $\frac{p_{s_+}^{(i)}}{p_{s_-}^{(i)}}\leq\frac{1}{1-e^2\eta_s}$ and $\frac{p_{s_-}^{(i)}}{p_s^{(i)}}\leq e^2$.*

Proof The proof follows by induction on the feedback arrival index. Let s be the first feedback to arrive. Before that, at s_-, we have $p_{s_-}^{(i)}=\frac{1}{K}$ and $\frac{p_s^{(i)}}{p_{s_-}^{(i)}}=1$ for all i. Then, the first update satisfies

$$\frac{p_{s_-}^{(i)}}{p_{s_+}^{(i)}}=\frac{\frac{1}{K}}{\frac{\frac{1}{K}e^{-\eta_s\tilde{l}_s^{(i)}}}{\sum_{j=1}^{K}\frac{1}{K}e^{-\eta_s\tilde{l}_s^{(j)}}}}\geq1-\frac{1}{K}+\frac{1}{K}e^{-\eta_s\frac{l_s^{(a_s)}}{\frac{1}{K}+\gamma_s}}\geq1+\frac{1}{K}\left(e^{-\eta_s K l_s^{(a_s)}}-1\right)\geq1-\eta_s l_s^{(a_s)}\geq1-e^2\eta_s. \quad (104)$$

Now let s be any arbitrary round for which the feedback arrives at time t. According to the inductive hypothesis, we have $\frac{p_{q_+}^{(i)}}{p_{q_-}^{(i)}}\leq\frac{1}{1-e^2\eta_s}$ for all $q\in\{r\in\mathcal{S}_t,r<s\}\cup\left\{\bigcup_{r=s}^{t-1}\mathcal{S}_r\right\}$. Define s_0 as the minimal q such that $s\leq q+d_q\leq t$ and $q\notin\mathcal{D}$ (if it exists). Then for all $i=1,...,K$

$$\frac{p_{s_-}^{(i)}}{p_s^{(i)}}=\prod_{r=s}^{t-1}\prod_{q\in\mathcal{S}_r}\prod_{q\in\mathcal{S}_t,q<s}\frac{p_{q_+}^{(i)}}{p_{q_-}^{(i)}}\underset{(a)}{\leq}\prod_{r=s}^{t-1}\prod_{q\in\mathcal{S}_r}\prod_{q\in\mathcal{S}_t,q<s}\left(1+\frac{e^2\eta_q}{1-e^2\eta_q}\right)$$
$$\underset{(b)}{\leq}\left(1+\frac{1}{e^{-2}\eta_{s_0}^{-1}-1}\right)^{d_{s_0}}\left(1+\frac{1}{e^{-2}\eta_s^{-1}-1}\right)^{d_s}\underset{(c)}{\leq}e^2 \quad (105)$$

where (a) uses the inductive hypothesis and (c) uses that by definition $d_{s_0}\leq e^{-2}\eta_{s_0}^{-1}-1$ and $d_s\leq e^{-2}\eta_s^{-1}-1$. If s_0 does not exist then the first factor is one (i.e., $d_{s_0}=0$). Inequality (b) uses that the product runs over all rounds $q \notin \mathcal{D}$ for which the feedback is received between s and s_-. Feedback from $q \in \mathcal{D}$ is discarded and has no effect on $p_{q+}^{(a_s)}$. The received feedback includes no more than d_{s_0} samples of rounds before s. This follows since there are at most d_{s_0} rounds between s_0 and s (since $s \le s_0 + d_{s_0}$ by definition), and each of them contributes at most one feedback that is received between s and s_-. We have $\eta_q \le \eta_{s_0}$ for each such round q, since η_t is non-increasing. It also includes no more than d_s feedback samples of rounds after s, since all these feedback samples are received before s_-, which occurs during round $t = s + d_s$. We have $\eta_r \le \eta_s$ for each such round r. We conclude that the update at s_-, occurring at time t using the feedback for a_s, satisfies:

$$\frac{p_{s_-}^{(i)}}{p_{s_+}^{(i)}} = \frac{\frac{e^{-\tilde{L}_{s_-}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}}{\frac{e^{-\tilde{L}_{s_+}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_+}^{(j)}}}} \ge \frac{\frac{e^{-\tilde{L}_{s_-}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}}{\frac{e^{-\tilde{L}_{s_-}^{(i)}} e^{-\eta_s \tilde{l}_s^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} e^{-\eta_s \tilde{l}_s^{(j)}}}} \ge \frac{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} e^{-\eta_s \tilde{l}_s^{(j)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}} \ge \frac{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} \left(1 - \eta_s \tilde{l}_s^{(j)}\right)}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}$$

$$= 1 - \eta_s \sum_{j=1}^K p_{s_-}^{(j)} \tilde{l}_s^{(j)} = 1 - \eta_s p_{s_-}^{(a_s)} \frac{l_s^{(a_s)}}{p_s^{(a_s)} + \gamma_s} \ge 1 - \eta_s \frac{p_{s_-}^{(a_s)}}{p_s^{(a_s)}} \underset{(a)}{\ge} 1 - e^2 \eta_s \quad (106)$$

where (a) follows from (105). Hence $\frac{p_{s_+}^{(i)}}{p_{s_-}^{(i)}} \le \frac{1}{1 - e^2 \eta_s}$ and the proof is complete. ■

The next lemma shows standard smoothness properties of the softmax function, and we provide it here for completeness.

Lemma 10. *Let* $h_i(\boldsymbol{x}) = \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}}$ *and* $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), ..., h_K(\boldsymbol{x}))$. *Then* $\forall \boldsymbol{x} \in \mathbb{R}^K$ *and* $\forall \Delta \in \mathbb{R}_+^K$

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x} + \Delta)\|_1 \le 2 \langle h(\boldsymbol{x}), \Delta \rangle . \quad (107)$$

Proof For all $\boldsymbol{x} \in \mathbb{R}^K$ and $\Delta \in \mathbb{R}_+^K$

$$h_i(\boldsymbol{x} + \Delta) - h_i(\boldsymbol{x}) = \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}} \underset{(a)}{\ge} \left(e^{-\Delta_i} - 1\right) h_i(\boldsymbol{x}) \underset{(b)}{\ge} -\Delta_i h_i(\boldsymbol{x}) \quad (108)$$

where (a) follows since $\sum_{j=1}^K e^{-x_j - \Delta_j} \le \sum_{j=1}^K e^{-x_j}$ and (b) since $1 - x \le e^{-x}$ for all $x \ge 0$. We also have for all $\boldsymbol{x} \in \mathbb{R}^K$ and $\Delta \in \mathbb{R}_+^K$ that

$$h_i(\boldsymbol{x} + \Delta) - h_i(\boldsymbol{x}) = \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}} \underset{(a)}{\le} \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j}} =$$
$$h_i(\boldsymbol{x} + \Delta) \left(1 - \frac{\sum_{j=1}^K e^{-x_j - \Delta_j}}{\sum_{l=1}^K e^{-x_l}}\right) = h_i(\boldsymbol{x} + \Delta) \frac{\sum_{j=1}^K e^{-x_j} \left(1 - e^{-\Delta_j}\right)}{\sum_{l=1}^K e^{-x_l}} \underset{(b)}{\le} h_i(\boldsymbol{x} + \Delta) \frac{\sum_{j=1}^K \Delta_j e^{-x_j}}{\sum_{l=1}^K e^{-x_l}} \quad (109)$$

where (a) uses $e^{-x_j} \ge e^{-x_j - \Delta_j}$ and (b) uses $1 - x \le e^{-x}$ for all $x \ge 0$. Combining (108) and (109):

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x} + \Delta)\|_1 = \sum_{i=1}^K |h_i(\boldsymbol{x}) - h_i(\boldsymbol{x} + \Delta)| \underset{(a)}{\le} \sum_{i=1}^K h_i(\boldsymbol{x} + \Delta) \left(\sum_{j=1}^K \frac{\Delta_j e^{-x_j}}{\sum_{l=1}^K e^{-x_l}}\right)$$
$$+ \sum_{i=1}^K \Delta_i h_i(\boldsymbol{x}) = \left(\sum_{j=1}^K \left(\Delta_j \frac{e^{-x_j}}{\sum_{l=1}^K e^{-x_l}}\right)\right) \sum_{i=1}^K h_i(\boldsymbol{x} + \Delta) + \langle h(\boldsymbol{x}), \Delta \rangle \underset{(b)}{=} 2 \langle h(\boldsymbol{x}), \Delta \rangle \quad (110)$$

where (a) uses $|h_i(\boldsymbol{x}+\Delta) - h_i(\boldsymbol{x})| \leq \max\left\{\Delta_i h_i(\boldsymbol{x}), h_i(\boldsymbol{x}+\Delta)\frac{\sum_{j=1}^K \Delta_j e^{-x_j}}{\sum_{j=1}^K e^{-x_j}}\right\}$ for all i, due to (108) and (109). Equality (b) uses that $\sum_{i=1}^K h_i(\boldsymbol{x}+\Delta) = 1$ by definition. ■

The next Lemma is taken from Neu (2015), and we provide a (very) slightly modified proof to verify that the same result holds even when the order of arrivals changes as a result of the delayed feedback.

Lemma 11. *Let $\tilde{l}_t^{(i)} = \frac{l_t^{(i)} 1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t}$. If $\left\{\alpha_t^{(i)}\right\}$ is a positive sequence such that $\alpha_t^{(i)} \leq 2\gamma_t$ for all t and all i, then*

$$\mathbb{P}\left(\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)} - l_t^{(i)}\right) > \ln\frac{1}{\delta}\right) \leq \delta. \tag{111}$$

Proof Define the filtration $\mathcal{G}_t = \sigma(\{\boldsymbol{a}_\tau \mid \tau < t\})$ and note that $\boldsymbol{l}_t \in \mathcal{G}_t$. We have

$$\tilde{l}_t^{(i)} = \frac{l_t^{(i)} 1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t} \leq \frac{l_t^{(i)} 1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t l_t^{(i)}} = \frac{1}{2\gamma_t}\frac{2\gamma_t \frac{l_t^{(i)}}{p_t^{(i)}} 1_{\{a_t=i\}}}{1+\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}} \underset{(a)}{\leq} \frac{1}{2\gamma_t}\log\left(1+2\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right) \tag{112}$$

where (a) uses $\frac{x}{1+\frac{x}{2}} \leq \log(1+x)$ which holds for $x \geq 0$. Then

$$\begin{aligned}
\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_t^{(i)}\tilde{l}_t^{(i)}} \mid \mathcal{G}_t\right\} &\leq \mathbb{E}\left\{e^{\sum_{i=1}^K \frac{\alpha_t^{(i)}}{2\gamma_t}\log\left(1+2\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right)} \mid \mathcal{G}_t\right\} \\
&\underset{(a)}{\leq} \mathbb{E}\left\{e^{\sum_{i=1}^K \log\left(1+\alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right)} \mid \mathcal{G}_t\right\} = \mathbb{E}\left\{\prod_{i=1}^K\left(1+\alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right) \mid \mathcal{G}_t\right\} \\
&\underset{(b)}{=} \mathbb{E}\left\{1+\sum_{i=1}^K \alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}} \mid \mathcal{G}_t\right\} = 1+\sum_{i=1}^K \alpha_t^{(i)} l_t^{(i)} \leq e^{\sum_{i=1}^K \alpha_t^{(i)} l_t^{(i)}}
\end{aligned} \tag{113}$$

where (a) uses $\alpha_t^{(i)} \leq 2\gamma_t$ and $x\log(1+y) \leq \log(1+xy)$ which holds for $y > -1$ and $0 \leq x \leq 1$. Inequality (b) uses that $1_{\{a_t=i\}}1_{\{a_t=j\}} = 0$ for all $i \neq j$.

Since $l_t^{(i)} \in \mathcal{G}_t$, then (113) yields $\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)} \mid \mathcal{G}_t\right\} \leq 1$ for all i. Let $S = \sum_{t=1}^T |\mathcal{S}_t|$. Let τ_l be the round for which the feedback is the l-th to be received, after arbitrarily arranging all the samples that were received at the same round (so $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_S$). Then

$$\begin{aligned}
&\mathbb{E}\left\{e^{\sum_{l=1}^S\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{e^{\sum_{l=1}^S\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)} \mid \mathcal{G}_{\tau_S}\right\}\right\} \underset{(a)}{=} \\
&\quad \mathbb{E}\left\{e^{\sum_{l=1}^{S-1}\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_{\tau_S}^{(i)}\left(\tilde{l}_{\tau_S}^{(i)}-l_{\tau_S}^{(i)}\right)} \mid \mathcal{G}_{\tau_S}\right\}\right\} \leq \mathbb{E}\left\{e^{\sum_{l=1}^{S-1}\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\right\}
\end{aligned} \tag{114}$$

where (a) uses $a_{\tau_1}, p_{\tau_1}, \ldots, a_{\tau_{S-1}}, p_{\tau_{S-1}} \in \mathcal{G}_{\tau_S}$. Iterating over (114) yields $\mathbb{E}\left\{e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)}\right\} \leq 1$, so by Markov's inequality

$$\begin{aligned}
\mathbb{P}\left(\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)} - l_t^{(i)}\right) > \ln\frac{1}{\delta}\right) &= \mathbb{P}\left(e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)} > \frac{1}{\delta}\right) \\
&\leq \delta\mathbb{E}\left\{e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)}\right\} \leq \delta.
\end{aligned} \tag{115}$$

■