A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Proceedings of Machine Learning Research vol 195:1–37, 2023 36th Annual Conference on Learning Theory

A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Dirk van der Hoeven DIRK@DIRKVANDERHOEVEN.COM
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands

Lukas Zierahn LUKASZIERAHN@GMAIL.COM
Università degli Studi di Milano, Italy

Tal Lancewicki LANCEWICKI@MAIL.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University, Israel

Aviv Rosenberg AVIVROS007@GMAIL.COM
Amazon Science

Nicoló Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano and Politecnico di Milano, Italy

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We derive a new analysis of Follow The Regularized Leader (FTRL) for online learning with delayed bandit feedback. By separating the cost of delayed feedback from that of bandit feedback, our analysis allows us to obtain new results in three important settings. On the one hand, we derive the first optimal (up to logarithmic factors) regret bounds for combinatorial semi-bandits with delay and adversarial Markov decision processes with delay (and known transition functions). On the other hand, we use our analysis to derive an efficient algorithm for linear bandits with delay achieving near-optimal regret bounds. Our novel regret decomposition shows that FTRL remains stable across multiple rounds under mild assumptions on the Hessian of the regularizer.

Keywords: Online learning, bandit feedback, delayed feedback

1. Introduction

Delayed feedback is a phenomenon that cannot be avoided in many applications of online learning. For example, in digital advertisement a conversion event may happen with some delay after an ad is shown to a user. In healthcare, the effect of a drug on a patient may take some time before it becomes observable (Eick, 1988). A consequence of delayed feedback is that sequential decision makers have to act before knowing the effect of their previous actions, where the effect of multiple past actions may be potentially observed all at once. These challenges pertain not only to the algorithms, but also to the way they are analyzed, which is the reason why standard (non-delayed) proof techniques fail in the presence of delayed feedback.

Due to its fundamental nature in online learning, delayed feedback has been extensively studied in several different scenarios, including full-information feedback (Weinberger and Ordentlich, 2002; Joulani et al., 2013; Quanrud and Khashabi, 2015; Joulani et al., 2016; Flaspohler et al., 2021) and bandit feedback (Cesa-Bianchi et al., 2016; Thune et al., 2019; Bistritz et al., 2019; Zimmert and Seldin, 2020; Ito et al., 2020a; Gyorgy and Joulani, 2021; Van Der Hoeven and Cesa-Bianchi, 2022; Masoudian et al., 2022). In this work, we focus on the more realistic case of bandit feedback; that is, when the only way for the learner to know the effect of an action is to execute it. We develop

a general framework for the analysis of delayed bandit feedback which we then apply to three important settings: combinatorial semi-bandits (which includes multi-armed bandits as a special case), adversarial Markov Decision Processes (MDPs), and linear bandits.

Our analysis, which is based on Follow The Regularized Leader (FTRL)—see, for example, (Orabona, 2019, Chapter 7), unifies previous analyses and sheds light on the impact of delayed bandit feedback in online learning. Our main insight is that one can separate the cost of delayed feedback and bandit feedback through a novel decomposition of the FTRL regret, which allows to separately bound these different regret components. This insight leads to new results in all of the settings we consider. We prove the first regret bounds for combinatorial semi-bandits with delays, which also turn out to be optimal for sufficiently large T (throughout the paper, by optimal we always mean optimal for sufficiently large T). We also prove the first regret bounds for adversarial MDPs with delays and known transitions, which are again optimal. Finally, we derive a computationally efficient algorithm for linear bandits, whose regret has an optimal dependence on delays.

We now formally introduce the setting of online learning with delayed bandit feedback studied in this paper. Online learning with delayed bandit feedback proceeds in rounds. In each round $t \in [T]$ the learner chooses (possibly in a randomized manner) an action $\boldsymbol{a}_t \in \mathcal{A} \subseteq \mathbb{R}^K$, suffers loss $\boldsymbol{a}_t^\top \boldsymbol{\ell}_t$, where $\boldsymbol{\ell}_t \in \mathbb{R}^K$ is bounded in some suitably chosen norm, and observes $\{\mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau) : \tau + d_\tau = t\}$, where $d_1, \dots, d_T$ is an unknown sequence of delays and $\mathcal{L}$ is an application-specific (possibly randomized) feedback function, encoding which information about $\boldsymbol{\ell}_\tau$ the learner sees based on the action $\boldsymbol{a}_\tau$. For example, in the combinatorial semi-bandit setting the learner observes all loss components corresponding to the non-zero elements of the action, whereas in the linear bandit setting the learner only observes the scalar $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$.

1.1. Contributions

New analysis. In section 3 we provide a novel analysis of FTRL under delayed bandit feedback. The main novelty is showing that we can decompose the regret into three main parts. The first part of the regret is standard, namely the pseudo-distance between the starting point of the algorithm and the optimal point in hindsight. The second part is the cost of delayed feedback. In our analysis, we show that the cost of delayed feedback is essentially the same as in the delayed full-information setting. The third part of the regret is the cost of bandit feedback, which is the same term that occurs in the standard analysis of FTRL for bandit feedback. A technical novelty is that we show that FTRL is stable across multiple rounds under some mild assumptions on the Hessian of the regularizer. In related work, Huang et al. (2023) provides an analysis of online mirror descent with delayed bandit feedback in several settings. However, their analysis does not lead to optimal bounds because it does not separate the cost of delayed and bandit feedback.

Combinatorial semi-bandits with delayed feedback. As far as we know we are the first to consider nonstochastic combinatorial semi-bandits under delayed feedback. In the combinatorial semi-bandit setting, we apply the newly gained insight from our analysis of FTRL to derive an optimal algorithm. We show that if $\max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_1 \leq B$, then the regret after T rounds is of order $\sqrt{B(KT + BD)\log(K)}$, where $D = \sum_{t=1}^{T} d_t$ is the total delay after T rounds. In the worst case, the delay is constant (i.e., $d_t = d$ for all t) and we provide a matching lower bound (up to logarithmic factors) showing that any learner must incur $\Omega(\sqrt{BT(K + Bd)})$ regret.

Adversarial Markov decision processes. Delayed feedback in adversarial (finite-horizon and episodic) MDPs was first studied by Lancewicki et al. (2022). Under full-information feedback,

where the agent observes the entire cost function at the end of the episode, they achieve optimal regret $\widetilde{O}(H\sqrt{T+D})$, where T is the number of episodes and H is the horizon. However, under the more realistic bandit feedback (where the only observed costs are those along the agent's trajectory), their regret bound scales with $T^{2/3}+D^{2/3}$, which is far from optimal. The current state-of-the-art guarantee under bandit feedback is by Jin et al. (2022) who achieve regret bound of $\widetilde{O}(H\sqrt{SAT}+H(HSA)^{1/4}\sqrt{D})$. However, there is still a $(HSA)^{1/4}$ factor gap on the second term compared to the lower bound of Lancewicki et al. (2022). Remarkably, the application of our FTRL analysis to adversarial MDPs allows us to close this gap and achieve the first optimal regret bound of $\widetilde{O}(H\sqrt{SAT}+H\sqrt{D})$ for the case of known transitions.

Linear bandits. In the linear bandit setting, Ito et al. (2020a) provide an analysis of continuous exponential weights (Cover, 1991; Vovk, 1990; Littlestone and Warmuth, 1994) with delayed bandit feedback and constant delay d that obtains the optimal $\widetilde{O}(K\sqrt{T}+\sqrt{dT})$ regret bound. One drawback is that the per-round runtime of continuous exponential weights is prohibitively large, although it is polynomial in K and T. Building on Scrible (Abernethy et al., 2008), we derive an algorithm that achieves a slightly suboptimal $\widetilde{O}(K^{3/2}\sqrt{T}+\sqrt{D})$ regret, but with a much better per-round running time of order K^3, provided a self-concordant barrier for the decision set can be efficiently computed. Huang et al. (2023) show an algorithm with a similar running time, but with a worse regret bound of $\widetilde{O}(K^{3/2}\sqrt{T}+K^2\sqrt{D})$.

1.2. Additional related work

Delayed feedback in stochastic models. Delayed feedback with stochastic losses were studied both in MDPs (Howson et al., 2021) and linear bandits (Vernade et al., 2020; Howson et al., 2022), as well as many other domains (Dudik et al., 2011; Agarwal and Duchi, 2012; Vernade et al., 2017, 2020; Pike-Burke et al., 2018; Cesa-Bianchi et al., 2018; Zhou et al., 2019; Gael et al., 2020; Lancewicki et al., 2021; Cohen et al., 2021). However, the adversarial losses considered in this work are much more general and induce many additional technical challenges.

Combinatorial semi-bandits with delayed feedback. Even though we are the first to study combinatorial semi-bandits with delayed feedback, a special case, namely multi-armed bandits with delayed feedback, is well understood. Neu et al. (2010, 2014) were among the first ones to study the impact of delayed feedback in the nonstochastic setting. Subsequently, Cesa-Bianchi et al. (2019) proved a $\Omega(\sqrt{KT}+\sqrt{dT\log(K)})$ lower bound when $d_t=d$ for all t. The matching upper bound was provided by Zimmert and Seldin (2020), but nearly matching upper bounds also exist Thune et al. (2019); Bistritz et al. (2019); Gyorgy and Joulani (2021); Van Der Hoeven and Cesa-Bianchi (2022). Conversely, (special cases of) combinatorial semi-bandits without delay have also received considerable attention (György et al., 2007; Kale et al., 2010; Uchiya et al., 2010; Cesa-Bianchi and Lugosi, 2012; Audibert et al., 2014; Combes et al., 2015; Lattimore et al., 2018; Zimmert et al., 2019).

Adversarial Markov decision processes. There is a rich literature on regret minimization in MDPs with non-delayed feedback (Even-Dar et al., 2009; Jaksch et al., 2010; Zimin and Neu, 2013; Dick et al., 2014; Rosenberg and Mansour, 2019b,a, 2021; Jin et al., 2020; Shani et al., 2020; Luo et al., 2021). Under delayed feedback, apart from the literature mentioned earlier, Dai et al. (2022) recently presented a Follow-The-Perturbed-Leader approach that can also handle delayed feedback in adversarial MDPs. However, their regret bound is slightly weaker than Jin et al. (2022) mentioned

earlier. Finally, a different line of work (Katsikopoulos and Engelbrecht, 2003; Walsh et al., 2009) consider delays in observing the current state, which is inherently different than our setting—for a thorough discussion on the differences between the models we refer the reader to Lancewicki et al. (2022).

Linear bandits. Early work in the non-delayed linear bandit setting suffered from suboptimal results in terms of T (McMahan and Blum, 2004; Awerbuch and Kleinberg, 2004; Dani and Hayes, 2006). Abernethy et al. (2008) were the first to prove a regret bound with optimal scaling in T. Subsequent works by (Bubeck and Eldan, 2015; Hazan and Karnin, 2016; Ito et al., 2020b; Zimmert and Lattimore, 2022) obtained the optimal $O(K\sqrt{T})$ regret bound.

2. Preliminaries

We denote by $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}^K$ the estimate of the loss $\boldsymbol{\ell}_t$ in round t. We will define a loss estimator for each application separately. We assume that delays $d_1, \ldots, d_T$ and losses $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ are both generated by an oblivious adversary. We focus on Follow The Regularized Leader (FTRL) which, in each round t, computes

$$\boldsymbol{w}_t = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{W}} \sum_{\tau \in o_t} \widehat{\boldsymbol{\ell}}_\tau^\top \boldsymbol{w} + R(\boldsymbol{w}), \tag{1}$$

where $\mathcal{W} \subseteq \mathbb{R}^K$ is a compact closed convex set, R is a twice-differentiable convex function such that $\nabla^2 R(\boldsymbol{w}) \succ 0\boldsymbol{I}$ for all $\boldsymbol{w} \in \mathcal{W}$, and $o_t = \{\tau : \tau + d_\tau < t\}$ is the set of indices of observed losses at the end of round $t-1$. Note that $\mathcal{W}$ and $\mathcal{A}$ do not necessarily coincide, as is the case of combinatorial semi-bandits for example. Similarly, $\boldsymbol{a}_t$ and $\boldsymbol{w}_t$ do not necessarily coincide. We will specify the relationship between $\boldsymbol{a}_t$ and $\boldsymbol{w}_t$ in each application. We define $m_t = [t-1] \setminus o_t$ to be the set of indices of losses that have not been observed at the end of round $t-1$ due to delay. As a simplifying assumption, we assume that $d_{\max} = \max_{t \in [T]} d_t \geq 1$ which is known to the learner. We also use the simplifying assumptions that $\sum_{t=1}^T |m_t| = D$ and T are both known to the learner. These assumptions are without loss of generality, as we may employ the standard doubling trick to overcome the need to know these parameters (Bistritz et al., 2019; Lancewicki et al., 2022), see also Appendix E. We also make use of the following notations for FTRL on cumulative loss $\boldsymbol{L}$, which we denote by

$$\boldsymbol{w}(\boldsymbol{L}) = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{L}^\top \boldsymbol{w} + R(\boldsymbol{w}).$$

In the remainder of the paper we use the following cumulative losses:

$$\widehat{\boldsymbol{L}}_t = \sum_{\tau \in o_t} \widehat{\boldsymbol{\ell}}_\tau, \qquad \overline{\boldsymbol{L}}_t^m = \widehat{\boldsymbol{L}}_t + \sum_{\tau \in m_t} \boldsymbol{\ell}_\tau, \qquad \widehat{\boldsymbol{L}}_t^m = \widehat{\boldsymbol{L}}_t + \sum_{\tau \in m_t} \widehat{\boldsymbol{\ell}}_\tau, \qquad \widehat{\boldsymbol{L}}_t^\star = \sum_{\tau \in [t]} \widehat{\boldsymbol{\ell}}_\tau$$

Note that $\widehat{\boldsymbol{L}}_t^\star = \widehat{\boldsymbol{L}}_t + \sum_{\tau \in [t] \setminus o_t} \widehat{\boldsymbol{\ell}}_\tau$, $\boldsymbol{w}(\widehat{\boldsymbol{L}}_t) = \boldsymbol{w}_t$ and that $\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)$ is equivalent to FTRL in the non-delayed setting.

Additional notations. We define a filtration of all random events observed by the learner up to round t as $\mathcal{F}_t = \left\{\left(\tau, \boldsymbol{a}_\tau, \mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau)\right) : \tau + d_\tau < t\right\}$. For a twice-differentiable function ϕ such that $\nabla^2 \phi(\boldsymbol{w}) \succ 0\boldsymbol{I}$ for all $\boldsymbol{w} \in \mathcal{W}$ we will denote by $\|\boldsymbol{L}\|_{\phi, \boldsymbol{w}} = \sqrt{\boldsymbol{L}^\top \left(\nabla^2 \phi(\boldsymbol{w})\right)^{-1} \boldsymbol{L}}$ and by

$\|\boldsymbol{L}\|^{\star}_{\phi,\boldsymbol{w}} = \sqrt{\boldsymbol{L}^\top \nabla^2 \phi(\boldsymbol{w}) \boldsymbol{L}}$. The Dikin ellipsoid with radius r around $\boldsymbol{w}$ induced by ϕ is defined as $\mathcal{D}_\phi(\boldsymbol{w}, r) = \{\boldsymbol{x} \in \mathcal{W} : \|\boldsymbol{x} - \boldsymbol{w}\|^{\star}_{\phi,\boldsymbol{w}} \le r\}$. The notations $\widetilde{O}(\cdot)$ and $\lesssim$ hide poly-logarithmic factors.

3. Analysis

We build on the analysis of Flaspohler et al. (2021) for delayed feedback in the full-information setting, where they observe that delayed feedback can be interpreted as poor hints in the sense of optimistic online learning (Rakhlin and Sridharan, 2013). Taking this idea one step further, we analyze what would happen had the algorithm received slightly different hints, and subsequently bound the difference between different instances of FTRL.

We assume that our loss estimates satisfy $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t | \mathcal{F}_t] = \boldsymbol{\ell}_t + \boldsymbol{b}_t$, where $\boldsymbol{b}_t$ is the estimator's bias. Our analysis relies on the following decomposition of the instantaneous regret

$$\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] = \underbrace{\sum_{t=1}^{T} -\mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right]}_{\text{bias}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t\right]}_{\text{cheating regret}} \tag{2}$$

$$+ \sum_{t=1}^{T} \Big(\mathbb{E}\big[\underbrace{(\boldsymbol{w}_t - \boldsymbol{w}(\overline{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t}_{H_1}\big] + \mathbb{E}\big[\underbrace{(\boldsymbol{w}(\overline{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t}_{H_2}\big] + \mathbb{E}\big[\underbrace{(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t}_{H_3}\big] \Big).$$

Suppose that $\boldsymbol{b}_t = \boldsymbol{0}$, i.e., $\widehat{\boldsymbol{\ell}}_t$ is an unbiased estimator of the loss. This implies that the bias term of the decomposition is 0. The cheating regret can be found in different forms in online learning—see, for example, the proof of (Shalev-Shwartz, 2012, Lemma 2.3) or (Gyorgy and Joulani, 2021, Equation 4)—and can be bounded using the standard be-the-leader lemma (Lemma 11 in Appendix A, see also, for example Theorem 3 of (Joulani et al., 2020)). Now, let us focus on the second line of Equation (2). Once simplified, the second line becomes $\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t\right]$,which can be recognised as the drift term of Gyorgy and Joulani (2021, Equation (4)). Normally, we would like to use standard tools from linear bandits or online convex optimization to bound such terms, such as for example (a variation of) Lemma 1 below, the proof of which can be found in Appendix A.

Lemma 1 *Suppose that $\nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$, $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$, and for some twice-differentiable convex R. Let $\boldsymbol{x} \in \mathcal{W}$ and $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^K$ such that $\boldsymbol{w}(\boldsymbol{L}'), \boldsymbol{w}(\boldsymbol{L}) \in \mathcal{D}_R(\boldsymbol{x}, \frac{1}{2})$, then $\|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^{\star}_{R,\boldsymbol{x}} \le 8\|\boldsymbol{L}' - \boldsymbol{L}\|_{R,\boldsymbol{x}}$.*

We can bound the drift term $(\boldsymbol{w}_t - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t$ by $\|\boldsymbol{w}_t - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star)\|^{\star}_{R,\boldsymbol{w}_t} \|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}$ using Hölder inequality, and then apply Lemma 1 to further bound the right-hand side. This would lead to the problematic term

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|\sum_{\tau \in [t] \setminus o_t} \widehat{\boldsymbol{\ell}}_\tau\right\|_{R,\boldsymbol{w}_t} \|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}\right] \le \sum_{t=1}^{T} \mathbb{E}\left[(1 + |m_t|) \max_{\tau \in [t] \setminus o_t} \|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R,\boldsymbol{w}_t}\right].$$

To see where the problem is, suppose we are in the multi-armed bandit setting, R is the negative entropy scaled by $\frac{1}{\eta}$, and $\widehat{\boldsymbol{\ell}}_t$ is the standard importance-weighted estimator. Then, the upper bound above is $O(\eta(1 + |m_t|)K)$, where K is the number of arms. Summing over T rounds, using a $\frac{\log(K)}{\eta}$ bound on the cheating regret, and tuning η, we see that this analysis delivers a regret of order

$O\big(\sqrt{K(T+D)\log(K)}\big)$ where $D = \sum_t |m_t|$. In the case of constant delay $d_t = d$, the bound becomes $O\big(\sqrt{KdT\log(K)}\big)$ which is known to be suboptimal, as the minimax regret in this case is of order $\max\big\{\sqrt{dT\log(K)}, \sqrt{KT}\big\}$ (Cesa-Bianchi et al., 2019; Zimmert and Seldin, 2020).

The intuition behind the suboptimality of the above analysis is that the cost of bandit feedback and the cost of delayed feedback are not separated. Indeed, the analysis of most lower bounds is split in two cases: a lower bound for bandit feedback without delay and a lower bound for delayed *full-information* feedback, see for example Cesa-Bianchi et al. (2019). Separating the impact of delayed feedback and bandit feedback is precisely why we bound the terms H_1, H_2, and H_3 of Equation (2) separately, which leads to Lemma 2 below, whose proof can be found in Appendix A.

Lemma 2 *Let R be convex and twice-differentiable such that $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. Let $\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} \leq \alpha \leq \frac{1}{16 d_{\max}}$ for all t and $\boldsymbol{w} \in \mathcal{W}$, and $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_t\|^2_{R,\boldsymbol{w}_t}\right] \leq \beta^2$ for all t, where $\boldsymbol{w}_t$ is given by (1). Suppose also that $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \frac{1}{64(1+d_{\max})}$ for all t with probability 1. Then for all $\boldsymbol{u} \in \mathcal{W}$*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 8\beta^2 T - \sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] + \sum_{t=1}^{T}\left(8\alpha^2|m_t| + 8\alpha\,\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R,\boldsymbol{w}_t}\right]\right).$$

The work of Van Der Hoeven and Cesa-Bianchi (2022) provides a similar result for the multi-armed bandit setting. However, that result does not apply to the more general linear bandit optimization setting, which Lemma 2 does. The result of Van Der Hoeven and Cesa-Bianchi (2022) can not be easily extend to our setting: their analysis relies on the fact that the constraint in the Lagrangian of the FTRL objective can be expressed in a simple manner in the multi-armed bandit setting, which is not possible in our setting.

To interpret Lemma 2, consider the multi-armed bandit setting with the standard importance-weighted estimator and regularizer $R(\boldsymbol{w}) = \sum_{i=1}^{K}\frac{1}{\eta}\boldsymbol{w}(i)\log(\boldsymbol{w}(i)) - \frac{1}{\gamma}\log(\boldsymbol{w}(i))$. The purpose of the log barrier term in the regularizer is to ensure stability of the iterates, as required by the assumptions of the lemma. In this case, if $\|\boldsymbol{\ell}_t\|_\infty \leq 1$, then α is $O(\sqrt{\eta})$. The quantity β^2 is a bound on the expectation of the squared local norm of the loss estimate, which is $O(\eta K)$. Thus, by choosing $u(i) = \widetilde{u}(i)\big(1 - 1/\frac{1}{T}\big) + \frac{1}{T}w_1(i)$, we have that the expected regret against $\widetilde{\boldsymbol{u}}$ is of order

$$\frac{1}{\eta}\log(K) + d_{\max}^2 K\ln(T) + \eta(KT + D) + \sum_{t=1}^{T}\alpha\,\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R,\boldsymbol{w}_t}\right], \tag{3}$$

where we used that $\sum_{t=1}^{T}|m_t| = D$. The $d_{\max}^2 K\ln(T)$ term in the above equation comes from the log-barrier part of R, which when tuned properly is able to ensure that the iterates of FTRL are close to each other. So far, it seems that we did not manage to separate the cost of delay and bandit feedback because of the final summation in (3). However, observe that due to the delay, for $\tau, \tau' \in m_t$, $\widehat{\boldsymbol{\ell}}_\tau$ and $\widehat{\boldsymbol{\ell}}_{\tau'}$ are independent random variables and $\boldsymbol{\ell}_\tau$ and $\boldsymbol{\ell}_{\tau'}$ are their means. Recall that the variance of the sum of independent random variables equals to the sum of their variances. Thus, by applying Jensen's inequality to the square root, and using that $\nabla^2 R(\boldsymbol{w}) \succeq \operatorname{diag}\big(\eta\boldsymbol{w}\big)^{-1}$, we can

see that

$$\alpha\,\mathbb{E}\left[\Big\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\Big\|_{R,\boldsymbol{w}_t}\right]\le 2\alpha\,\mathbb{E}\left[\Big\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\Big\|_{R,\boldsymbol{w}_\tau}\right]$$
$$\le 2\alpha\sqrt{\mathbb{E}\left[\sum_{i=1}^{K}\eta\boldsymbol{w}_\tau(i)\left(\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau(i)-\widehat{\boldsymbol{\ell}}_\tau(i))\right)^2\right]}=2\alpha\sqrt{\mathbb{E}\left[\sum_{\tau\in m_t}\sum_{i=1}^{K}\eta\boldsymbol{w}_\tau(i)\big(\boldsymbol{\ell}_\tau(i)-\widehat{\boldsymbol{\ell}}_\tau(i)\big)^2\right]}$$
$$\le\sqrt{\alpha^2|m_t|\eta K}\,,$$

where the first inequality is due to Lemma 10 in Appendix A, a new result that proves the multi-round stability of FTRL iterates under certain conditions, which can be applied for sufficiently small γ. Recalling that α is $O(\sqrt{\eta})$ and using $\sqrt{\eta|m_t|\eta K}\le\frac{1}{2}(\eta|m_t|+\eta K)$ we can see that (3) is in fact of order $\log(K)/\eta+d_{\max}^2K\ln(T)+\eta(KT+D)$, which gives a $O\big(\sqrt{(KT+D)\log(K)}+d_{\max}^2K\ln(T)\big)$ bound for an appropriately tuned η.

To conclude, as long as loss estimates $\widehat{\boldsymbol{\ell}}_\tau$ and $\widehat{\boldsymbol{\ell}}_{\tau'}$ are independent for $\tau,\tau'\in m_t$, Lemma 2 implies that we have effectively split the cost of delayed feedback and bandit feedback. We formalize the above in Corollary 3, whose proof can be found in Appendix A.

Corollary 3 *Under the same assumptions as in Lemma 2, suppose that* $\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t]=\boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau)^\top\big(\nabla^2R(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t))\big)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'}-\boldsymbol{\ell}_{\tau'})\,\middle|\,\mathcal{F}_t\right]=0$ *for all* $t\in[T]$ *and all* $\tau,\tau'\in m_t$ *where* $\tau'\neq\tau$. *Then for all* $\boldsymbol{u}\in\mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right]\le R(\boldsymbol{u})-R(\boldsymbol{w}_1)+16\beta^2T+16\alpha^2D\,.$$

4. Combinatorial Bandits

In this section, we demonstrate how to apply our generic FTRL approach to combinatorial bandits (CMAB) with delayed feedback. This yields the first algorithm to achieve optimal regret in that setting. We start with the description of the model as an instance of our general setting.

Delayed combinatorial bandits with semi-bandit feedback is an instance of the online learning framework where $\boldsymbol{\ell}_t\in[-1,1]^K$, $\mathcal{A}\subseteq\{0,1\}^K$, $\mathcal{W}=\mathrm{Conv}(\mathcal{A})$ (the convex hull of $\mathcal{A}$), and the feedback function is $\mathcal{L}(\boldsymbol{\ell}_\tau,\boldsymbol{a}_\tau)=\boldsymbol{a}_\tau\odot\boldsymbol{\ell}_\tau$ where $\odot$ is the Hadamard (elementwise) vector product.

We define the pseudo-regret in this setting as

$$\mathcal{R}_T=\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t-\boldsymbol{a}^*)^\top\boldsymbol{\ell}_t\right],$$

with $\boldsymbol{a}^*=\arg\min_{\boldsymbol{a}\in\mathcal{A}}\sum_{t=1}^{T}\boldsymbol{a}^\top\boldsymbol{\ell}_t$. The algorithm we use in this setting (Algorithm 1) is inspired by the algorithm of Audibert et al. (2014). In any given round t, Algorithm 1 first computes $\boldsymbol{w}_t$, the solution of the FTRL optimization problem (Eq. (1)) over the convex hull of the action set. Then, it constructs a probability distribution $\boldsymbol{p}_t$ over $\mathcal{A}$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}]=\boldsymbol{w}_t$. The estimator of loss is

Algorithm 1: Delayed FTRL for combinatorial bandits

Input: $\gamma \in (0,1)$, η.
for $t \in [T]$ **do**
  Compute $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \sum_{\tau\in o_t} \widehat{\boldsymbol{\ell}}_\tau^\top \boldsymbol{w} + R(\boldsymbol{w})$ with R as in equation (4).
  Find probability distribution $\boldsymbol{p}_t$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t$.
  Draw and play $\boldsymbol{a}_t \sim \boldsymbol{p}_t$.
  Observe loss $\boldsymbol{a}_\tau \odot l_\tau$ and compute $\widehat{\boldsymbol{\ell}}_\tau(i) = \frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(i)}$ for all $\tau \in o_t$.
end for

given by $\widehat{\boldsymbol{\ell}}_t(i) = \frac{\boldsymbol{a}_t(i)\boldsymbol{\ell}_t(i)}{\boldsymbol{w}_t(i)}$, which is clearly unbiased. We use the regularizer

$$R(\boldsymbol{w}) = \sum_{i=1}^{K} \Big(\frac{1}{\eta}\boldsymbol{w}(i)\log(\boldsymbol{w}(i)) - \frac{1}{\gamma}\log(\boldsymbol{w}(i))\Big), \tag{4}$$

where $\eta > 0$ and $\gamma > 0$ are hyperparameters. We now state the main results of this section.

Theorem 4 *Let* $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_1 \le B$. *Running Delayed FTRL for combinatorial bandits (Algorithm 1) with* $\gamma = \frac{1}{64^2 B(1+d_{\max})^2}$ *and* $\eta = \min\left\{\frac{1}{16^2 B d_{\max}^2}, \sqrt{\frac{B\left(1+\ln\left(\frac{K}{B}\right)\right)}{16(KT+BD)}}\right\}$ *guarantees,*

$$\mathcal{R}_T \le 12\sqrt{B(KT+BD)\ln\left(\frac{K}{B}\right)} + 64^2 B(1+d_{\max})^2 K\ln(T) + 2^9 B^2 d_{\max}^2 \ln\left(\frac{K}{B}\right).$$

We sketch the proof of Theorem 4 (see Appendix B for a full proof) which follows from an application of Corollary 3. To apply this corollary we first need to verify its assumptions. The assumptions on $\nabla^2 R(\boldsymbol{w})$ are implied by Lemma 16 in Appendix F. Next, we set $\alpha = \sqrt{\eta B}$ and verify that $\frac{1}{16 d_{\max}} \ge \alpha \ge \|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}}$. Indeed, $\boldsymbol{w} \in \mathrm{Conv}(\mathcal{A})$ implies $\sum_{i=1}^K \boldsymbol{w}(i) \le B$. This, together with $|\boldsymbol{\ell}(i)| \le 1$ and $\eta + \gamma\boldsymbol{w}(i) \ge \gamma\boldsymbol{w}(i)$, implies

$$\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} = \sqrt{\boldsymbol{\ell}_t^\top \left(\nabla^2 R(\boldsymbol{w})\right)^{-1}\boldsymbol{\ell}_t} = \sqrt{\sum_{i=1}^{K} \boldsymbol{\ell}_t(i)^2 \frac{\eta\gamma\boldsymbol{w}^2(i)}{\eta+\gamma\boldsymbol{w}(i)}} \le \sqrt{\eta B} = \alpha\,.$$

Our choice of η then implies $\alpha \le \frac{1}{16 d_{\max}}$. Similarly, by setting $\beta^2 = \eta K$ we fulfill the condition $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2 | \mathcal{F}_t\right] \le \beta^2$. Finally, the condition on γ allows us to verify the assumption $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \le \frac{1}{64(1+d_{\max})}$. Since all the assumptions are verified, we can now apply Corollary 3 and finish the proof of Theorem 4. The Algorithm is computationally efficient for a range of action-sets including m-sets and spanning trees. In general the algorithm is efficient whenever OSMD of Audibert et al. (2014) is efficient and we refer to that paper for more details.

We now state a lower bound for the delayed combinatorial semi-bandit setting. This implies that, ignoring terms that are logarithmic in T, the result of Theorem 4 is optimal. The proof of our lower bound follows from standard arguments in the delayed bandit feedback literature and can be found in Appendix B.

Theorem 5 *Suppose that $d_t = d$ for all t and that $B \leq K/2$. Then for any algorithm there exists a sequence of losses such that*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{a}^\star)^\top \boldsymbol{\ell}_t\right] = \Omega\left(\max\left\{\sqrt{BKT}, B\sqrt{dT}\right\}\right) \quad .$$

5. Adversarial Markov Decision Processes

In this section, we apply our FTRL approach to adversarial Markov Decision Processes (MDPs) where the transition function is known to the learner in advance. We show that it yields the first algorithm that handles delay optimally in this setting. We start with a presentation of the model and regret minimization framework.

A finite-horizon episodic adversarial MDP is defined by $\mathcal{M} = (\mathscr{S}, \mathscr{A}, H, p, \{\boldsymbol{\ell}_t\}_{t=1}^T, s_{\text{init}})$, where $\mathscr{S}$ and $\mathscr{A}$ are finite state and action spaces of sizes S and A, respectively, H is the horizon, T is the number of episodes, and $s_{\text{init}} \in \mathscr{S}$ is the initial state. $p = \{p_h : \mathscr{S} \times \mathscr{A} \to \Delta_{\mathscr{S}}\}_{h=1}^H$ is the *transition function* such that $p_h(s' \mid s, a)$ is the probability of moving to s' when taking action a in state s at time h. $\{\boldsymbol{\ell}_{t,h} : \mathscr{S} \times \mathscr{A} \to [0,1]\}_{t=1,h=1}^{T,H}$ are *cost functions* chosen by an *oblivious adversary*, where $\boldsymbol{\ell}_{t,h}(s,a)$ is the cost of taking action a in state s at time h of episode t. For the ease of presentation, we slightly abuse notation and treat each set of loss functions $\{\boldsymbol{\ell}_{t,h} : \mathscr{S} \times \mathscr{A} \to [0,1]\}_{h=1}^H$ simply as a vector $\boldsymbol{\ell}_t \in [0,1]^{HSA}$.

The learner interacts with the environment over T episodes. At the beginning of episode t, it picks a policy $\pi_t = \{\pi_{t,h} : \mathscr{S} \to \Delta_{\mathscr{A}}\}_{h=1}^H$, and starts in the initial state $s_{t,1} = s_{\text{init}}$. At each time $h \in [H]$, it observes the current state $s_{t,h} \in \mathscr{S}$, draws an action from the policy $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$ and transitions to the next state $s_{t,h+1} \sim p_h(\cdot \mid s_{t,h}, a_{t,h})$. The feedback of episode t contains the cost function over the agent's trajectory $\{\boldsymbol{\ell}_{t,h}(s_{t,h}, a_{t,h})\}_{h=1}^H$, i.e., bandit feedback, and is observed only at the end of episode $t + d_t$. The learner's goal is to minimize the value of its policies, where $V_{t,h}^\pi(s) = \mathbb{E}[\sum_{h'=h}^H \boldsymbol{\ell}_{t,h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi, p]$ is the value function of policy π with respect to the cost $\boldsymbol{\ell}_t$. The performance is measured by the *regret*, defined as the difference between the cumulative expected cost of the learner and the best fixed policy in hindsight: $\mathcal{R}_T = \sum_{t=1}^T V_{t,1}^{\pi_t}(s_{\text{init}}) - \min_{\pi \in \Pi} \sum_{t=1}^T V_{t,1}^{\pi}(s_{\text{init}})$.

In order present the adversarial MDP model as an instance of the general online learning framework we use the notion of *occupancy measures*. Given a policy π and a transition function p', the *occupancy measure* $\boldsymbol{w}^{\pi,p'} \in [0,1]^{HS^2A}$ is a vector, where $\boldsymbol{w}_h^{\pi,p'}(s,a,s')$ is the probability to visit state s at time h, take action a and transition to state s'. We also denote $\boldsymbol{w}_h^{\pi,p'}(s,a) = \sum_{s'} \boldsymbol{w}_h^{\pi,p'}(s,a,s')$ and $\boldsymbol{w}_h^{\pi,p'}(s) = \sum_a \boldsymbol{w}_h^{\pi,p'}(s,a)$. By Rosenberg and Mansour (2019b) (see also Zimin and Neu (2013); Dick et al. (2014)), the occupancy measure encodes the policy and the transition function through the relations $\pi_h(a \mid s) = \boldsymbol{w}_h^{\pi,p'}(s,a) / \boldsymbol{w}_h^{\pi,p'}(s)$; $p'_h(s' \mid s, a) = \boldsymbol{w}_h^{\pi,p'}(s,a,s') / \boldsymbol{w}_h^{\pi,p'}(s,a)$. The set of all occupancy measures with respect to an MDP $\mathcal{M}$ is denoted by $\Delta(\mathcal{M})$, and the set of all policies by $\Pi = \left\{\{\pi_h : \mathscr{S} \to \Delta_{\mathscr{A}}\}_{h=1}^H\right\}$. Importantly, the value of a policy from the initial state (i.e., the expected loss in an episode) can be written as the dot product between its occupancy measure and the cost function, i.e., $\langle \boldsymbol{w}^{\pi,p'}, \boldsymbol{\ell} \rangle = \sum_{h,s,a} \boldsymbol{w}_h^{\pi,p'}(s,a) \boldsymbol{\ell}_h(s,a)$. Thus, the regret becomes $\mathcal{R}_T = \sum_{t=1}^T \langle \boldsymbol{w}^{\pi_t,p}, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{w} \in \Delta(\mathcal{M})} \sum_{t=1}^T \langle \boldsymbol{w}, \boldsymbol{\ell}_t \rangle$. Whenever p' is omitted from the notation $\boldsymbol{w}^{\pi,p'}$, it is understood to be the true transition function p.

Algorithm 2: Delayed FTRL for adversarial MDPs

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

Compute $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{w} + R(\boldsymbol{w})$, and policy $\pi_{t,h}(a \mid s) = \frac{\boldsymbol{w}_{t,h}(s,a)}{\boldsymbol{w}_{t,h}(s)} \ \ \forall (s,a,h)$.

Play episode t with policy π_t, and observe feedback $\{\boldsymbol{\ell}_{t,h}(s_{\tau,h}, a_{\tau,h})\}_{h=1}^H$ for all $\tau + d_\tau = t$.

Compute upper occupancy bound $\boldsymbol{u}_{\tau,h}(s,a) = \max_{\widehat{p} \in \mathcal{P}} \boldsymbol{w}_h^{\pi_\tau, \widehat{p}}(s,a)$.

Compute loss estimator $\widehat{\boldsymbol{\ell}}_{\tau,h}(s,a) = \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}\boldsymbol{\ell}_{\tau,h}(s,a)}{\boldsymbol{u}_{\tau,h}(s,a)}$ and update $\widehat{L}_t$.

end for

With that in hand, adversarial MDPs is an instance of the online learning framework where $\boldsymbol{\ell}_t \in [0,1]^{HSA}$, $\mathcal{A}$ as the set of occupancy measures $\Delta(\mathcal{M})$ and the feedback $\mathcal{L}(\boldsymbol{w}^{\pi_\tau}, \boldsymbol{\ell}_\tau)$ is the loss over the trajectory $\{\boldsymbol{\ell}_{\tau,h}(s_{\tau,h}, a_{\tau,h})\}_{h=1}^H$. $\mathcal{W}$ is a (slightly modified) set of occupancy measures which we will define later.

Next, we present our FTRL algorithm (Algorithm 2), based on the general framework presented in Section 3. To satisfy the stability conditions required for Lemma 2, we employ a hybrid regularization of negative entropy and log-barrier similar to the combinatorial bandit case: $R(\boldsymbol{w}) = \frac{1}{\eta} \sum_{h,s,a,s'} \boldsymbol{w}_h(s,a,s') \log \boldsymbol{w}_h(s,a,s') - \frac{1}{\gamma} \sum_{h,s,a,s'} \log \boldsymbol{w}_h(s,a,s')$. The main difference is that some of the elements of the occupancy measures may be 0 regardless of the chosen policy (e.g., if $p_h(s' \mid s,a) = 0$, then $\boldsymbol{w}_h^\pi(s,a,s') = 0$), in which case the regularization is not well-defined. To avoid that, we first augment the set of occupancy measures as follows: $\Delta(\mathcal{P}) = \{\boldsymbol{w}^{\pi,\widehat{p}} : \pi \in \Pi, \widehat{p} \in \mathcal{P}\}$ where $\mathcal{P} = \left\{\{\widehat{p}_h\}_{h=1}^H : \forall h, \ \|\widehat{p}_h - p_h\|_\infty \le \frac{1}{THSA}\right\}$. Then, we intersect it with $\Omega = \left\{\boldsymbol{w} \in [0,1]^{HS^2A} : \forall (h,s,a,s'). \ \boldsymbol{w}_h(s,a,s') \ge \frac{1}{T^3H^2S^4A^2}\right\}$. This construction allows us to establish the following properties (the proof can be found in Appendix C):

Lemma 6 *Let $\mathcal{W} = \Delta(\mathcal{P}) \cap \Omega$. It holds that $\mathcal{W}$ is non-empty and,*

1. *For any $\boldsymbol{w} \in \Delta(\mathcal{M})$, there exists $\widetilde{\boldsymbol{w}} \in \mathcal{W}$ such that $\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_1 \le \frac{2H}{T}$.*
2. *Given $\boldsymbol{w} \in \mathcal{W}$, let π be defined by $\pi_h(a \mid s) = \frac{\boldsymbol{w}_h(s,a)}{\boldsymbol{w}_h(s)}$ and $\boldsymbol{u}_h(s,a) = \max_{\widehat{p} \in \mathcal{P}} \boldsymbol{w}_h^{\pi,\widehat{p}}(s,a)$. Then, $\|\boldsymbol{w}^\pi - \boldsymbol{w}\|_1 \le \frac{2H}{T}$ and $\|\boldsymbol{u} - \boldsymbol{w}\|_1 \le \frac{4H^2S}{T}$.*

With that in hand, the regularization R is well-defined on the the domain $\mathcal{W}$ and bounded by $\widetilde{O}(\frac{1}{\eta} + \frac{HS^2A}{\gamma})$. Moreover, we are guaranteed that given the iterate $\boldsymbol{w}_t$ and the corresponding policy $\pi_{t,h}(a \mid s) = \boldsymbol{w}_{t,h}(s,a)/\boldsymbol{w}_{t,h}(s)$, the true occupancy measure $\boldsymbol{w}^{\pi_t}$ is close to $\boldsymbol{w}_t$ up to a small error. Next, to keep the local norm of the estimator small (which affects the guarantee of Lemma 2), we introduce a slightly biased importance-sampling estimator (inspired by Jin et al. (2020)) defined by $\widehat{\boldsymbol{\ell}}_{t,h}(s,a) = \frac{\mathbb{I}\{s_{t,h}=s, a_{t,h}=a\}\boldsymbol{\ell}_{t,h}(s,a)}{\boldsymbol{u}_{t,h}(s,a)}$ where $\boldsymbol{u}_{t,h}(s,a) = \max_{\widehat{p} \in \mathcal{P}} \boldsymbol{w}_h^{\pi_t,\widehat{p}}(s,a)$. Recall that the local norm is evaluated at $\boldsymbol{w}_t$ and note that the expectation of the indicator $\mathbb{I}\{s_{t,h} = s, a_{t,h} = a\}$ is $w_h^{\pi_t}(s,a)$. Thus, the fact that $\boldsymbol{u}_{t,h}(s,a)$ upper bounds both $\boldsymbol{w}_{t,h}(s,a)$ and $\boldsymbol{w}_h^{\pi_t}(s,a)$ would allow us to keep local norm small. In addition, using the second property in Lemma 6, we can also show that the estimator's bias is only of order $1/T$ (ignoring S, H factors). Finally, note that $\mathcal{W}$ is a convex set defined by linear constrains, and thus the optimization can be solved efficiently (Rosenberg and Mansour, 2019b; Lee et al., 2020). In addition, $\boldsymbol{u}_t$ can be computed efficiently as well using dynamic programming (Jin et al., 2020).

Algorithm 3: Delayed FTRL for linear bandits

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

Compute $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{w} + R(\boldsymbol{w})$, where $R(\boldsymbol{w}) = \frac{1}{\eta}\|\boldsymbol{w}\|_2^2 + \frac{1}{\gamma}\Psi(\boldsymbol{w})$ and Ψ is a ν-self concordant barrier for $\mathcal{W}$.

Play $\boldsymbol{a}_t = \boldsymbol{w}_t + \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1/2} \boldsymbol{v}_t$, where v_t is uniformly sampled from the unit sphere.

Observe $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$ for $\tau : \tau + d_\tau = t$.

Compute loss estimators $\widehat{\boldsymbol{\ell}}_\tau = K \boldsymbol{\ell}_\tau^\top \boldsymbol{a}_\tau \left(\nabla^2 R(\boldsymbol{w}_\tau)\right)^{1/2} \boldsymbol{v}_\tau$ for $\tau : \tau + d_\tau = t$ and update $\widehat{L}_t$.

end for

We finish this section with the regret bound of Algorithm 2. The proof relies on the following regret decomposition

$$\mathcal{R}_T = \sum_{t=1}^{T} \langle \boldsymbol{w}^{\pi_t} - \boldsymbol{w}^\star, \boldsymbol{\ell}_t \rangle = \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{w}^{\pi_t} - \boldsymbol{w}_t, \boldsymbol{\ell}_t \rangle}_{\text{ERROR}} + \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star, \boldsymbol{\ell}_t \rangle}_{\text{REG}} + \underbrace{\sum_{t=1}^{T} \langle \widetilde{\boldsymbol{w}}^\star - \boldsymbol{w}^\star, \boldsymbol{\ell}_t \rangle}_{\text{SHIFT-PENALTY}},$$

where, by Lemma 6, ERROR is bounded by $2H$ and $\widetilde{w}^\star \in \mathcal{W}$ exists such that SHIFT-PENALTY is bounded by $2H$. Finally, REG is bounded by utilizing Lemma 2. Due to lack of space we defer the full details of the proof to Appendix C.

Theorem 7 *Running Delayed FTRL for adversarial MDPs (Algorithm 2) with* $\gamma = \frac{1}{4096H(1+d_{max})^2}$ *and* $\eta = \min\left\{\frac{1}{256H(1+d_{max})^2}, \frac{1}{\sqrt{(SAT+D)\log(HSAT)}}\right\}$ *guarantees*

$$\mathbb{E}[\mathcal{R}_T] \leq 10H\sqrt{SAT\log(HSAT)} + 10H\sqrt{D\log(HSAT)} + 7 \cdot 10^5 H^2 S^2 A(1+d_{max})^2.$$

Remarkably, the above regret bound matches the lower bound of Lancewicki et al. (2022) (up to poly-logarithmic factors), making it the first optimal regret for adversarial MDPs with delayed bandit feedback.

6. Linear Bandits

In this section, we show how to apply our analysis of FTRL to linear bandits with delayed feedback. Linear bandits with delayed feedback is an instance of our general setting where $\boldsymbol{\ell}_t \in \mathbb{R}^K$ such that $\max_t \|\boldsymbol{\ell}_t\|_2 \leq 1$, $\mathcal{A} = \mathcal{W} \subset \mathbb{R}^K$, and the feedback function is $\mathcal{L}(\boldsymbol{\ell}, \boldsymbol{a}) = \boldsymbol{\ell}^\top \boldsymbol{a}$. Additionally, we assume that $\mathcal{W} \subseteq \mathcal{B}(B)$, where $\mathcal{B}(B)$ is an L_2 ball with radius B.

Our algorithm for the linear bandit setting is inspired by Abernethy et al. (2008), who show a regularizer delivering nearly optimal bounds for the linear bandit setting with an efficient algorithm. For the delayed linear bandit setting we use a regularizer of the form $R(\boldsymbol{w}) = \frac{1}{\eta}\|\boldsymbol{w}\|_2^2 + \frac{1}{\gamma}\Psi(\boldsymbol{w})$, where Ψ is a ν-self-concordant barrier function for $\mathcal{W}$. Recall that a thrice-differentiable function Ψ is called self-concordant if it is convex and satisfies $|\nabla^3 \Psi(\boldsymbol{w})[\boldsymbol{h}, \boldsymbol{h}, \boldsymbol{h}]| \leq 2\left(\nabla^2 \Psi(\boldsymbol{w})[\boldsymbol{h}, \boldsymbol{h}]\right)^{3/2}$, where $\nabla^3 \Psi(\boldsymbol{w})[\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3] = \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3}|_{t_1=t_2=t_3=0} \Psi(\boldsymbol{w} + t_1\boldsymbol{h}_1 + t_2\boldsymbol{h}_2 + t_3\boldsymbol{h}_3)$. A self-concordant function Ψ is a ν-self-concordant barrier if $|\nabla \Psi(\boldsymbol{w})[\boldsymbol{h}]| \leq \sqrt{\nu \nabla^2 \Psi(\boldsymbol{w})[\boldsymbol{h}, \boldsymbol{h}]}$. As a specific example, the log barrier, $-\log(x)$, is 1-self-concordant for the non-negative reals. For a thorough introduction to self-concordant barriers, we refer the reader to Nesterov and Nemirovskii (1994).

Here, we only recall the most important properties, which can be found in (Nemirovski and Todd, 2008, Section 2).

The first property will allow us to satisfy the stability condition of the Hessian in Lemma 2: For $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$, we have that if $\|\boldsymbol{w} - \boldsymbol{w}'\|^{\star}_{\Psi,\boldsymbol{w}} < 1$ then

$$\left(1 - \|\boldsymbol{w} - \boldsymbol{w}'\|^{\star}_{\Psi,\boldsymbol{w}}\right)^2 \nabla^2\Psi(\boldsymbol{w}) \preceq \Psi(\boldsymbol{w}') \preceq \left(1 - \|\boldsymbol{w} - \boldsymbol{w}'\|^{\star}_{\Psi,\boldsymbol{w}}\right)^{-2} \nabla^2\Psi(\boldsymbol{w}) \,. \tag{5}$$

Next, given $\boldsymbol{y} \in \mathcal{W}$ denote by $\pi_{\boldsymbol{y}}(\boldsymbol{x}) = \inf\{z \geq 0 : \boldsymbol{y} + z^{-1}(\boldsymbol{x} - \boldsymbol{y}) \in \mathcal{W}\}$ the Minkowsky function. We denote by $\mathcal{W}_\delta = \{\boldsymbol{w} : \pi_{\boldsymbol{w}_1}(\boldsymbol{w}) \leq (1+\delta)^{-1}\}$, where $\delta > 0$. If Ψ is a ν-self-concordant barrier, then for any $\boldsymbol{w} \in \mathcal{W}_\delta$

$$\Psi(\boldsymbol{w}) - \Psi(\boldsymbol{w}_1) \leq \nu \ln\left((1+\delta)\delta^{-1}\right) . \tag{6}$$

This property will essentially allow us to show that for any benchmark point $\widetilde{\boldsymbol{u}} \in \mathcal{W}$ there is a sufficiently close u such that the penalty term in the regret (namely, $R(\boldsymbol{u}) - R(\boldsymbol{w}_1)$) is nicely bounded (see Eq. (19) in the proof). Finally, we turn to the way we choose the played action $\boldsymbol{a}_t \in \mathcal{A}$ and the construction of the estimator. For that we use the fact that, $\mathcal{D}_\Psi(\boldsymbol{w}, 1) \subseteq \mathcal{W} = \mathcal{A}$ for any $\boldsymbol{w} \in \mathcal{W}$. Now, let $\boldsymbol{v}$ be in the K-dimensional unit sphere, denoted by $\mathcal{S}$. For any $\boldsymbol{w} \in \mathcal{W}$, we have that $\boldsymbol{a} = \boldsymbol{w} + \left(\nabla^2\Psi(\boldsymbol{w})\right)^{-1/2}\boldsymbol{v} \in \mathcal{A}$ because $\|\boldsymbol{a} - \boldsymbol{w}\|_{\Psi,\boldsymbol{w}} = 1$. For adversarial linear bandits with delayed feedback, we use $\boldsymbol{a}_t = \boldsymbol{w}_t + \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1/2}\boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is sampled i.i.d. from the uniform distribution over the unit sphere. Note that $\boldsymbol{a}_t \in \mathcal{W}$ still holds. As for the loss estimate, we use $\widehat{\boldsymbol{\ell}}_t = K\boldsymbol{\ell}_t^\top \boldsymbol{a}_t \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{1/2}\boldsymbol{v}_t$, which can be seen to an unbiased estimator for $\boldsymbol{\ell}_t$ after observing that $\mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_t^\top|\mathcal{F}_t] = \frac{1}{K}\boldsymbol{I}$.

Theorem 8 *Let $\boldsymbol{u} \in \mathcal{W}$. Running Delayed FTRL for linear bandits (Algorithm 3) with $\gamma = \min\left\{\frac{1}{(64BK(1+d_{\max}))^2}, \sqrt{\frac{\nu\ln\left(1+\sqrt{T}\right)}{16(BK)^2T}}\right\}$ and $\eta = \min\left\{\frac{1}{(16d_{\max})^2}, \sqrt{\frac{B^2}{16D}}\right\}$ guarantees,*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top\boldsymbol{\ell}_t\right] \leq 14BK\sqrt{\nu T\ln(T)} + 8B\sqrt{D} + 2^{14}\nu B^2K^2(1+d_{\max})^2\ln(T).$$

The proof is similar to the proof of Theorem 4 and follows from an application of Corollary 3, which we can apply after verifying its assumptions (see the full proof in Appendix D).

Let us interpret the result of Theorem 8. It is known that a K-self-concordant barrier exists for all convex and closed $\mathcal{W}$ (Bubeck and Eldan, 2015; Chewi, 2021). Therefore, for $d_t = d$, our algorithm always guarantees a $O(K^{3/2}B\sqrt{T\ln(T)}) + B\sqrt{dT})$ bound for delayed bandit linear optimization. This makes our bound slightly suboptimal, as Ito et al. (2020a) show that the minimax regret is $\Theta(KB\sqrt{T\ln(T)}) + B\sqrt{dT})$. This trade-off between running time and regret bounds also exists in non-delayed linear bandits: algorithms that obtain the optimal regret bound (Bubeck and Eldan, 2015; Hazan and Karnin, 2016; Van der Hoeven et al., 2018; Ito et al., 2020b; Zimmert and Lattimore, 2022) have running time polynomial in both T and K, whereas slightly suboptimal algorithms in terms of regret, such as Scrible (Abernethy et al., 2008), have $O(K^3)$ runtime, given that a self-concordant barrier for the set of interest can be efficiently computed. For particular action sets, there exist algorithms that avoid this trade-off, see (Bubeck et al., 2012). There also exist domains for which ν can be considerably smaller. For example, $\Psi(\boldsymbol{w}) = -\log(1 - \|\boldsymbol{w}\|_2^2)$ is a 1-self-concordant barrier for the L_2 unit ball, in which case our bound is $O(KB\sqrt{T\ln(T)}) + B\sqrt{dT})$.

This matches the upper bound of Ito et al. (2020a) with less stringent assumptions, since Ito et al. (2020a) assume constant delays whereas Algorithm 3 can handle variable delays.

Efficient implementation. Abernethy et al. (2008, section 9) provide an approximation of FTRL with self-concordant barrier regularizers with a $O(K^3)$ per-round running time. Seemingly this implies that we can not use the results of Abernethy et al. (2008) to approximate our algorithm since our regularizer R is not a barrier, but only a self-concordant function on $\mathcal{W}$—see also (Nemirovski, 2004, Proposition 2.1.1). However, even though Abernethy et al. (2008) assume that the regularizer is a self-concordant barrier, the properties used to prove that the approximation has a suitable regret bound rely on properties of self-concordant functions—see (Nemirovski, 2004, Chapter 2, Statement IX). Thus, we can use the approximation of Abernethy et al. (2008) to obtain a $O(K^3)$ per round running time approximation of our algorithm.

7. Future Work and Discussion

We provided a new analysis of FTRL with delayed bandit feedback which leads to several new results for combinatorial semi-bandits, adversarial Markov decision processes, and linear bandits. The main downside of our approach is the additive $d_{\max}^2 \log(T)$ term in our bounds. Even though it is a lower order term, it prevents us from employing the skipping technique of Thune et al. (2019). Therefore, an important open problem is removing the $d_{\max}^2 \log(T)$ term and replacing it with $d_{\max} \log(T)$ or removing it altogether. Another direction for future work is MDPs with unknown transitions. Extending our ideas to that setting is not straightforward due the fact that standard analyses for this prwoblem use a changing domain $\mathcal{W}$. Finally, in the linear bandits setting it would be interesting to see whether the results for non-delayed feedback and the specific choice of $\mathcal{W}$ used by Bubeck et al. (2012) can be transferred to the delayed feedback setting.

Acknowledgments

This work was mostly done while DvdH was at the University of Milan partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and partially supported by Netherlands Organization for Scientific Research (NWO), grant number VI.Vidi.192.095. LZ and NCB are partially supported by the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence) and by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence). TL is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), the Yandex Initiative for Machine Learning at Tel Aviv University and a grant from the Tel Aviv University Center for AI and Data Science (TAD).

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

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

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

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*, pages 11349–11358, 2019.

Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In *Conference on Learning Theory*, pages 279–279, 2015.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622, 2016.

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.

Sinho Chewi. The entropic barrier is n-self-concordant. *arXiv preprint arXiv:2112.10947*, 2021.

Alon Cohen, Amit Daniely, Yoel Drori, Tomer Koren, and Mariano Schain. Asynchronous stochastic optimization robust to arbitrary delays. *Advances in Neural Information Processing Systems*, 34:9024–9035, 2021.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, and Alexandre Proutiere. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, 2015.

Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. *arXiv preprint arXiv:2205.13451*, 2022.

Varsha Dani and Thomas P Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *SODA*, volume 6, pages 937–943, 2006.

Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pages 512–520. PMLR, 2014.

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

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

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

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

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

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997, 2021.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10), 2007.

Elad Hazan and Zohar Karnin. Volumetric spanners: an efficient exploration basis for learning. *Journal of Machine Learning Research*, 2016.

Dirk van der Hoeven, Tim van Erven, and Wojciech Kotlowski. The many faces of exponential weights in online learning. In *Conference On Learning Theory*, 2018.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in episodic reinforcement learning. *arXiv preprint arXiv:2111.07615*, 2021.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. *arXiv preprint arXiv:2207.10786*, 2022.

Jiatai Huang, Yan Dai, and Longbo Huang. Banker online mirror descent: A universal approach for delayed online bandit learning. *arXiv preprint arXiv:2301.10500*, 2023.

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

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. In *Advances in Neural Information Processing Systems*, pages 2028–2038, 2020b.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869, 2020.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13172*, 2022.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

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

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138, 2020.

Satyen Kale, Lev Reyzin, and Robert E Schapire. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems*, 2010.

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

Wouter M Koolen, Manfred K Warmuth, and Jyrki Kivinen. Hedging structured concepts. In *Conference on Learning Theory*, pages 93–105, 2010.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978, 2021.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial Markov decision processes with delayed feedback. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 7281–7289, 2022.

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

Tor Lattimore, Branislav Kveton, Shuai Li, and Csaba Szepesvari. Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems*, 2018.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Advances in Neural Information Processing Systems*, pages 15522–15533, 2020.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems*, 2021.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. *arXiv preprint arXiv:2206.14906*, 2022.

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Conference on Learning Theory*, pages 109–123, 2004.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 42(16):3215–3224, 2004.

Arkadi S Nemirovski and Michael J Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

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

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

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

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

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2209–2218, 2019a.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486, 2019b.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *International Joint Conference on Artificial Intelligence*, pages 2936–2942, 2021.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613, 2020.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2019.

Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for adversarial bandit problems with multiple plays. In *International Conference on Algorithmic Learning Theory*, pages 375–389, 2010.

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

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*, pages 9712–9721, 2020.

Volodimir G Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 1990.

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

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

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

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2013.

Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312, 2022.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294, 2020.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Appendix A. Deferred Proofs of Analysis (Section 3)

Lemma 2 *Let R be convex and twice-differentiable such that $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. Let $\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} \leq \alpha \leq \frac{1}{16 d_{\max}}$ for all t and $\boldsymbol{w} \in \mathcal{W}$, and $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_t\|^2_{R,\boldsymbol{w}_t}\right] \leq \beta^2$ for all t, where $\boldsymbol{w}_t$ is given by (1). Suppose also that $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \frac{1}{64(1+d_{\max})}$ for all t with probability 1. Then for all $\boldsymbol{u} \in \mathcal{W}$*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 8\beta^2 T - \sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] + \sum_{t=1}^{T}\left(8\alpha^2|m_t| + 8\alpha\,\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R,\boldsymbol{w}_t}\right]\right).$$

Proof By Lemma 10, $\boldsymbol{w}_t \in \mathcal{D}_R(\boldsymbol{w}_\tau, \frac{1}{2})$ for all $\tau \in [t] \setminus o_t = m_t \cup \{t\} \subseteq \{t - d_{\max}, \dots, t\}$ and all t. We can use this to show

$$\sum_{\tau\in m_t}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_t} \leq \sum_{\tau\in[t]\setminus o_t}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_t} \leq 2\sum_{\tau\in[t]\setminus o_t}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_\tau} \leq \frac{1}{16},$$

by employing $4\nabla^2 R(\boldsymbol{w}_\tau) \succeq \nabla^2 R(\boldsymbol{w}_t)$ and the assumption on $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}$. By Lemma 9 and our assumptions on $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}$ and $\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}_t}$, we can thus conclude that $\boldsymbol{w}(\overline{\boldsymbol{L}}_t^m), \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m), \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$. By Hölder's inequality and Lemma 1

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{w}(\overline{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] &\leq \mathbb{E}\left[\|\boldsymbol{w}_t - \boldsymbol{w}(\overline{\boldsymbol{L}}_t^m)\|^\star_{R,\boldsymbol{w}_t}\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}_t}\right]\\ &\leq \mathbb{E}\left[8\|\widehat{\boldsymbol{L}}_t - \overline{\boldsymbol{L}}_t^m\|_{R,\boldsymbol{w}_t}\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}_t}\right]\\ &= \mathbb{E}\left[8\left\|\sum_{\tau\in m_t}\boldsymbol{\ell}_\tau\right\|_{R,\boldsymbol{w}_t}\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}_t}\right] \leq 8\alpha^2|m_t|, \end{aligned} \tag{7}$$

where the last inequality is due to the triangle inequality and the assumptions on $\|\boldsymbol{\ell}_\tau\|_{R,\boldsymbol{w}}$. Similarly we bound

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{w}(\overline{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] &\leq 8\alpha\,\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R,\boldsymbol{w}_t}\right]\\ \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t\right] &\leq 8\beta^2. \end{aligned} \tag{8}$$

By Lemma 11 we have that

$$\sum_{t=1}^{T}(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1). \tag{9}$$

Thus, by combining equations (2), (7), (8), and (9),

$$\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] = \sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t\right] - \sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right]$$
$$+ \sum_{t=1}^{T} \Big(\mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{w}(\overline{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] + \mathbb{E}\left[(\boldsymbol{w}(\overline{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] + \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t\right] \Big)$$
$$\leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 8\beta^2 T + \sum_{t=1}^{T} \left(8\alpha^2 |m_t| + 8\alpha\, \mathbb{E}\left[\Big\| \sum_{\tau \in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau) \Big\|_{R, \boldsymbol{w}_t} \right] \right)$$
$$- \sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] ,$$

which concludes the proof. ■

Corollary 3 *Under the same assumptions as in Lemma 2, suppose that* $\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau | \mathcal{F}_t] = \boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top \left(\nabla^2 R(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t))\right)^{-1} (\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,\middle|\, \mathcal{F}_t \right] = 0$ *for all* $t \in [T]$ *and all* $\tau, \tau' \in m_t$ *where* $\tau' \neq \tau$. *Then for all* $\boldsymbol{u} \in \mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t \right] \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 16\beta^2 T + 16\alpha^2 D \,.$$

Proof We are looking to control $\mathbb{E}\left[\| \sum_{\tau \in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau) \|_{R, \boldsymbol{w}_t} \right]$ for a given $t \in [T]$. We start by considering

$$\mathbb{E}\left[\Big\| \sum_{\tau \in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau) \Big\|_{R, \boldsymbol{w}_t}^2 \right] = \sum_{\tau \in m_t} \mathbb{E}\left[\|\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau\|_{R, \boldsymbol{w}_t}^2 \right]$$
$$= \sum_{\tau \in m_t} \left(\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|_{R, \boldsymbol{w}_t}^2 \right] - \mathbb{E}\left[\|\boldsymbol{\ell}_\tau\|_{R, \boldsymbol{w}_t}^2 \right] \right)$$
$$\leq \sum_{\tau \in m_t} \mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|_{R, \boldsymbol{w}_t}^2 \right] ,$$

where we used that $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top \left(\nabla^2 R(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t))\right)^{-1} (\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,|\, \mathcal{F}_t \right] = 0$ for $\tau \neq \tau'$ in the first equality, and that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau \,|\, \mathcal{F}_t] = \boldsymbol{\ell}_\tau$ in the second equality. In turn, the above together with Jensen's inequality implies that

$$\mathbb{E}\left[\Big\| \sum_{\tau \in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau) \Big\|_{R, \boldsymbol{w}_t} \right] \leq \sqrt{ \sum_{\tau \in m_t} \mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|_{R, \boldsymbol{w}_t}^2 \right] }$$
$$\leq \sqrt{ 4 \sum_{\tau \in m_t} \mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|_{R, \boldsymbol{w}_\tau}^2 \right] } \leq \sqrt{4|m_t|\beta^2} \,, \tag{10}$$

where in the second inequality we used Lemma 10 together with $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. Finally, the third inequality of 10 is due to the assumptions of Lemma 2. We conclude by substituting this bound into the results of Lemma 2,

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] &\leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 8\beta^2 T + \sum_{t=1}^{T}\left(8\alpha^2|m_t| + 16\sqrt{|m_t|\alpha^2\beta^2}\right) \\ &\leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 16\beta^2 T + 16\alpha^2 \sum_{t=1}^{T}|m_t|,\end{aligned}$$

where in the last inequality we used that $\sqrt{ab} \leq \frac{1}{2}(a+b)$ for $a, b > 0$. ■

Lemma 1 *Suppose that $\nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$, $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$, and for some twice-differentiable convex R. Let $\boldsymbol{x} \in \mathcal{W}$ and $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^K$ such that $\boldsymbol{w}(\boldsymbol{L}'), \boldsymbol{w}(\boldsymbol{L}) \in \mathcal{D}_R(\boldsymbol{x}, \frac{1}{2})$, then $\|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^\star_{R,\boldsymbol{x}} \leq 8\|\boldsymbol{L}' - \boldsymbol{L}\|_{R,\boldsymbol{x}}$.*

Proof By Taylor's theorem and the optimality of $\boldsymbol{w}(\boldsymbol{L}')$ we have that for some ζ on the line segment between $\boldsymbol{w}(\boldsymbol{L}')$ and $\boldsymbol{w}(\boldsymbol{L})$

$$\begin{aligned}&\boldsymbol{L}'^\top \boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L})) - \boldsymbol{L}'^\top \boldsymbol{w}(\boldsymbol{L}') - R(\boldsymbol{w}(\boldsymbol{L}')) \\ &\geq \frac{1}{2}(\boldsymbol{w}(\boldsymbol{L}') - \boldsymbol{w}(\boldsymbol{L}))^\top \nabla^2 R(\zeta)(\boldsymbol{w}(\boldsymbol{L}') - \boldsymbol{w}(\boldsymbol{L})) \\ &\geq \frac{1}{8}(\boldsymbol{w}(\boldsymbol{L}') - \boldsymbol{w}(\boldsymbol{L}))^\top \nabla^2 R(\boldsymbol{x})(\boldsymbol{w}(\boldsymbol{L}') - \boldsymbol{w}(\boldsymbol{L})),\end{aligned}$$

where the last inequality is due the assumption on $\nabla^2 R(\boldsymbol{w})$, which is applicable because if $\boldsymbol{w}(\boldsymbol{L}'), \boldsymbol{w}(\boldsymbol{L}) \in \mathcal{D}_R(\boldsymbol{x}, \frac{1}{2})$ it implies that the line segment between $\boldsymbol{w}(\boldsymbol{L}')$ and $\boldsymbol{w}(\boldsymbol{L})$ is also in $\mathcal{D}_R(\boldsymbol{x}, \frac{1}{2})$. Thus $\zeta \in \mathcal{D}_R(\boldsymbol{x}, \frac{1}{2})$.

By Taylor's theorem we have that

$$\begin{aligned}&\boldsymbol{L}'^\top \boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L})) - \boldsymbol{L}'^\top \boldsymbol{w}(\boldsymbol{L}') - R(\boldsymbol{w}(\boldsymbol{L}')) \\ &= (\boldsymbol{L}' - \boldsymbol{L})^\top(\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')) + \boldsymbol{L}^\top \boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L})) - \boldsymbol{L}^\top \boldsymbol{w}(\boldsymbol{L}') - R(\boldsymbol{w}(\boldsymbol{L}')) \\ &\leq (\boldsymbol{L}' - \boldsymbol{L})^\top(\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')) \\ &\leq \|\boldsymbol{L}' - \boldsymbol{L}\|_{R,\boldsymbol{x}} \|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^\star_{R,\boldsymbol{x}}\,,\end{aligned}$$

where the first inequality is due to the optimality of $\boldsymbol{w}(\boldsymbol{L})$ and the second inequality is Hölder's inequality. Thus, we may conclude that

$$\|\boldsymbol{L}' - \boldsymbol{L}\|_{R,\boldsymbol{x}} \|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^\star_{R,\boldsymbol{x}} \geq \frac{1}{8}\left(\|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^\star_{R,\boldsymbol{x}}\right)^2,$$

which concludes the proof after multiplying both sides of the above inequality by $\frac{8}{\|\boldsymbol{w}(\boldsymbol{L}) - \boldsymbol{w}(\boldsymbol{L}')\|^\star_{R,\boldsymbol{x}}}$.
■

Lemma 9 *Suppose that $\nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$ and for R strictly convex and twice differentiable. Let $z \subset \mathbb{N}$ be a finite set, and define $\boldsymbol{L}' = \boldsymbol{L} + \sum_{\tau \in z} \boldsymbol{y}_\tau$, where $\boldsymbol{y}_\tau \in \mathbb{R}^K$. If $\sum_{\tau \in z} \|\boldsymbol{y}_\tau\|_{R,\boldsymbol{w}(\boldsymbol{L})} \leq \frac{1}{16}$, then $\boldsymbol{w}(\boldsymbol{L}') \in \mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$.*

Proof Because of the strict convexity of R, to show that $\boldsymbol{w}(\boldsymbol{L}') \in \mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ it suffices to show that for all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$

$$\boldsymbol{L}'^{\top}\boldsymbol{x} + R(\boldsymbol{x}) \geq \boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L})). \tag{11}$$

To see why the strict convexity of R is sufficient, suppose that all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ indeed satisfy equation (11). For the sake of contradiction suppose that $\boldsymbol{w}(\boldsymbol{L}')$ is not in $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$. Let $\boldsymbol{z} = (1-a)\boldsymbol{w}(\boldsymbol{L}) + a\boldsymbol{w}(\boldsymbol{L}')$ be the point on the boundary of $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ on the segment between $\boldsymbol{w}(\boldsymbol{L})$ and $\boldsymbol{w}(\boldsymbol{L}')$. Then

$$\begin{aligned}
\boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L})) &\leq \boldsymbol{L}'^{\top}\boldsymbol{z} + R(\boldsymbol{z}) \\
&< (1-a)(\boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L}))) + a(\boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}') + R(\boldsymbol{w}(\boldsymbol{L}'))) \\
&\leq \boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) + R(\boldsymbol{w}(\boldsymbol{L}))\ ,
\end{aligned}$$

where the last inequality is by definition of $\boldsymbol{w}(\boldsymbol{L}')$. Thus, we have a contradiction, which implies that if all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ satisfy equation (11) then $\boldsymbol{w}(\boldsymbol{L}') \in \mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$.

We proceed by showing that all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_R(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ satisfy equation (11). Denote by $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{w}(\boldsymbol{L})$. Using Taylor's theorem, there exists ζ on the segment between $\boldsymbol{x}$ and $\boldsymbol{w}(\boldsymbol{L})$ such that

$$\begin{aligned}
&\boldsymbol{L}'^{\top}\boldsymbol{x} + R(\boldsymbol{x}) - \boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) - R(\boldsymbol{w}(\boldsymbol{L})) \\
&= (\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + (\boldsymbol{L} + \nabla R(\boldsymbol{w}(\boldsymbol{L})))^{\top}\boldsymbol{h} + \tfrac{1}{2}\boldsymbol{h}^{\top}\nabla^2 R(\zeta)\boldsymbol{h} \\
&\geq (\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + \tfrac{1}{2}\boldsymbol{h}^{\top}\nabla^2 R(\zeta)\boldsymbol{h} \\
&\geq (\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + \tfrac{1}{8}\boldsymbol{h}^{\top}\nabla^2 R(\boldsymbol{w}(\boldsymbol{L}))\boldsymbol{h}
\end{aligned} \tag{12}$$

where the first inequality is due to the optimality of $\boldsymbol{w}(\boldsymbol{L})$ and the second inequality is because $\zeta, \boldsymbol{w}(\boldsymbol{L}) \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. Thus, by applying Hölder's inequality we can see that

$$\begin{aligned}
&\boldsymbol{L}'^{\top}\boldsymbol{x} + R(\boldsymbol{x}) - \boldsymbol{L}'^{\top}\boldsymbol{w}(\boldsymbol{L}) - R(\boldsymbol{w}(\boldsymbol{L})) \\
&\geq (\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + \tfrac{1}{8}\boldsymbol{h}^{\top}\nabla^2 R(\boldsymbol{w}(\boldsymbol{L}))\boldsymbol{h} \\
&\geq -\sum_{\tau \in z} \|\boldsymbol{y}_\tau\|_{R,\boldsymbol{w}(\boldsymbol{L})} \|\boldsymbol{h}\|^{\star}_{R,\boldsymbol{w}(\boldsymbol{L})} + \tfrac{1}{8}\boldsymbol{h}^{\top}\nabla^2 R(\boldsymbol{w}(\boldsymbol{L}))\boldsymbol{h} \\
&= -\frac{1}{2}\sum_{\tau \in z} \|\boldsymbol{y}_\tau\|_{R,\boldsymbol{w}(\boldsymbol{L})} + \tfrac{1}{32} \\
&\geq 0
\end{aligned}$$

where the equality is due to the fact that $\|\boldsymbol{h}\|^{\star}_{R,\boldsymbol{w}(\boldsymbol{L})} = \frac{1}{2}$, as $\boldsymbol{x}$ is on the boundary of $\mathcal{D}(\boldsymbol{w}(\boldsymbol{L}), \frac{1}{2})$ and the final inequality is due to the assumption that $\sum_\tau \|\boldsymbol{y}_\tau\|_{R,\boldsymbol{w}(\boldsymbol{L})} \leq \frac{1}{16}$. ■

Lemma 10 *Suppose that* $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ *for all* $\boldsymbol{w} \in \mathcal{W}$ *and* $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. *Also suppose that* $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \frac{1}{64(1+d_{\max})}$ *for all* t *with probability 1. Then, for all* t *and all* $\delta \in [d_{\max}+1]$, *we have that* $\boldsymbol{w}_{t+\delta} \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$.

Proof We prove the statement by induction on t. For the base case, we need to show that for all $\delta \in [1+d_{\max}]$, $\boldsymbol{w}_{1+\delta} \in \mathcal{D}_R(\boldsymbol{w}_1, \frac{1}{2})$. We start by showing that $\boldsymbol{w}_2 \in \mathcal{D}_R(\boldsymbol{w}_1, \frac{1}{2})$ where $\widehat{\boldsymbol{L}}_2 = \widehat{\boldsymbol{L}}_1 + \widehat{\boldsymbol{\ell}}_1 \mathbb{I}\{1 \in o_2\}$. Since $\|\widehat{\boldsymbol{\ell}}_1\|_{R,\boldsymbol{w}_1} \leq \frac{1}{16}$ by assumption, this follows from Lemma 9. Now, $\widehat{\boldsymbol{L}}_3 = \widehat{\boldsymbol{L}}_1 + \widehat{\boldsymbol{\ell}}_1 \mathbb{I}\{1 \in o_3 \setminus o_2\} + \widehat{\boldsymbol{\ell}}_2 \mathbb{I}\{2 \in o_3\}$. Since $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$ and $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \frac{1}{64(1+d_{\max})}$ for all t, we have that

$$\|\widehat{\boldsymbol{\ell}}_1\|_{R,\boldsymbol{w}_1} + \|\widehat{\boldsymbol{\ell}}_2\|_{R,\boldsymbol{w}_1} \leq \|\widehat{\boldsymbol{\ell}}_1\|_{R,\boldsymbol{w}_1} + 2\|\widehat{\boldsymbol{\ell}}_2\|_{R,\boldsymbol{w}_2} \leq \frac{1}{16}$$

and Lemma 9 implies that $\boldsymbol{w}_3 \in \mathcal{D}_R(\boldsymbol{w}_1, \frac{1}{2})$. We can now repeat this argument to show that $\boldsymbol{w}_{1+\delta} \in \mathcal{D}_R(\boldsymbol{w}_1, \frac{1}{2})$ for $\delta \in [d_{\max}+1]$, which completes the proof for the base case.

For the induction step, assume that $\boldsymbol{w}_t \in \mathcal{D}_R(\boldsymbol{w}_\tau, \frac{1}{2})$ for $\tau \in \{t-d_{\max}-1, \dots, t-1\}$. Recall that $\widehat{\boldsymbol{L}}_{t+1} = \widehat{\boldsymbol{L}}_t + \sum_{\tau \in o_{t+1} \setminus o_t} \widehat{\boldsymbol{\ell}}_\tau$. Since $o_{t+1} \setminus o_t \subseteq \{t-d_{\max}, \dots, t-1, t\}$, we have that

$$\sum_{\tau \in o_{t+1} \setminus o_t} \|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_t} \leq 2 \sum_{\tau \in o_{t+1} \setminus o_t} \|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_\tau} \leq \frac{|o_{t+1} \setminus o_t|}{32(d_{\max}+1)} \leq \frac{1}{32}\,,$$

where in the first inequality we used that $\nabla^2 R(\boldsymbol{w}_t) \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w}_\tau)$ if $\boldsymbol{w}_t \in \mathcal{D}_R(\boldsymbol{w}_\tau, \frac{1}{2})$ which for $\tau \in \{t-d_{\max}, \dots, t-1\}$ is true by the inductive assumption, and in the second inequality we used the assumption $\|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_\tau} \leq \frac{1}{64(1+d_{\max})}$. Then Lemma 9 implies $\boldsymbol{w}_{t+1} \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$ which in turn implies $\nabla^2 R(\boldsymbol{w}_t) \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w}_{t+1})$. Using this last inequality, we can see that

$$\begin{aligned}\sum_{\tau \in o_{t+2} \setminus o_t} \|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_t} &\leq \sum_{\tau \in o_{t+2} \setminus (o_t \cup \{t+1\})} \|\widehat{\boldsymbol{\ell}}_\tau\|_{R,\boldsymbol{w}_t} + \|\widehat{\boldsymbol{\ell}}_{t+1}\|_{R,\boldsymbol{w}_t} \\ &\leq \frac{1}{32} + 2\|\widehat{\boldsymbol{\ell}}_{t+1}\|_{R,\boldsymbol{w}_{t+1}} \leq \frac{1}{16}\,.\end{aligned}$$

Thus, by Lemma 9, $\boldsymbol{w}_{t+2} \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$. We can repeat this argument to show that for all $\delta \in [1+d_{\max}]$, $\boldsymbol{w}_{t+\delta} \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$.

■

Lemma 11 (Be-The-Leader Lemma) *For any fixed* $\boldsymbol{u} \in \mathcal{W}$ *we have that*

$$\sum_{t=1}^{T} \widehat{\boldsymbol{\ell}}_t^\top (\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u}) \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1)$$

Proof We will prove the statement by induction on T. The base case holds by definition of $\boldsymbol{w}_1$. For the induction step, assume that

$$\sum_{t=1}^{T-1} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) + R(\boldsymbol{w}_1) \leq \sum_{t=1}^{T-1} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{w} + R(\boldsymbol{w})$$

for any $\boldsymbol{w} \in \mathcal{W}$. Adding $\widehat{\boldsymbol{\ell}}_T^\top \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star)$ to both sides of the above inequality and setting $\boldsymbol{w} = \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star)$ on the right-hand side of the above inequality we find

$$\begin{aligned}\sum_{t=1}^{T} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^\star) + R(\boldsymbol{w}_1) &\leq \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{W}} \sum_{t=1}^{T} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{w} + R(\boldsymbol{w}) \\ &\leq \sum_{t=1}^{T} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{u} + R(\boldsymbol{u}),\end{aligned}$$

which proves the statement after reordering the above inequality. ■

Appendix B. Deferred Proof for Combinatorial Semi-Bandits (Section 4)

Theorem 4 *Let* $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_1 \leq B$. *Running Delayed FTRL for combinatorial bandits (Algorithm 1) with* $\gamma = \frac{1}{64^2 B(1+d_{\max})^2}$ *and* $\eta = \min\left\{\frac{1}{16^2 B d_{\max}^2}, \sqrt{\frac{B\left(1+\ln\left(\frac{K}{B}\right)\right)}{16(KT+BD)}}\right\}$ *guarantees,*

$$\mathcal{R}_T \leq 12\sqrt{B(KT+BD)\ln\left(\frac{K}{B}\right)} + 64^2 B(1+d_{\max})^2 K\ln(T) + 2^9 B^2 d_{\max}^2 \ln\left(\frac{K}{B}\right).$$

Proof We are looking to apply Corollary 3. We start by showing that indeed $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w}\in\mathcal{W}$ and $\boldsymbol{w}'\in\mathcal{D}_R(\boldsymbol{w},\frac{1}{2})$. With ϕ, as defined in Lemma 16,

$$\left(\nabla^2 R(\boldsymbol{w})\right)(i,i) = \frac{1}{\gamma \boldsymbol{w}^2(i)} + \frac{1}{\eta \boldsymbol{w}(i)} \geq \frac{1}{\gamma \boldsymbol{w}^2(i)} = \left(\nabla^2\phi(\boldsymbol{w})\right)(i,i)\,,$$

and we conclude that $\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2\phi(\boldsymbol{w})$ as both matrices are diagonal. Together with $\mathcal{W}\subseteq\mathbb{R}_+^K$, R being strictly convex, and $\gamma\in(0,1)$, we are in a position to apply Lemma 16 to conclude that $\frac{1}{2}\boldsymbol{w}(i) \leq \boldsymbol{w}'(i) \leq 2\boldsymbol{w}(i)$ for all $\boldsymbol{w}'\in\mathcal{D}_R(\boldsymbol{w},\frac{1}{2})$ and all $i\in[K]$. We can use this fact to show that

$$\left(\nabla^2 R(\boldsymbol{w}')\right)(i,i) = \frac{1}{\gamma \boldsymbol{w}'^2(i)} + \frac{1}{\eta \boldsymbol{w}'(i)} \geq \frac{1}{\gamma 4\boldsymbol{w}^2(i)} + \frac{1}{\eta 2\boldsymbol{w}(i)} \geq \frac{1}{4}\left(\nabla^2 R(\boldsymbol{w})\right)(i,i)$$

and

$$\left(\nabla^2 R(\boldsymbol{w}')\right)(i,i) = \frac{1}{\gamma \boldsymbol{w}'^2(i)} + \frac{1}{\eta \boldsymbol{w}'(i)} \leq \frac{1}{\gamma \frac{1}{4}\boldsymbol{w}^2(i)} + \frac{1}{\eta \frac{1}{2}\boldsymbol{w}(i)} \leq 4\left(\nabla^2 R(\boldsymbol{w})\right)(i,i)\,.$$

With $\nabla^2 R(\boldsymbol{w})$ being a diagonal matrix, we can conclude that $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w}\in\mathcal{W}$ and $\boldsymbol{w}'\in\mathcal{D}_R(\boldsymbol{w},\frac{1}{2})$.

As the next step, we find an appropriate α such that $\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} \leq \alpha \leq \frac{1}{16 d_{\max}}$. For that we use $\sum_{i=1}^K \boldsymbol{w}(i) \leq B$, $\|\boldsymbol{\ell}\|_\infty \leq 1$, and $\eta + \gamma\boldsymbol{w}(i) \geq \gamma\boldsymbol{w}(i)$ to bound

$$\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} = \sqrt{\sum_{i=1}^K \boldsymbol{\ell}_t(i)^2 \frac{\eta\gamma\boldsymbol{w}^2(i)}{\eta+\gamma\boldsymbol{w}(i)}} \leq \underbrace{\sqrt{\eta B}}_{\alpha}\,. \tag{13}$$

and it is clear that $\alpha \leq \frac{1}{16 d_{\max}}$, by $\eta \leq \frac{1}{16^2 B d_{\max}^2}$. Next is $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2\right] \leq \beta^2$, for which we commence by using the tower rule.

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2\right] = \mathbb{E}_{\mathcal{F}_t}\left[\mathbb{E}_{\boldsymbol{a}_t}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2 | \mathcal{F}_t\right]\right]\,.$$

Next we consider $\mathbb{E}_{\boldsymbol{a}_t}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2 | \mathcal{F}_t\right]$ in isolation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}_t}\left[\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t}^2 | \mathcal{F}_t\right] &= \mathbb{E}_{\boldsymbol{a}_t}\left[\sum_{i=1}^K \left(\frac{\boldsymbol{a}_t(i)\boldsymbol{\ell}_t(i)}{\boldsymbol{w}_t(i)}\right)^2 \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1}(i,i) | \mathcal{F}_t\right] \\ &= \sum_{i=1}^K \frac{\boldsymbol{\ell}_t(i)^2}{\boldsymbol{w}_t(i)} \frac{\eta\gamma\boldsymbol{w}_t^2(i)}{\eta+\gamma\boldsymbol{w}_t(i)} \leq \underbrace{\eta K}_{\beta^2}\,,\end{aligned} \tag{14}$$

where we also used that $\mathbb{E}_{\boldsymbol{a}_t}[\boldsymbol{a}_t] = \boldsymbol{w}_t$, and $\eta + \gamma \boldsymbol{w}(i) \geq \gamma \boldsymbol{w}(i)$. Next is showing that $\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \frac{1}{64(1+d_{\max})}$:

$$\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}(\widehat{L}_t)} = \sqrt{\sum_{i=1}^{K} \left(\frac{\boldsymbol{a}_t(i)\boldsymbol{\ell}_t(i)}{\boldsymbol{w}_t(i)}\right)^2 \frac{\eta\gamma \boldsymbol{w}_t^2(i)}{\eta + \gamma \boldsymbol{w}_t(i)}} \leq \sqrt{\gamma B} \leq \frac{1}{64(1+d_{\max})} ,$$

where we used that $\gamma \leq \frac{1}{64^2 B(1+d_{\max})^2}$. Finally we show that $\widehat{\boldsymbol{\ell}}_\tau$ and $\widehat{\boldsymbol{\ell}}_{\tau'}$ are independent for all $\tau, \tau' \in m_t$ where $\tau' \neq \tau$. Recall that $\widehat{\boldsymbol{\ell}}_\tau(i) = \frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(i)}$, for all i. Conditioned on the observed history $\mathcal{F}_t$, the only random element of $\widehat{\boldsymbol{\ell}}_{\tau'}$ is $\boldsymbol{a}_{\tau'} \sim \boldsymbol{p}_{\tau'}$. Since $\widehat{\boldsymbol{\ell}}_\tau$ can not be used in round t to compute $\boldsymbol{p}_{\tau'}$ we have that $\widehat{\boldsymbol{\ell}}_{\tau'}$ is independent of $\widehat{\boldsymbol{\ell}}_\tau$. We conclude that

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1} (\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,\middle|\, \mathcal{F}_t\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{\ell}}_\tau}\left[\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top \left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1} (\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,\middle|\, \mathcal{F}_t, \widehat{\boldsymbol{\ell}}_\tau\right]\right] = 0
\end{aligned}$$

where we used that $\widehat{\boldsymbol{\ell}}_{\tau'}$ is and unbiased estimator of $\boldsymbol{\ell}_{\tau'}$. Now we are able to apply Corollary 3, from which it follows that for any $\boldsymbol{u} \in \mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 16\eta KT + 16\eta BD , \tag{15}$$

by substituting α from equation (13) and β^2 from equation (14). The next step is finding an appropriate bound on R. We can do this for all $\boldsymbol{u} \in \mathcal{W}$ for the negative entropy component as follows

$$\begin{aligned}
-\sum_{i=1}^{K} \boldsymbol{u}(i) \ln \boldsymbol{u}(i) &= \|\boldsymbol{u}\|_1 \sum_{i=1}^{K} \frac{\boldsymbol{u}(i)}{\|\boldsymbol{u}\|_1} \ln \frac{1}{\boldsymbol{w}(i)} \\
&\leq \|\boldsymbol{u}\|_1 \ln\left(\sum_{i=1}^{K} \frac{\boldsymbol{u}(i)}{\|\boldsymbol{u}\|_1} \frac{1}{\boldsymbol{u}(i)}\right) \\
&\leq \|\boldsymbol{u}\|_1 \ln\left(\frac{K}{\|\boldsymbol{u}\|_1}\right) + \|\boldsymbol{u}\|_1 \leq B\left(1 + \ln\left(\frac{K}{B}\right)\right),
\end{aligned} \tag{16}$$

where we used Jensen's inequality in the second step and the fact that $x \ln(\frac{K}{x}) + x$ is in increasing on $x \in [1, K]$ in the last inequality. The negative logarithm component however is unbounded and tends to infinity when any element of $\boldsymbol{u}$ tends to 0. Thus we cannot compare to $\boldsymbol{a}^*$ directly, which might lie on the boundary on $\mathcal{W}$. Instead we define $\boldsymbol{u} = (1-\theta)\boldsymbol{a}^* + \theta \boldsymbol{w}_1$ for an $\theta \in [0, 1]$. θ now acts as a trade-off between an upper bound on the regularizer and an additional bias-like term that stems from comparing $\boldsymbol{a}^*$ to $\boldsymbol{u}$ in terms of pseudo-regret. We now bound the negative logarithm of $\boldsymbol{u}$ by simply using $(1-\theta)\boldsymbol{a}^*(i) \geq 0$

$$\ln(\boldsymbol{w}_1(i)) - \ln(\boldsymbol{u}(i)) = \ln\left(\frac{\boldsymbol{w}_1(i)}{(1-\theta)\boldsymbol{a}^*(i) + \theta \boldsymbol{w}_1(i)}\right) \leq \ln\left(\frac{1}{\theta}\right). \tag{17}$$

Combining equation (16) and equation (17) allows us to bound $R(\boldsymbol{u}) - R(\boldsymbol{w}_1)$

$$\begin{aligned} R(\boldsymbol{u}) - R(\boldsymbol{w}_1) &= \sum_{i=1}^{K} \left(\frac{\boldsymbol{u}(i)}{\eta} \ln(\boldsymbol{u}(i)) - \frac{1}{\gamma} \ln(\boldsymbol{u}(i)) \right) - \sum_{i=1}^{K} \left(\frac{\boldsymbol{w}_1(i)}{\eta} \ln(\boldsymbol{w}_1(i)) - \frac{1}{\gamma} \ln(\boldsymbol{w}_1(i)) \right) \\ &\leq \frac{B\left(1 + \ln\left(\frac{K}{B}\right)\right)}{\eta} + \frac{K \ln\left(\frac{1}{\theta}\right)}{\gamma} , \end{aligned} \tag{18}$$

where we applied that $\ln(x) \leq 0$ for all $x \in (0, 1)$. To finish the proof, we start from the regret

$$\begin{aligned} \mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + \theta\, \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{w}_1 - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + 2\theta BT , \end{aligned}$$

where we bound $(\boldsymbol{w}_1 - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t \leq 2B$ in the inequality. We continue by using equation (15) and equation (18)

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + 2\theta BT &\leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 16\eta KT + 16\eta BD + 2\theta BT \\ &\leq \frac{B\left(1 + \ln\left(\frac{K}{B}\right)\right)}{\eta} + \frac{K \ln\left(\frac{1}{\theta}\right)}{\gamma} + 16\eta KT + 16\eta BD + 2\theta BT \\ &= \frac{B\left(1 + \ln\left(\frac{K}{B}\right)\right)}{\eta} + \frac{K \ln\left(T\right)}{\gamma} + 16\eta KT + 16\eta BD + 2B , \end{aligned}$$

where in the equality we set $\theta = \frac{1}{T}$. Substituting

$$\gamma = \frac{1}{64^2 B(1 + d_{\max})^2} \qquad \text{and} \qquad \eta = \min\left\{ \frac{1}{16^2 B d_{\max}^2}, \frac{\sqrt{B\left(1 + \ln\left(\frac{K}{B}\right)\right)}}{4\sqrt{(BD + KT)}} \right\}$$

yields

$$\begin{aligned} \mathcal{R}_T \leq\ & 8\sqrt{B\left(1 + \ln\left(\frac{K}{B}\right)\right)(KT + BD)} \\ & + 64^2 B(1 + d_{\max})^2 K \ln\left(T\right) + 2^8 B^2 d_{\max}^2 \left(1 + \ln\left(\frac{K}{B}\right)\right) . \end{aligned}$$

■

Theorem 5 *Suppose that $d_t = d$ for all t and that $B \leq K/2$. Then for any algorithm there exists a sequence of losses such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{a}^\star)^\top \boldsymbol{\ell}_t\right] = \Omega\left(\max\left\{\sqrt{BKT}, B\sqrt{dT}\right\}\right) .$$

Proof By Audibert et al. (2014), we have that any algorithm without delay must suffer at least $\Omega(\sqrt{BKT})$ regret in the combinatorial semi-bandit setting.

Next, we assume full information feedback, which is easier from the point of view of the algorithm. We take inspiration from Langford et al. (2009, Lemma 3). For simplicity we will assume that T/d is an integer. We divide the T rounds into T/d blocks of d rounds. We take the losses of the lower bound for B-sets in (Koolen et al., 2010, Section 4), which states that any algorithm in the full information setting must suffer at least $\Omega(B\sqrt{T'})$ regret after T' rounds. We take the loss of the first round of the lower bound (Koolen et al., 2010) and copy it d times, which we use as the losses for the first block. We repeat this process for the remaining blocks. Since the algorithm can not respond to the copied losses, we must have that any algorithm must suffer at least $\Omega(dB\sqrt{T/d}) = \Omega(B\sqrt{dT})$ regret, which completes the proof. ■

Appendix C. Deferred Proofs for Adversarial MDPs (Section 5)

Theorem 7 *Running Delayed FTRL for adversarial MDPs (Algorithm 2) with* $\gamma = \frac{1}{4096H(1+d_{max})^2}$ *and* $\eta = \min\left\{\frac{1}{256H(1+d_{max})^2}, \frac{1}{\sqrt{(SAT+D)\log(HSAT)}}\right\}$ *guarantees*

$$\mathbb{E}[\mathcal{R}_T] \le 10H\sqrt{SAT\log(HSAT)} + 10H\sqrt{D\log(HSAT)} + 7\cdot 10^5 H^2S^2A(1+d_{max})^2.$$

Proof First, we decompose

$$\mathcal{R}_T = \sum_{t=1}^{T}\langle \boldsymbol{w}^{\pi_t} - \boldsymbol{w}^\star, \boldsymbol{\ell}_t\rangle = \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}^{\pi_t} - \boldsymbol{w}_t, \boldsymbol{\ell}_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star, \boldsymbol{\ell}_t\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^{T}\langle \widetilde{\boldsymbol{w}}^\star - \boldsymbol{w}^\star, \boldsymbol{\ell}_t\rangle}_{\textsc{Shift-penalty}},$$

where, by Lemma 6, Error is bounded by $2H$ and $\widetilde{w}^\star \in \mathcal{W}$ exists such that Shift-penalty is bounded by $2H$. For Reg we use Lemma 2. Much like in the proof of Lemma 4, $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$. For any t and $\boldsymbol{w} \in \mathcal{W}$

$$\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} \le \sqrt{\eta\sum_{h,s,a}\boldsymbol{w}_h(s,a)\boldsymbol{\ell}_h(s,a)^2} \le \sqrt{\eta\sum_{h,s,a}\boldsymbol{w}_h(s,a)} = \sqrt{\eta H} =: \alpha \le \frac{1}{16(1+d_{max})},$$

where the last inequality is by the definition of η. For any t,

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{\ell}}_t\|^2_{R,\boldsymbol{w}_t}] &= \eta\mathbb{E}\left[\sum_{h,s,a}\boldsymbol{w}_{t,h}(s,a)\widehat{\boldsymbol{\ell}}_{t,h}(s,a)^2\right] \le \eta\mathbb{E}\left[\sum_{h,s,a}\frac{\mathbb{E}\left[\mathbb{I}\{s_{t,h}=s, a_{t,h}=a\}\mid \mathcal{F}_t\right]}{\boldsymbol{u}_{t,h}(s,a)}\right]\\ &= \eta\mathbb{E}\left[\sum_{h,s,a}\frac{\boldsymbol{w}_h^{\pi_t}(s,a)}{\boldsymbol{u}_{t,h}(s,a)}\right] \le \eta HSA =: \beta^2,\end{aligned}$$

where the inequalities follow since $\boldsymbol{u}_{t,h}(s,a) = \max_{\widehat{p}\in\mathcal{P}}\boldsymbol{w}_h^{\pi_\tau,\widehat{p}}(s,a) \ge \max\{\boldsymbol{w}_{t,h}(s,a), \boldsymbol{w}_h^{\pi_t}(s,a)\}$. Finally, for all t,

$$\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \le \sqrt{\gamma\sum_{h,s,a}\boldsymbol{w}_{t,h}(s,a)^2\widehat{\boldsymbol{\ell}}_h(s,a)^2} \le \sqrt{\gamma\sum_{h,s,a}\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}} = \sqrt{\gamma H} \le \frac{1}{64(1+d_{max})},$$

where the second is since $\boldsymbol{u}_{t,h}(s,a) \ge \boldsymbol{w}_h^{\pi_t}(s,a)$ and the last is by definition of γ. Thus, applying Lemma 2 with $b_t = \mathbb{E}[\widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t \mid \mathcal{F}_t]$, we get

$$\begin{aligned}\textsc{Reg} \le{}& \underbrace{R(\widetilde{w}^\star) - R(w_1)}_{\textsc{Penalty}} + 8\eta HSAT + 8\eta H(T+D)\\ &+ \underbrace{\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)^\top(\boldsymbol{\ell}_t - \widehat{\boldsymbol{\ell}}_t)]}_{\textsc{Bias}_1} + \underbrace{\sum_{t=1}^{T}\mathbb{E}[\widetilde{\boldsymbol{w}^\star}^\top(\widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t)]}_{\textsc{Bias}_2} + 8\sqrt{\eta H}\underbrace{\sum_{t=1}^{T}\mathbb{E}[\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\|_{R,\boldsymbol{w}_t}]}_{\textsc{Drift}}.\end{aligned}$$

Using standard arguments, PENRALTY $\leq \frac{4HS^2A\log(HSAT)}{\gamma}+\frac{2H\log(SAT)}{\eta}$ since $\boldsymbol{w}_1, \widetilde{\boldsymbol{w}}^\star \in \Omega$. Recall that by definition $\boldsymbol{u}_{t,h}(s,a) \geq \boldsymbol{w}_h^{\pi_t}(s,a)$. Thus, $\mathbb{E}[\widehat{\boldsymbol{\ell}}_{t,h}(s,a) \mid \mathcal{F}_t] \leq \boldsymbol{\ell}_t$ and $\text{BIAS}_2 \leq 0$. BIAS_1 is the bias of the estimator due to the fact that we use an upper confidence bound on the occupancy measure instead of the actual occupancy measure. By Lemma 12 we have $\text{BIAS}_1 \leq 8H^2S$. Finally, for the DRIFT term, for each t,

$$\begin{aligned}
\mathbb{E}\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|_{R,\boldsymbol{w}_t} &= \mathbb{E}\sqrt{(\sum_{\tau\in m_t}\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R(\boldsymbol{w}_t)(\sum_{\tau\in m_t}\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)} \\
&= \mathbb{E}\sqrt{\sum_{\tau\in m_t}\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R,\boldsymbol{w}_t}+\sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R(\boldsymbol{w}_t)(\boldsymbol{\ell}_{\tau'}-\widehat{\boldsymbol{\ell}}_{\tau'})} \\
&\leq \sqrt{\sum_{\tau\in m_t}\mathbb{E}\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R,\boldsymbol{w}_t}}+\mathbb{E}\sqrt{\sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R(\boldsymbol{w}_t)(\boldsymbol{\ell}_{\tau'}-\widehat{\boldsymbol{\ell}}_{\tau'})},
\end{aligned}$$

where the inequality is by $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and Jensen. For the first term, by Lemma 10,

$$\begin{aligned}
\sum_{\tau\in m_t}\mathbb{E}\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R,\boldsymbol{w}_t} &\leq 4\sum_{\tau\in m_t}\mathbb{E}\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R,\boldsymbol{w}_\tau} \leq 4\sum_{\tau\in m_t}\mathbb{E}\|\boldsymbol{\ell}_\tau\|^2_{R,\boldsymbol{w}_\tau}+4\sum_{\tau\in m_t}\mathbb{E}\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R,\boldsymbol{w}_\tau} \\
&\leq 4|m_t|(\alpha^2+\beta^2) \leq 8\eta HSA|m_t|.
\end{aligned}$$

In Lemma 13 we bound the second term similarly to BIAS_1 by $4\sqrt{\frac{H^2S}{T}}$ due to the estimator's small bias. Overall,

$$\begin{aligned}
\sqrt{\eta H}\cdot\text{DRIFT} &\leq 8\eta H\sum_{t=1}^T\sqrt{SA|m_t|}+4\sqrt{\eta H^3ST} \leq 8\eta H\sum_{t=1}^T\sqrt{S^2A^2+|m_t|^2}+4\sqrt{\eta H^3ST} \\
&\leq 8\eta HSAT+8\eta H\sum_{t=1}^T|m_t|+4\sqrt{\eta H^3ST} \leq 8\eta HSAT+8\eta HD+4\sqrt{\eta H^3ST},
\end{aligned}$$

where we used $ab\leq a^2+b^2$ and $\sum_{t=1}^T|m_t|=D$. Finally, we sum all the different terms. ■

Lemma 6 *Let $\mathcal{W}=\Delta(\mathcal{P})\cap\Omega$. It holds that $\mathcal{W}$ is non-empty and,*

1. *For any $\boldsymbol{w}\in\Delta(\mathcal{M})$, there exists $\widetilde{\boldsymbol{w}}\in\mathcal{W}$ such that $\|\boldsymbol{w}-\widetilde{\boldsymbol{w}}\|_1\leq\frac{2H}{T}$.*
2. *Given $\boldsymbol{w}\in\mathcal{W}$, let π be defined by $\pi_h(a\mid s)=\frac{\boldsymbol{w}_h(s,a)}{\boldsymbol{w}_h(s)}$ and $\boldsymbol{u}_h(s,a)=\max_{\widehat{p}\in\mathcal{P}}\boldsymbol{w}_h^{\pi,\widehat{p}}(s,a)$. Then, $\|\boldsymbol{w}^\pi-\boldsymbol{w}\|_1\leq\frac{2H}{T}$ and $\|\boldsymbol{u}-\boldsymbol{w}\|_1\leq\frac{4H^2S}{T}$.*

Proof

1. Define $\widetilde{p}=\{\widetilde{p}_h:\mathscr{S}\times\mathscr{A}\to\Delta_{\mathscr{S}}\}_{h=1}^H$ by $\widetilde{p}_h(s'\mid s,a)=(1-\frac{1}{THSA})p_h(s'\mid s,a)+\frac{1}{THS^2A}$ and notice that $\widetilde{p}\in\mathcal{P}$ since $|p_h(s'\mid s,a)-\widetilde{p}_h(s'\mid s,a)|\leq\frac{1}{THSA}$. Next, let π_u be the uniformly random policy, and define $\widetilde{\boldsymbol{w}}=(1-\frac{1}{T})\boldsymbol{w}+\frac{1}{T}\boldsymbol{w}^{\pi_u,\widetilde{p}}$. It holds that $\widetilde{\boldsymbol{w}}\in\Delta(\mathcal{P})$

because $\Delta(\mathcal{P})$ is a convex set. Moreover, notice that $\boldsymbol{w}_h^{\pi_u,\widetilde{p}}(s,a,s') \geq \frac{1}{(THS^2A)^2A}$ which implies that $\widetilde{\boldsymbol{w}}_h(s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$. Thus, $\widetilde{\boldsymbol{w}} \in \mathcal{W}$. Finally,

$$\begin{aligned}
\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_1 &= \sum_{h,s,a,s'} \left|\boldsymbol{w}_h(s,a,s') - \widetilde{\boldsymbol{w}}_h(s,a,s')\right| \\
&= \sum_{h,s,a,s'} \left|\frac{1}{T}\boldsymbol{w}_h(s,a,s') - \frac{1}{T}\boldsymbol{w}_h^{\pi_u,\widetilde{p}}(s,a,s')\right| \\
&\leq \frac{1}{T}\sum_{h,s,a,s'} \boldsymbol{w}_h(s,a,s') + \frac{1}{T}\sum_{h,s,a,s'} \boldsymbol{w}_h^{\pi_u,\widetilde{p}}(s,a,s') = \frac{2H}{T}.
\end{aligned}$$

2. Define loss function $\widetilde{\boldsymbol{\ell}}_h(s,a) = \operatorname{sign}(\boldsymbol{w}_h^{\pi}(s,a) - \boldsymbol{w}_{t,h}(s,a))$ and note that $\|\boldsymbol{w}^{\pi} - \boldsymbol{w}\|_1 = V_1^{\pi,p,\widetilde{\boldsymbol{\ell}}}(s_{\text{init}}) - V_1^{\pi,\widehat{p},\widetilde{\boldsymbol{\ell}}}(s_{\text{init}})$ for some $\widehat{p} \in \mathcal{P}$. Combining Lemma 15[1] and the fact that $\|p - \widehat{p}\|_\infty \leq \frac{1}{THSA}$ proves that $\|\boldsymbol{w}^{\pi} - \boldsymbol{w}\|_1 \leq \frac{2H}{T}$. Now, let $\widehat{p}^{h,s}$ be the transition function that corresponds to $\boldsymbol{u}_h(s)$. We have that, $\|\widehat{p}^{h,s} - \widehat{p}\|_\infty \leq \|\widehat{p}^{h,s} - p\|_\infty + \|p - \widehat{p}\|_\infty \leq \frac{2}{THSA}$. Thus, using the same argument as the above,

$$\|\boldsymbol{u} - \boldsymbol{w}\|_1 \leq \sum_{h,s} \|\boldsymbol{w}^{\pi,\widehat{p}^{h,s}} - \boldsymbol{w}\|_1 \leq \frac{4H^2S}{T}.$$

■

Lemma 12 (BIAS$_1$) *When running Delayed FTRL for adversarial MDPs we have,*

$$\text{BIAS}_1 = \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)^\top(\boldsymbol{\ell}_t - \widehat{\boldsymbol{\ell}}_t)] \leq 8H^2S.$$

Proof Let $\mathcal{G}_t$ be the history of all episodes in $[t-1]$, and note that $\boldsymbol{w}_t, \boldsymbol{u}_t$ and $\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)$ are all determined by $\mathcal{G}_t$. Therefore,

$$\begin{aligned}
\text{BIAS}_1 &= \mathbb{E}\left[\sum_{t,h,s,a} \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)_h(s,a)(\boldsymbol{\ell}_{t,h}(s,a) - \mathbb{E}[\widehat{\boldsymbol{\ell}}_{t,h}(s,a) \mid \mathcal{G}_t])\right] \\
&= \mathbb{E}\left[\sum_{t,h,s,a} \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)_h(s,a)\boldsymbol{\ell}_{t,h}(s,a)\left(1 - \frac{\boldsymbol{w}_h^{\pi_t}(s,a)}{\boldsymbol{u}_{t,h}(s,a)}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t,h,s,a} \boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)_h(s,a)\frac{|\boldsymbol{u}_{t,h}(s,a) - \boldsymbol{w}_h^{\pi_t}(s,a)|}{\boldsymbol{u}_{t,h}(s,a)}\right]
\end{aligned}$$

1. We note that Even-Dar et al. (2009) (see also Shani et al. (2020)) apply the value difference lemma (Lemma 15) on positive losses (or rewards), where here we use loss function supported in $[-1,1]$. However, the proof of the lemma in fact holds for any loss functions $\widetilde{\boldsymbol{\ell}}, \boldsymbol{\ell} \subseteq \mathbb{R}^{HSA}$.

Now, as in the proof of Lemma 2, $\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m) \in \mathcal{D}_R(\boldsymbol{w}_t, \frac{1}{2})$. Thus, by Lemma 16, $\boldsymbol{w}(\widehat{\boldsymbol{L}}_t^m)_h(s,a) \leq 2\boldsymbol{w}_{t,h}(s,a) \leq 2\boldsymbol{u}_{t,h}(s,a)$. Therefore,

$$\textsc{Bias}_1 \leq 2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{u}_t - \boldsymbol{w}^{\pi_t}\|_1] \leq 8H^2S$$

where the last is by article 2 in Lemma 6. ■

Lemma 13 *When running Delayed FTRL for adversarial MDPs, for any t,*

$$\mathbb{E}\sqrt{\sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R(\boldsymbol{w}_t)(\boldsymbol{\ell}_{\tau'} - \widehat{\boldsymbol{\ell}}_{\tau'})} \leq 4\sqrt{\frac{H^2S}{T}}.$$

Proof We first apply the law of total expectation and Jensen inequality,

$$\begin{aligned}
&\mathbb{E}\sqrt{\sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R(\boldsymbol{w}_t)(\boldsymbol{\ell}_{\tau'} - \widehat{\boldsymbol{\ell}}_{\tau'})} \\
&\leq \mathbb{E}\sqrt{\eta\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_{t,h}(s,a)(\boldsymbol{\ell}_{\tau,h}(s,a) - \mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau,h}(s,a)\mid\mathcal{G}_t])(\boldsymbol{\ell}_{\tau',h}(s,a) - \mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau',h}(s,a)\mid\mathcal{G}_t])}
\end{aligned}$$

Now, let $\mathcal{G}_t$ be the history of all episodes in $[t-1]$, and note that $\boldsymbol{\ell}_{\tau',h}(s,a) - \mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau',h}(s,a)\mid\mathcal{G}_t] \in [0,1]$. Thus, we can further bound the above by,

$$\begin{aligned}
&\mathbb{E}\sqrt{\eta\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_{t,h}(s,a)(\boldsymbol{\ell}_{\tau,h}(s,a) - \mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau,h}(s,a)\mid\mathcal{G}_t])} \\
&\leq \mathbb{E}\sqrt{\eta|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{w}_{t,h}(s,a)\boldsymbol{\ell}_{\tau,h}(s,a)\frac{\boldsymbol{u}_{\tau,h}(s,a) - \boldsymbol{w}_{\tau,h}(s,a)}{\boldsymbol{u}_{\tau,h}(s,a)}} \\
&\leq 2\mathbb{E}\sqrt{\eta|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{w}_{\tau,h}(s,a)\frac{\boldsymbol{u}_{\tau,h}(s,a) - \boldsymbol{w}_{\tau,h}(s,a)}{\boldsymbol{u}_{\tau,h}(s,a)}} \\
&\leq 2\mathbb{E}\sqrt{\eta|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{u}_{\tau,h}(s,a) - \boldsymbol{w}_{\tau,h}(s,a)} \\
&= 2\mathbb{E}\sqrt{\eta|m_t|\sum_{\tau\in m_t}\|\boldsymbol{u}_\tau - \boldsymbol{w}_\tau\|_1} \\
&\leq 4\sqrt{\eta|m_t|^2\frac{H^2S}{T}} \leq 4\sqrt{\frac{H^2S}{T}},
\end{aligned}$$

where the third inequality is by Lemma 16, the forth is since $\boldsymbol{w}_{\tau,h}(s,a) \leq \boldsymbol{u}_{\tau,h}(s,a)$, the fifth inequality is by Lemma 6, and the last is since $\eta \leq \frac{1}{d_{max}^2} \leq \frac{1}{|m_t|^2}$. ■

Appendix D. Deferred Proofs for Linear Bandits (Section 6)

Theorem 8 *Let $\boldsymbol{u} \in \mathcal{W}$. Running Delayed FTRL for linear bandits (Algorithm 3) with $\gamma = \min\left\{\frac{1}{(64BK(1+d_{\max}))^2}, \sqrt{\frac{\nu \ln\left(1+\sqrt{T}\right)}{16(BK)^2T}}\right\}$ and $\eta = \min\left\{\frac{1}{(16d_{\max})^2}, \sqrt{\frac{B^2}{16D}}\right\}$ guarantees,*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq 14BK\sqrt{\nu T \ln(T)} + 8B\sqrt{D} + 2^{14}\nu B^2K^2(1+d_{\max})^2\ln(T).$$

Proof We start by verifying the assumptions of Corollary 3. Using $\mathbb{E}[\boldsymbol{v}_t] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_t^\top] = \frac{1}{K}\boldsymbol{I}$ we see that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t] = \boldsymbol{\ell}_t$. Observe that the distribution of $\widehat{\boldsymbol{\ell}}_{\tau'}$ is fully determined given $\mathcal{F}_t$ because $\mathcal{F}_{\tau'} \subseteq \mathcal{F}_t$. Furthermore, since $\widehat{\boldsymbol{\ell}}_\tau$ can not be used in round τ' because τ is not available in round t due to the delay, we must have that $\widehat{\boldsymbol{\ell}}_{\tau'}$ is independent of $\widehat{\boldsymbol{\ell}}_\tau$. Thus, by the tower rule

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top\left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,|\, \mathcal{F}_t\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{\ell}}_\tau}\left[\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top\left(\nabla^2 R(\boldsymbol{w}_t)\right)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,|\, \mathcal{F}_t, \widehat{\boldsymbol{\ell}}_\tau\right]\right] = 0\,,
\end{aligned}$$

where we used that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}|\mathcal{F}_t] = \mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t, \widehat{\boldsymbol{\ell}}_\tau] = \boldsymbol{\ell}_t$. Next, observe that because $\nabla^2 R(\boldsymbol{w}) \succeq \frac{1}{\gamma}\nabla^2\Psi(\boldsymbol{w})$ we have that

$$\|\widehat{\boldsymbol{\ell}}_t\|^2_{R,\boldsymbol{w}_t} \leq \gamma K^2(\boldsymbol{\ell}_t^\top \boldsymbol{a}_t)^2 \boldsymbol{v}_t^\top \boldsymbol{v}_t = \gamma K^2(\boldsymbol{\ell}_t^\top \boldsymbol{a}_t)^2\,.$$

Since $\|\boldsymbol{\ell}_t\|_2 \leq 1$ and $\mathcal{W} \subseteq \mathcal{B}(B)$, we have that $(\boldsymbol{\ell}_t^\top \boldsymbol{a}_t)^2 \leq B^2$ and thus

$$\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t} \leq \underbrace{\sqrt{\gamma(BK)^2}}_{\beta} \leq \frac{1}{64(1+d_{\max})}\,,$$

where the last inequality is because $\gamma \leq (64BK(1+d_{\max}))^{-2}$. Because $\nabla^2 R(\boldsymbol{w}) \succeq \frac{1}{\eta}\boldsymbol{I}$ and $\|\boldsymbol{\ell}_t\|_2 \leq 1$ we have that

$$\|\boldsymbol{\ell}_t\|_{R,\boldsymbol{w}} \leq \underbrace{\sqrt{\eta}}_{\alpha} \leq \frac{1}{16d_{\max}}\,,$$

where the last inequality is because $\eta \leq (16d_{\max})^{-2}$. The final assumption to verify is $4\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 R(\boldsymbol{w}') \succeq \frac{1}{4}\nabla^2 R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$, which is immediate due to equation (5).

Pick $\widetilde{\boldsymbol{u}} \in \mathcal{W}$ and $\delta > 0$. Set $\boldsymbol{u} = \frac{\widetilde{\boldsymbol{u}} - \boldsymbol{w}_1}{1+\delta} + \boldsymbol{w}_1 \in \mathcal{W}_\delta$. Using equation (6) and $\mathcal{W} \subseteq \mathcal{B}(B)$ we have that

$$R(\boldsymbol{u}) - R(\boldsymbol{w}_1) \leq \frac{B^2}{\eta} + \frac{\nu}{\gamma}\ln\left(\frac{1+\delta}{\delta}\right)\,. \tag{19}$$

Furthermore, by using that $\mathcal{W} \in \mathcal{B}(B)$ and $\|\boldsymbol{\ell}_t\|_2 \leq 1$, we have that

$$\sum_{t=1}^{T}(\widetilde{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\ell}_t = \sum_{t=1}^{T}\left(1 - \frac{1}{1+\delta}\right)(\widetilde{\boldsymbol{u}} - \boldsymbol{w}_1)^\top \boldsymbol{\ell}_t \leq 2TB\left(\frac{\delta}{1+\delta}\right)\,.$$

Thus, by setting $\delta = \frac{1}{\sqrt{T}}$ and then applying Corollary 3 with $\beta^2 = \gamma(BK)^2$ we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \widetilde{\boldsymbol{u}})^\top \boldsymbol{\ell}_t\right] &\le \mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + 2B\sqrt{T} \\
&\le R(\boldsymbol{u}) - R(\boldsymbol{w}_1) + 16\gamma(BK)^2 + 16\eta\sum_{t=1}^{T}|m_t| + 2B\sqrt{T} \\
&\le \frac{B^2}{\eta} + \frac{\nu}{\gamma}\ln\left(1+\sqrt{T}\right) + 16\gamma(BK)^2 T + 16\eta D + 2B\sqrt{T} \qquad \text{(using (19))} \\
&\le 8B\sqrt{D} + 256 d_{\max}^2 \\
&\quad + 8BK\sqrt{\nu T \ln\left(1+\sqrt{T}\right)} + \left(64BK(1+d_{\max})\right)^2 \nu \ln\left(1+\sqrt{T}\right) + 2B\sqrt{T}\,,
\end{aligned}$$

where in the last step we used our choices for η and γ. ■

Algorithm 4: Doubling procedure

Input: T, D and algorithm ALG (for known T, D and d_{max}).
Set epoch index $e = 1$ and initialize ALG with T, D and 2^e as d_{max}.
for $t = 1, ..., T$ **do**
  if $\max_{j \in o_t} d_j \geq 2^e$ **then**
    Start a new epoch $e = e + 1$, and re-initiate ALG with T, D and 2^e as d_{max}.
  end if
  Play according to ALG.
end for

Appendix E. Doubling with Delayed Feedback

In this section we show how to handle unknown problem parameters. For simplicity of presentation we assume that only d_{max} is unknown. The case of unknown T and D can be done in a similar fashion (e.g., see Bistritz et al. (2019); Lancewicki et al. (2022)).

Theorem 14 *Let ALG be an algorithm for known T, D and d_{max} and assume that ALG guarantees regret of $R_{T,D}(d_{max})$ whenever initiated properly. Then, running Algorithm 4 with unknown d_{max} guarantees regret,*

$$\mathcal{R}_T \leq 2R_{T,D}(2d_{max}) \log T + 2Md_{max} \log T,$$

where $M = \max_{t \in [T], \boldsymbol{a}, \widetilde{\boldsymbol{a}} \in \mathcal{A}} (\boldsymbol{a} - \widetilde{\boldsymbol{a}})^\top \boldsymbol{\ell}_t$ is the maximal regret per round (e.g., in Section 5, $M \leq H$).

Proof Let $\mathcal{T}_e = \{t : 2^{e-1} \leq \max_{j \in o_t} d_j \leq 2^e\}$ be the set of indices of epoch e, and let $\widetilde{\mathcal{T}}_e = \{t \in \mathcal{T}_e : d_t \leq 2^e\}$ be the indices of epoch e with delay $\leq 2^e$. The regret in rounds $t \in \widetilde{\mathcal{T}}_e$ is at most $R_{T,D}(2^e) \leq R_{T,D}(2d_{max})$ since the maximal delay in these rounds is indeed bounded by 2^e. In addition, the regret in $\mathcal{T}_e \backslash \widetilde{\mathcal{T}}_e$ is at most Md_{max} since $|\mathcal{T}_e \backslash \widetilde{\mathcal{T}}_e| \leq d_{max}$. Thus, the total regret in epoch e is at most,

$$\underbrace{R_{T,D}(2d_{max})}_{\text{Regret in } \widetilde{\mathcal{T}}_e} + \underbrace{Md_{max}}_{\text{Regret in } \mathcal{T}_e \backslash \widetilde{\mathcal{T}}_e} .$$

Finally, the total number of epochs is at most $\log d_{max} + 1 \leq 2 \log T$ and thus, the total regret is bounded by,

$$\mathcal{R}_T \leq 2R_{T,D}(2d_{max}) \log T + 2Md_{max} \log T.$$

■

Appendix F. Auxiliary Lemmas

Lemma 15 (Value Difference Lemma Even-Dar et al. (2009)) *For any two triplets $(\pi, p, \boldsymbol{\ell})$ and $(\widetilde{\pi}, \widetilde{p}, \widetilde{\boldsymbol{\ell}})$ of policy, transition and cost function,*

$$\begin{aligned} V_1^{\pi,p,\boldsymbol{\ell}}(s_{init}) - V_1^{\widetilde{\pi},\widetilde{p},\widetilde{\boldsymbol{\ell}}}(s_{init}) = &\sum_{h=1}^{H} \mathbb{E}_{s\sim \boldsymbol{w}_h^{\widetilde{\pi},\widetilde{p}}} \Big[\Big\langle \pi_h(\cdot \mid s) - \widetilde{\pi}_h(\cdot \mid s), Q_h^{\pi,p,\boldsymbol{\ell}}(s,\cdot) \Big\rangle \Big] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s,a\sim \boldsymbol{w}_h^{\widetilde{\pi},\widetilde{p}}} \Big[\boldsymbol{\ell}(s,a) - \widetilde{\boldsymbol{\ell}}(s,a) \Big] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s,a\sim \boldsymbol{w}_h^{\widetilde{\pi},\widetilde{p}}} \Big[\Big\langle p_h(\cdot \mid s) - \widetilde{p}_h(\cdot \mid s), V_{h+1}^{\pi,p,\boldsymbol{\ell}} \Big\rangle \Big] \end{aligned}$$

Lemma 16 *Let $\mathcal{W} \subseteq \{\boldsymbol{w} \in \mathbb{R}^n : \forall i \in [n],\ \boldsymbol{w}(i) > 0\}$. Let $R : \mathcal{W} \to \mathbb{R}$ be some twice-differentiable convex function, and let $\phi(\boldsymbol{w}) = -\frac{1}{\gamma}\sum_{i=1}^{n} \log \boldsymbol{w}(i)$ be the log barrier with $\gamma \in (0,1)$. Assume that for any $\boldsymbol{w} \in \mathcal{W}$, $\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 \phi(\boldsymbol{w})$. Then for any $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$,*

$$\forall i \in [n], \qquad \frac{1}{2} w(i) \le w'(i) \le 2w(i).$$

Proof Since $\nabla^2 R(\boldsymbol{w}) \succeq \nabla^2 \phi(\boldsymbol{w})$, for any $\boldsymbol{w}' \in \mathcal{D}_R(\boldsymbol{w}, \frac{1}{2})$,

$$(\|\boldsymbol{w}' - \boldsymbol{w}\|_{\phi,\boldsymbol{w}}^*)^2 \le (\|\boldsymbol{w}' - \boldsymbol{w}\|_{R,\boldsymbol{w}}^*)^2 \le \frac{1}{4}.$$

On the other hand,

$$(\|w' - w\|_{\phi,\boldsymbol{w}}^*)^2 = \sum_{j=1}^{n} \frac{(\boldsymbol{w}'(j) - \boldsymbol{w}(j))^2}{\gamma \boldsymbol{w}(j)^2} \ge \frac{(\boldsymbol{w}'(i) - \boldsymbol{w}(i))^2}{\gamma \boldsymbol{w}(i)^2} \ge \frac{(\boldsymbol{w}'(i) - \boldsymbol{w}(i))^2}{\boldsymbol{w}(i)^2}.$$

Thus, $|\boldsymbol{w}'(i) - \boldsymbol{w}(i)| \le \frac{1}{2}\boldsymbol{w}(i)$ which implies that $\frac{1}{2}\boldsymbol{w}(i) \le \boldsymbol{w}'(i) \le 2\boldsymbol{w}(i)$. ■