POLITECNICO DI TORINO
Repository ISTITUZIONALE

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

Original
A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs / Zierahn, L., Van Der Hoeven, D., Lancewicki, T., Rosenberg, A., Cesa-Bianchi, N.. - In: JOURNAL OF MACHINE LEARNING RESEARCH. - ISSN 1532-4435. - 26:(2025), pp. 1-60.

Availability:
This version is available at: 11583/2998193 since: 2025-03-10T08:47:52Z

Publisher:
JMLR

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

Journal of Machine Learning Research 26 (2025) 1-60 Submitted 4/24; Revised 9/24; Published 3/25

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

Lukas Zierahn LUKAS.ZIERAHN@GMAIL.COM
Centrum Wiskunde & Informatica and Booking.com, The Netherlands

Dirk van der Hoeven DIRK@DIRKVANDERHOEVEN.COM
Mathemathical Institute, Leiden University, The Netherlands

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

Aviv Rosenberg AVIVROS@GOOGLE.COM
Google Research

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

Editor: Ambuj Tewari

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 four important settings. We derive the first optimal (up to logarithmic factors) regret bounds for combinatorial semi-bandits with delay and adversarial Markov Decision Processes with delay (both known and unknown transition functions). Furthermore, we use our analysis to develop an efficient algorithm for linear bandits with delay achieving near-optimal regret bounds. In order to derive these results we show that FTRL remains stable across multiple rounds under mild assumptions on the regularizer.

Keywords: Online learning, bandit feedback, delayed feedback, Markov Decision Processes, combinatorial semi-bandits

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

©2025 Lukas Zierahn, Dirk van der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/24-0496.html.

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 bandit feedback setting; 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), linear bandits, and adversarial Markov Decision Processes (MDPs).

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 us to separately control 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 provide the first optimal regret bounds for adversarial MDPs with delays and known transitions. For adversarial MDPs with delays and unknown transitions we provide the state-of-the-art results. 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 $\boldsymbol{a}_t \in \mathcal{A} \subseteq \mathbb{R}^K$, where $\mathcal{A}$ is an action set of dimension K. The learner subsequently 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 $d_1, \ldots, d_T$ 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 $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$. 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.

1.1 Contributions

This work and the work it extends (Van der Hoeven et al., 2023) have the following main 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) provide 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.

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 $\widetilde{O}(K^{3/2}\sqrt{T}+K^2\sqrt{D})$.

Adversarial Markov Decision Processes. Delayed feedback in adversarial (finite-horizon and episodic) MDPs was first studied by Lancewicki et al. (2022a). Under full-information feedback, where the agent observes the entire cost function at the end of the episode, they achieve the optimal regret bound: $\widetilde{O}(H\sqrt{T+D})$, where T is the number of episodes and H is the horizon. However, with bandit feedback (where the only observed costs are those along the agent's trajectory), their regret bound is of order $T^{2/3}+D^{2/3}$. The current state-of-the-art guarantees under delayed bandit feedback are by Jin et al. (2022) and Lancewicki et al. (2023) who achieve a regret bound of $\widetilde{O}(H\sqrt{SAT}+H(HSA)^{1/4}\sqrt{D})$ and $\widetilde{O}(H^2\sqrt{SAT}+H^3\sqrt{D})$ in the known transition setting, and a regret bound of $\widetilde{O}(H^2S\sqrt{AT}+H(HSA)^{1/4}\sqrt{D})$ and $\widetilde{O}(H^3S\sqrt{AT}+H^3\sqrt{D})$ in the unknown transition setting, respectively. Here, S is the number of states in the MDP and A the number of actions. However, there is still a gap compared to the lower bound of Lancewicki et al. (2022a). 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. Moreover, our bound of $\widetilde{O}(H^2S\sqrt{AT}+H\sqrt{D})$ for unknown transitions, achieves the first optimal regret in the delay term and matches the best known regret bound (even for the standard non-delayed setting) in the other term.

1.2 Additional related work

Combinatorial semi-bandits with delayed feedback. Stochastic combinatorial semi-bandits have first been introduced by Gai et al. (2012) but featured an undesirable dependency on the reciprocal of the square of the smallest gap between arms, which was improved by Chen et al. (2013) by removing the square. The first matching upper and lower bounds are due to Kveton et al. (2015) by using an upper confidence bound (UCB) based approach, though bounds using Thompson sampling are also known Wen et al. (2015). A special case of the stochastic combinatorial semi-bandit setting with delayed feedback, namely stochastic multi-armed bandits with delayed feedback, has been studied in many different variations (Dudik et al., 2011; Agarwal and Duchi, 2012; Pike-Burke et al., 2018; Zhou et al., 2019; Gael et al., 2020; Lancewicki et al., 2021; Cohen et al., 2021).

In the nonstochastic combinatorial semi-bandit setting there have been several results. Adversarial online path-finding problems, a special case of semi-bandits, has been studied by György et al. (2007) achieving an sub-optimal upper bound, an optimal upper bound for m-sets is due to Kale et al. (2010) and Uchiya et al. (2010). The optimal bound for semi-bandits in general, which we recover in the non-delayed setting, is due to Audibert et al. (2014). 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 that of Jin et al. (2022) mentioned earlier. Finally, a different line of work (Katsikopoulos and Engelbrecht, 2003; Walsh et al., 2009) considers 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. (2022a). A stochastic version of MDPs with delayed feedback has been studied by (Howson et al., 2023a).

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). Dani et al. (2007); 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. Stochastic linear bandits with delayed feedback has been studied by (Vernade et al., 2020; Howson et al., 2023b).

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 focus on Follow The Regularized Leader (FTRL) and define the FTRL prediction given a sum of losses $\boldsymbol{L}$ (or estimated losses $\widehat{\boldsymbol{L}}$) as follows,

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

where $\mathcal{W} \subseteq \mathbb{R}^K$ is a compact closed convex set, R_t is a twice-differentiable strongly convex function. Note that the domain $\mathcal{W}$ and the action set $\mathcal{A}$ do not necessarily coincide, as is the case of

combinatorial semi-bandits for example, where $\mathcal{W}$ is the convex hull of $\mathcal{A}$, i.e. $\mathcal{W} = \mathrm{Conv}(\mathcal{A})$. Similarly, $\boldsymbol{a}_t$ and $\boldsymbol{w}_t(\widehat{\boldsymbol{L}})$ do not necessarily coincide. We will specify the relationship between $\boldsymbol{a}_t$ and $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$ in each application.

If we define $\widetilde{o}_t = \{\tau : \tau + d_\tau < t\}$ as the set of all losses available at the beginning of round t, then FTRL predicts $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$, where $\widehat{\boldsymbol{L}}_t = \sum_{\tau \in \widetilde{o}_t} \widehat{\boldsymbol{\ell}}_\tau$. We then use the notation $[N] = \{1, \dots, N\}$ and define $\widetilde{m}_t = [t-1] \setminus \widetilde{o}_t$ to be the set of indices of losses that have not been observed at the start of round t 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. This assumption is without loss of generality, as we may employ the standard doubling trick to overcome the need to know this parameter (Bistritz et al., 2019; Lancewicki et al., 2022a), see also Appendix E.

Additional notations. We denote by $\boldsymbol{w}_t(\boldsymbol{L}, i)$ the i-th element of the vector $\boldsymbol{w}_t(\boldsymbol{L})$. We define a filtration of all random events observed by the learner up to round t as $\mathcal{F}_t = \left\{\left(\tau, \boldsymbol{a}_\tau, \mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau)\right) : \tau + d_\tau < t\right\}$ and we use $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. For a twice-differentiable function ϕ such that $\nabla^2\phi(\boldsymbol{v}) \succ 0\boldsymbol{I}$ for all $\boldsymbol{v} \in \mathcal{W}$ we denote by $\|\boldsymbol{L}\|_{\phi,\boldsymbol{v}} = \sqrt{\boldsymbol{L}^\top \left(\nabla^2\phi(\boldsymbol{v})\right)^{-1} \boldsymbol{L}}$ and by $\|\boldsymbol{L}\|^\star_{\phi,\boldsymbol{v}} = \sqrt{\boldsymbol{L}^\top \nabla^2\phi(\boldsymbol{v}) \boldsymbol{L}}$. The Dikin ellipsoid with radius r around $\boldsymbol{v}$ induced by ϕ is defined as $\mathcal{D}_\phi(\boldsymbol{v}, r) = \{\boldsymbol{x} \in \mathcal{W} : \|\boldsymbol{x} - \boldsymbol{v}\|^\star_{\phi,\boldsymbol{v}} \leq r\}$. The notation $\widetilde{O}(\cdot)$ hides poly-logarithmic factors, whereas $\lesssim$ denotes inequalities that hide constant factors.

Changing domains. Some settings require changing domains. In the MDP setting with unknown transitions, the domain is related to the estimate of the transition function and as we update and become more confident in our estimates, we may wish to shrink the domain. We overload the notation slightly and define

$$\boldsymbol{w}_t(\boldsymbol{L}) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{W}_t} \boldsymbol{L}^\top \boldsymbol{v} + R_t(\boldsymbol{v}), \tag{1}$$

where we require all $\mathcal{W}_t$ to be compact closed convex sets. Our analysis requires that if we observe the feedback from round τ in timestep t, then the corresponding iterate of FTRL, $\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)$, must be in the same Dikin ellipsoid as the current iterate $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$. To ensure that condition holds the domains of timestep τ and timestep t have to agree. If that is not the case, we have to skip round τ, which means trivially bounding the regret of round τ with an appropriate constant value (like the length of the episode in the MDP setting) and not building a loss estimator using the information of round τ. We define $\Lambda \subseteq [T]$ to be the set of rounds that we skip and $\overline{\Lambda} = [T] \setminus \Lambda$ be the rounds that we do not skip. Since we chose not to use the loss estimators of skipped rounds, we intersect the set of observed losses and the set of missing losses at time t with the rounds that we did not skip: $o_t = \widetilde{o}_t \cap \overline{\Lambda}$, $m_t = \widetilde{m}_t \cap \overline{\Lambda}$. When we observe the loss of round τ, we know if we have changed the domain since τ and thus o_t is well defined and non-random given the history $\mathcal{F}_t$. The same is not true for m_t, which can depend on future rounds. This is not a problem for the algorithms considered here, as m_t is a quantity only used in the analysis and for tuning the learning rates, where $|\widetilde{m}_t|$ can be used as an upper bound for $|m_t|$. The constraints that must be fulfilled to use changing domains are formalised in Assumption 1.

Assumption 1 *For all $t \in [T]$ we assume that o_t is non-random given the history $\mathcal{F}_t$ and that $\mathcal{W}_t = \mathcal{W}_\tau$ for all $\tau \in o_t \setminus o_{t-1}$. We also assume that $\mathcal{W}_t \subseteq \mathcal{W}_{t-1}$ is a compact convex set such that $\mathcal{W}_T$ is non-empty.*

If the domain is constant and no rounds are skipped then Assumption 1 reduces to the standard assumption that $\mathcal{W}$ is compact, convex, and non-empty as in that case $o_t = \widetilde{o}_t$ and $m_t = \widetilde{m}_t$.

In the remainder of the paper we use the following notation for cumulative loss estimates:

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

Note that $\widehat{\boldsymbol{L}}_t^\star = \widehat{\boldsymbol{L}}_t^m + \widehat{\boldsymbol{\ell}}_t$ and that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)$ is equivalent to FTRL in the non-delayed setting. We also make the following regularity assumptions on the regularizer R_t.

Assumption 2 *Let R_t be the regularizer associated with equation* (1) *and let $\kappa > 0$. Suppose that for all $t \in [T]$*

(a) $4\nabla^2 R_t(\boldsymbol{v}) \succeq \nabla^2 R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{v})$ *for all $\boldsymbol{v} \in \mathcal{W}_t$ and $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$.*

(b) $\kappa \left(\nabla R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)) - \nabla R_{t+\delta}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\right)^\top \boldsymbol{x} \le \frac{\sqrt{\kappa}}{32}\sqrt{\boldsymbol{x}^\top \nabla^2 R_{t+\delta}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\boldsymbol{x}}$ *for all $\boldsymbol{x} \in \mathbb{R}^d$ and all $\delta \in [d_{\max}]$.*

(c) $R_t(\boldsymbol{v}) \le R_{t'}(\boldsymbol{v})$ *and* $\nabla^2 R_t(\boldsymbol{v}) \preceq \nabla^2 R_{t'}(\boldsymbol{v})$ *for all $\boldsymbol{v} \in \mathcal{W}_t$ all $t \le t'$.*

Assumption 2(a) allows us to relate the Hessian of the regularizer at different iterates of FTRL, which is crucial in our analysis. Since essentially all regularizers we use in this paper are approximately self-concordant, assumption 2(a) is almost automatically satisfied (Nemirovski, 2004), see also equation 18. Assumption (b) tells us that the regularizer should not change too much between rounds and is used show that the different iterates of FTRL are close to each other. As we will see, assumption (b) can be verified for most standard regularizers given that the learning rate does not change too much between rounds. Assumption 2(c) is a technical assumption and is satisfied by almost all standard regularizers, including those that we use in this paper.

3. Analysis

In this section we establish general results that are then applied to combinatorial bandits, MDPs, and linear bandits in the next sections. First, we give a broad overview of the proof ideas and then prove the statements rigorously.

3.1 Overview

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 change between different instances of FTRL.

Suppose for a moment that the domain $\mathcal{W}_t = \mathcal{W}$ is constant, we are not skipping any rounds $\Lambda = \emptyset$, and 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. Let

$\boldsymbol{u} \in \mathbb{R}^K$ be any comparator. Our analysis relies on the following decomposition of the regret

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

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

If $\widehat{\boldsymbol{\ell}}_t$ is an unbiased estimator of the loss then $\boldsymbol{b}_t = \boldsymbol{0}$, which implies that the bias term of the decomposition is also 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 18 in Appendix F), see also (Joulani et al., 2020, Theorem 3). Now we focus on the second line of Equation (2). Typically, H_1 and H_2 are analysed simultaneously and referred to as "drift", for example, see (Gyorgy and Joulani, 2021). We split the drift into H_1 and H_2 because we want to analyze the cost of delay and the cost of bandit feedback separately.

H_1 can be interpreted as capturing the influence of the missing observations. H_2 captures the influence of knowing the loss estimated one step in advance against running a non-delayed version of FTRL. To bound H_1 and H_2 we will use the same tools. First we need to relate the differences between the predictions of the different FTRL instances to the losses used in computing the different FTRL iterates. Lemma 5 states that if $\boldsymbol{w}_t(\boldsymbol{L}'), \boldsymbol{w}_t(\boldsymbol{L}) \in \mathcal{D}_R(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$ for some $\kappa > 0$, some $\boldsymbol{v} \in \mathcal{W}$, and some $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^K$, and the regularizer is sufficiently nice, then $\|\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}} \leq 8\|\boldsymbol{L}' - \boldsymbol{L}\|_{R_t,\boldsymbol{v}}$. In order to apply this result for different $\boldsymbol{w}_t(\cdot)$, we require them to lie in the same Dikin ellipsoid, and Lemma 6 establishes machinery to allow us to determine when that is the case. Specifically, if $\boldsymbol{L}' = \boldsymbol{L} + \sum_{\tau \in z} \widehat{\boldsymbol{\ell}}_\tau$ for some finite set z and $\sum_{\tau \in z} \kappa \|\widehat{\boldsymbol{\ell}}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} \leq \frac{1}{32}$, then $\boldsymbol{w}_{t'}(\boldsymbol{L}') \in \mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$. We apply the last result in Lemma 7 to establish that $\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\kappa}})$ for all $\tau \in m_t$, which in turn allows us to conclude that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m), \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\kappa}})$. Thus, we can repeatedly apply $\|\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}} \leq 8\|\boldsymbol{L}' - \boldsymbol{L}\|_{R_t,\boldsymbol{v}}$ due to Lemma 5, which leads to Lemma 3.

Lemma 3 *Suppose that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t|\mathcal{F}_t] = \boldsymbol{\ell}_t + \boldsymbol{b}_t$ and suppose that Assumption 1 and Assumption 2 hold. Let $t \in [T]$ and $\tau \in m_t \cup \{t\}$. Suppose that $\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \alpha_t$ and $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right] \leq \beta_t^2$. Suppose that for all $t, t' \in [T]$ $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$. Then for all $\boldsymbol{u} \in \mathcal{W}_T$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq \mathbb{E}\left[\sum_{t\in\Lambda}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + R_T(\boldsymbol{u}) - \min_{v\in\mathcal{W}_1} R_1(\boldsymbol{v}) + \sum_{t\in\overline{\Lambda}} 8\beta_t^2$$
$$- \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] + \sum_{t\in\overline{\Lambda}} \left(8\alpha_t^2 |m_t| + 8\alpha_t \,\mathbb{E}\left[\left\| \sum_{\tau\in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]\right) .$$

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 setting we consider

here as their analysis relies on the fact that in the multi-armed bandit setting the constraint in the Lagrangian of the FTRL objective can be expressed in a simple manner, which is not possible in our setting.

To interpret Lemma 3, consider the multi-armed bandit setting with the standard importance-weighted estimator, no skipping $\Lambda = \emptyset$, and regularizer $R(\boldsymbol{v}) = \sum_{i=1}^{K} \frac{1}{\eta}\boldsymbol{v}(i)\log(\boldsymbol{v}(i)) - \frac{1}{\gamma}\log(\boldsymbol{v}(i))$. 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 $\|\boldsymbol{\ell}_t\|_\infty \le 1$, then α_t is $O(\sqrt{\eta})$. The quantity β_t^2 is a bound on the expectation of the squared local norm of the loss estimate, which is $O(\eta K)$. Thus, by choosing $\boldsymbol{u} = \big(1 - \frac{1}{T}\big)\widetilde{\boldsymbol{u}} + \frac{1}{T}\arg\min_{v\in\mathcal{W}} R(\boldsymbol{v})$, we have that the expected regret against $\widetilde{\boldsymbol{u}}$ is of order

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

where we used $\sum_{t=1}^{T}|m_t| = D$. The $d_{\max}K\ln(T)$ term in the above equation comes from the log-barrier part of R, which—when properly tuned—ensures that the FTRL iterates 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, due to the delay, if $\tau, \tau' \in m_t$, then $\widehat{\boldsymbol{\ell}}_\tau$ 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 $\nabla^2 R(\boldsymbol{v}) \succeq \operatorname{diag}\big(\eta\boldsymbol{v}\big)^{-1}$, we can see that

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

where the second inequality is due to Lemma 7, a new result that proves the multi-round stability of FTRL iterates under certain conditions, which can be applied for sufficiently small γ. By 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 $\log(K)/\eta + d_{\max}K\ln(T) + \eta(KT + D)$, which gives a $O\big(\sqrt{(KT + D)\log(K)} + d_{\max}K\ln(T)\big)$ 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 3 implies that we have effectively split the cost of delayed feedback and bandit feedback. We formalize the above in Corollary 4, whose proof can be found in Section 3.2.

Corollary 4 *Under the same assumptions as in Lemma 3, suppose that* $\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t] = \boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top\big(\nabla^2 R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'})\,\middle|\,\mathcal{F}_t\right] = 0$ *for all* $t \in [T]$ *and all* $\tau, \tau' \in m_t$ *where*

$\tau' \neq \tau$. Let $\Lambda = \emptyset$ and let $\mathcal{W}_t = \mathcal{W}$. Then for all $\boldsymbol{u} \in \mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R_T(\boldsymbol{u}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v}) + 16\sum_{t=1}^{T}\beta_t^2 + 16\sum_{t=1}^{T}\alpha_t^2|m_t| \,.$$

3.2 Analysis Details

In this section we present the proofs of Lemma 3 and Corollary 4. We start by developing the necessary tools in Lemmas 5, 6, and 7. Beginning with the former two, both of which are standard and can be found in various forms in the literature.

Lemma 5 *Suppose that Assumption 2 holds. Let $\boldsymbol{v} \in \mathcal{W}_t$ and $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^K$ such that $\boldsymbol{w}_t(\boldsymbol{L}'), \boldsymbol{w}_t(\boldsymbol{L}) \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$, then $\|\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}} \leq 8\|\boldsymbol{L}' - \boldsymbol{L}\|_{R_t,\boldsymbol{v}}$.*

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

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

where the last inequality is due the assumption on $\nabla^2 R_t(\boldsymbol{v})$, which is applicable because if $\boldsymbol{w}_t(\boldsymbol{L}'), \boldsymbol{w}_t(\boldsymbol{L}) \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$ the line segment between $\boldsymbol{w}_t(\boldsymbol{L}')$ and $\boldsymbol{w}_t(\boldsymbol{L})$ is also in $\mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$. Thus $\zeta \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$.

By adding and subtracting $\boldsymbol{L}^\top(\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}'))$ we have that

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

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

$$\|\boldsymbol{L}' - \boldsymbol{L}\|_{R_t,\boldsymbol{v}} \|\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}} \geq \frac{1}{8}\Big(\|\boldsymbol{w}_t(\boldsymbol{L}) - \boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}}\Big)^2 \,,$$

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

Lemma 6 *Suppose that Assumption 2 holds. 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 $\sum_{\tau\in z} \sqrt{\kappa}\|\boldsymbol{y}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} \leq \frac{1}{32}$ and $\mathcal{W}_t = \mathcal{W}_{t'}$, then $\boldsymbol{w}_{t'}(\boldsymbol{L}') \in \mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$.*

Proof Because of the strict convexity of all R_t, to show that $\boldsymbol{w}_{t'}(\boldsymbol{L}') \in \mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ it suffices to show that for all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$

$$\boldsymbol{L'}^\top \boldsymbol{x} + R_{t'}(\boldsymbol{x}) \geq \boldsymbol{L'}^\top \boldsymbol{w}_t(\boldsymbol{L}) + R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})). \tag{4}$$

To see why the strict convexity of $R_{t'}$ is sufficient, suppose that all $\boldsymbol{x}$ that are on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ indeed satisfy (4). For the sake of contradiction suppose that $\boldsymbol{w}_{t'}(\boldsymbol{L}')$ is not in $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$. Let $\boldsymbol{z} = (1-a)\boldsymbol{w}_t(\boldsymbol{L}) + a\boldsymbol{w}_{t'}(\boldsymbol{L}')$ be the point on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ on the segment between $\boldsymbol{w}_t(\boldsymbol{L})$ and $\boldsymbol{w}_{t'}(\boldsymbol{L}')$. Then

$$\begin{aligned}
&\boldsymbol{L'}^\top \boldsymbol{w}_t(\boldsymbol{L}) + R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) \\
&\leq \boldsymbol{L'}^\top \boldsymbol{z} + R_{t'}(\boldsymbol{z}) \\
&< (1-a)(\boldsymbol{L'}^\top \boldsymbol{w}_t(\boldsymbol{L}) + R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))) + a(\boldsymbol{L'}^\top \boldsymbol{w}_{t'}(\boldsymbol{L}') + R_{t'}(\boldsymbol{w}_{t'}(\boldsymbol{L}'))) \\
&\leq \boldsymbol{L'}^\top \boldsymbol{w}_t(\boldsymbol{L}) + R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\,,
\end{aligned}$$

where the first inequality holds because we assumed (4) to be true and z is on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ and the last inequality is by definition of $\boldsymbol{w}_{t'}(\boldsymbol{L}')$ and the assumption that $\mathcal{W}_t = \mathcal{W}_{t'}$. Thus, we have a contradiction, which implies that if all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ satisfy (4), then $\boldsymbol{w}_{t'}(\boldsymbol{L}') \in \mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$. We proceed by showing that all $\boldsymbol{x}$ on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ satisfy (4). Let $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{w}_t(\boldsymbol{L})$. Note that

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

We have

$$\begin{aligned}
\Big(\kappa\boldsymbol{L} + \kappa\nabla R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^\top \boldsymbol{h} &= \Big(\kappa\boldsymbol{L} + \kappa\nabla R_t(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^\top \boldsymbol{h} + \kappa\Big(\nabla R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) - \nabla R_t(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^\top \boldsymbol{h} \\
&\geq \kappa\Big(\nabla R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) - \nabla R_t(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^\top \boldsymbol{h} \\
&\geq -\frac{1}{32}\sqrt{\kappa\boldsymbol{h}^\top\nabla^2 R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\boldsymbol{h}} = -\frac{1}{64}\,,
\end{aligned}$$

where the first inequality is due to the optimality of $\boldsymbol{w}_t(\boldsymbol{L})$, the second inequality is per Assumption 2(a), implying that $(\nabla\kappa R_t(\boldsymbol{w}_t(\boldsymbol{L})) - \nabla\kappa R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})))^\top \boldsymbol{x} \leq \frac{1}{32}\sqrt{\kappa\boldsymbol{x}^\top\nabla^2 R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\boldsymbol{x}}$, and the last equality is due to the fact that $\boldsymbol{h}$ is a point on the boundary of $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$ and thus $\|\boldsymbol{h}\|^\star_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} = \frac{1}{2\sqrt{\kappa}}$. Using Taylor's theorem, there exists ζ on the segment between $\boldsymbol{x}$ and $\boldsymbol{w}_t(\boldsymbol{L})$ such that

$$\begin{aligned}
&\kappa\boldsymbol{L'}^\top \boldsymbol{x} + \kappa R_{t'}(\boldsymbol{x}) - \kappa\boldsymbol{L'}^\top \boldsymbol{w}_t(\boldsymbol{L}) - \kappa R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) \\
&= \kappa(\boldsymbol{L'} - \boldsymbol{L})^\top \boldsymbol{h} + \Big(\kappa\boldsymbol{L} + \kappa\nabla R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^\top \boldsymbol{h} + \frac{\kappa}{2}\boldsymbol{h}^\top\nabla^2 R_{t'}(\zeta)\boldsymbol{h} \\
&\geq \kappa(\boldsymbol{L'} - \boldsymbol{L})^\top \boldsymbol{h} - \frac{1}{64} + \frac{1}{2}\boldsymbol{h}^\top\nabla^2 R_{t'}(\zeta)\boldsymbol{h} \\
&\geq \kappa(\boldsymbol{L'} - \boldsymbol{L})^\top \boldsymbol{h} - \frac{1}{64} + \frac{\kappa}{8}\boldsymbol{h}^\top\nabla^2 R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\boldsymbol{h} \\
&= \kappa(\boldsymbol{L'} - \boldsymbol{L})^\top \boldsymbol{h} + \frac{1}{64}
\end{aligned} \tag{5}$$

where we also used Assumption 2(a) and that $\zeta \in \mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$. Thus, by applying Hölder's inequality we can see that

$$\begin{aligned}\kappa \boldsymbol{L}'^{\top}\boldsymbol{x} + \kappa R_{t'}(\boldsymbol{x}) - \kappa \boldsymbol{L}'^{\top}\boldsymbol{w}_t(\boldsymbol{L}) - \kappa R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) &\geq -\sum_{\tau \in z} \kappa \|\boldsymbol{y}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} \|\boldsymbol{h}\|^{\star}_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} + \frac{1}{64} \\ &= -\frac{1}{2}\sum_{\tau \in z} \sqrt{\kappa}\|\boldsymbol{y}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} + \frac{1}{64} \geq 0\,,\end{aligned}$$

where the equality is due to the fact that $\|\boldsymbol{h}\|^{\star}_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} = \frac{1}{2\sqrt{\kappa}}$ and the final inequality is due to the assumption that $\sum_\tau \sqrt{\kappa}\|\boldsymbol{y}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})} \leq \frac{1}{32}$. ■

The following lemma states that if the local norms of the loss estimates, $\|\widehat{\boldsymbol{\ell}}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$, are small enough, the iterates of FTRL are close across multiple rounds. This is a crucial ingredient in our analysis, as this allows us to use Assumption (a) to control the variance term in Lemma 3. This lemma might be of independent interest.

Lemma 7 *Suppose that Assumption 1 and Assumption 2 hold. Also suppose that for all $t, t' \in [T]$, $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$. Then, for all $t \in [T]$ and all $\tau \in m_t$ we have that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$.*

Proof We will prove the statement by induction. Assume that there exists a $t \in [T]$ such that for all $\tau < t$ and all $s \in m_\tau$, it holds that $\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau) \in \mathcal{D}_{R_\tau}(\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$. Now pick any $s \in m_t$. For the induction step we need to show that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$. The goal is to apply Lemma 6 for which we start by decomposing $o_t \setminus o_s$ into the losses that were already missing at timestep s (and were observed later) and the losses that we incurred and observed after the round s,

$$\begin{aligned}\sum_{\tau \in o_t \setminus o_s} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s)} &= \sum_{\substack{\tau \in o_t \setminus o_s \\ \tau \geq s}} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s)} + \sum_{\tau \in m_s \setminus m_t} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s)} \\ &\leq 2\sum_{\substack{\tau \in o_t \setminus o_s \\ \tau \geq s}} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)} + 2\sum_{\tau \in m_s \setminus m_t} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)} \\ &= 2\sum_{\tau \in o_t \setminus o_s} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\,,\end{aligned}$$

For the inequality, we are applying Lemma 17 using the fact that $\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau) \in \mathcal{D}_{R_\tau}(\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$ for $\tau \in m_s$ and $\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s) \in \mathcal{D}_{R_s}(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$ for $\tau \geq s$ (where we follow $s \in m_\tau$ which follows from $s \in m_t$ and $t \geq \tau$), both of which hold by the inductive assumption. We continue:

$$2\sum_{\tau \in o_t \setminus o_s} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)} \leq \frac{2|o_t \setminus o_s|}{128 d_{\max}} \leq \frac{1}{32}\,,$$

where the first inequality is per the assumption and the second inequality follows by counting the number of elements in $o_t \setminus o_s$, which we do as

$$|o_t \setminus o_s| \leq |\{\widehat{\boldsymbol{\ell}}_{t-2d_{\max}}, \ldots, \widehat{\boldsymbol{\ell}}_{t-1}\}| = 2d_{\max}\,.$$

Since we have now established that $\sum_{\tau \in o_t \setminus o_s} \sqrt{\kappa} \|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t, \boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s)} \leq \frac{1}{128 d_{\max}}$ we can apply Lemma 6 to conclude that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_s(\widehat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$ as $\mathcal{W}_t = \mathcal{W}_s$, which holds by Assumption 1. That completes the induction step as we have chosen s arbitrarily. For the basis of induction it is sufficient to note that $\boldsymbol{w}_1(\widehat{\boldsymbol{L}}_1) \in \mathcal{D}_{R_1}(\boldsymbol{w}_1(\widehat{\boldsymbol{L}}_1), \frac{1}{2\sqrt{\kappa}})$ holds trivially. ■

Now that we have gathered the necessary tools, we can prove our main Lemma.

Lemma 3 (RESTATED) *Suppose that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t | \mathcal{F}_t] = \boldsymbol{\ell}_t + \boldsymbol{b}_t$ and suppose that Assumption 1 and Assumption 2 hold. Let $t \in [T]$ and $\tau \in m_t \cup \{t\}$. Suppose that $\|\boldsymbol{\ell}_\tau\|_{R_t, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \alpha_t$ and $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t, \boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right] \leq \beta_t^2$. Suppose that for all $t, t' \in [T]$ $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$. Then for all $\boldsymbol{u} \in \mathcal{W}_T$,*

$$\mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq \mathbb{E}\left[\sum_{t \in \Lambda} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + R_T(\boldsymbol{u}) - \min_{v \in \mathcal{W}_1} R_1(\boldsymbol{v}) + \sum_{t \in \overline{\Lambda}} 8\beta_t^2$$
$$- \sum_{t \in \overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] + \sum_{t \in \overline{\Lambda}} \left(8\alpha_t^2 |m_t| + 8\alpha_t \, \mathbb{E}\left[\left\| \sum_{\tau \in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau) \right\|_{R_t, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]\right) .$$

Proof The first step of the proof is to establish some base facts including that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$ for all $\tau \in m_t$ and $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m), \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\kappa}})$.

Since $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$, by Lemma 7 we may conclude that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$ for all $\tau \in m_t$ and all t, which is also a prerequisite for Lemma 17. We also note that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\kappa}})$ holds trivially. Now we can conclude that

$$\sum_{\tau \in m_t} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \sum_{\tau \in m_t \cup \{t\}} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq 2 \sum_{\tau \in m_t \cup \{t\}} \sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_\tau\|_{R_t, \boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)} \leq \frac{1}{32} ,$$

where we used Lemma 17 in the second inequality and the assumption on $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$ alongside the fact that $|m_t \cup \{t\}| \leq d_{\max} + 1 \leq 2 d_{\max}$ in the third inequality. By Lemma 6 and Assumption 1 we now know that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m), \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2})$.

We decompose the regret as follows

$$\sum_{t=1}^T \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] = \underbrace{\mathbb{E}\left[\sum_{t \in \Lambda} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right]}_{\text{skipped rounds}} + \underbrace{\sum_{t \in \overline{\Lambda}} - \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right]}_{\text{bias}} \tag{6}$$
$$+ \underbrace{\sum_{t \in \overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t\right]}_{\text{cheating regret}} + \sum_{t \in \overline{\Lambda}} \Big(\mathbb{E}\big[\underbrace{(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t}_{H_1}\big] + \mathbb{E}\big[\underbrace{(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t}_{H_2}\big] \Big) .$$

By Hölder's inequality and Lemma 5

$$\begin{aligned}
H_1 &= \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] \\
&\le \mathbb{E}\left[\|\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)\|^\star_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&\le \mathbb{E}\left[8\|\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^m\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&= \mathbb{E}\left[8\left\|\sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&= \mathbb{E}\left[8\left\|\sum_{\tau\in m_t} (\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau) + \sum_{\tau\in m_t} \boldsymbol{\ell}_\tau\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&\le \mathbb{E}\left[8\left(\left\|\sum_{\tau\in m_t} (\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} + \left\|\sum_{\tau\in m_t} \boldsymbol{\ell}_\tau\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right) \|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&\le 8\alpha_t\, \mathbb{E}\left[\left\|\sum_{\tau\in m_t} (\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] + 8\alpha_t^2 |m_t|\,, \qquad (7)
\end{aligned}$$

where the last inequality is due to the triangle inequality and the assumptions on $\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$. Similarly we bound

$$H_2 = \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t\right] \le 8\beta_t^2\,. \qquad (8)$$

By Lemma 18 we have that

$$\text{cheating regret} = \sum_{t\in\overline{\Lambda}} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t \le R_T(\boldsymbol{u}) - \min_{\boldsymbol{v}\in\mathcal{W}_1} R_1(\boldsymbol{v}). \qquad (9)$$

By combining equations (7), (8), and (9) with the regret decomposition and leaving the skipped rounds and bias untouched we find

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] = \mathbb{E}\left[\sum_{t\in\Lambda} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + \sum_{t\in\overline{\Lambda}} -\mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] \\
&+ \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u})^\top \widehat{\boldsymbol{\ell}}_t\right] + \sum_{t\in\overline{\Lambda}} \Big(\mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m))^\top \boldsymbol{\ell}_t\right] + \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^\star))^\top \widehat{\boldsymbol{\ell}}_t\right]\Big) \\
&\le \mathbb{E}\left[\sum_{t\in\Lambda} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] + R_T(\boldsymbol{u}) - \min_{v\in\mathcal{W}_1} R_1(\boldsymbol{v}) + \sum_{t\in\overline{\Lambda}} 8\beta_t^2 \\
&- \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t\right] + \sum_{t\in\overline{\Lambda}} \left(8\alpha_t^2 |m_t| + 8\alpha_t\, \mathbb{E}\left[\left\|\sum_{\tau\in m_t} (\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]\right)\,.
\end{aligned}$$

which concludes the proof. ■

We conclude this section with the proof of Corollary 4.

Corollary 4 (RESTATED) *Under the same assumptions as in Lemma 3, suppose that* $\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t] = \boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top \left(\nabla^2 R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\right)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,\middle|\, \mathcal{F}_t\right] = 0$ *for all* $t \in [T]$ *and all* $\tau, \tau' \in m_t$ *where* $\tau' \neq \tau$. *Let* $\Lambda = \emptyset$ *and let* $\mathcal{W}_t = \mathcal{W}$. *Then for all* $\boldsymbol{u} \in \mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R_T(\boldsymbol{u}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v}) + 16\sum_{t=1}^{T}\beta_t^2 + 16\sum_{t=1}^{T}\alpha_t^2|m_t| \,.$$

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

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] &= \sum_{\tau\in m_t}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] \\
&= \sum_{\tau\in m_t}\left(\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] - \mathbb{E}\left[\left\|\boldsymbol{\ell}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]\right) \\
&\leq \sum_{\tau\in m_t}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right],
\end{aligned}$$

where we used that $\mathbb{E}\left[(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)^\top\left(\nabla^2 R_t(\boldsymbol{w}(\widehat{\boldsymbol{L}}_t))\right)^{-1}(\widehat{\boldsymbol{\ell}}_{\tau'} - \boldsymbol{\ell}_{\tau'}) \,|\, \mathcal{F}_t\right] = 0$ 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

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\widehat{\boldsymbol{\ell}}_\tau - \boldsymbol{\ell}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] &\leq \sqrt{\sum_{\tau\in m_t}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]} \\
&\leq \sqrt{4\sum_{\tau\in m_t}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right]} \leq \sqrt{4|m_t|\beta_t^2}\,, \qquad (10)
\end{aligned}$$

where in the second inequality we used Lemma 7 together with Lemma 17. Finally, the third inequality of (10) is due to the assumptions of Lemma 3. We conclude by substituting this bound into the results of Lemma 3,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \boldsymbol{u})^\top\boldsymbol{\ell}_t\right] &\leq R_T(\boldsymbol{u}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v}) + 8\sum_{t=1}^{T}\beta_t^2 + \sum_{t=1}^{T}\left(8\alpha_t^2|m_t| + 16\sqrt{|m_t|\alpha_t^2\beta_t^2}\right) \\
&\leq R_T(\boldsymbol{u}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v}) + 16\sum_{t=1}^{T}\beta_t^2 + 16\alpha_t^2\sum_{t=1}^{T}|m_t|,
\end{aligned}$$

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

4. Combinatorial Semi-Bandits

In this section, we demonstrate how to apply our generic FTRL approach to combinatorial semi-bandits (CMAB) with delayed feedback. As outlined in the introduction, combinatorial semi-bandits extend multi-armed bandits to be able to efficiently deal with combinatorial decision spaces

Algorithm 1: Delayed FTRL for combinatorial semi-bandits

Input: Regularizers $\{R_t\}_{t\geq 1}$ defined in (11), including hyperparams $\gamma \in (0,1)$ and $\{\eta_t\}_{t\geq 1}$.
for $t \in [T]$ **do**
  Observe $\boldsymbol{a}_\tau \odot \boldsymbol{\ell}_\tau$ for $\tau \in o_t \setminus o_{t-1}$.
  Find loss estimators $\widehat{\boldsymbol{\ell}}_\tau(i) = \frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau, i)}$ for new observations $\tau \in o_t \setminus o_{t-1}$.
  Compute $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v}\in\mathcal{W}} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$.
  Find probability distribution $\boldsymbol{p}_t$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$.
  Draw and play $\boldsymbol{a}_t \sim \boldsymbol{p}_t$.
end for

and have been used in portfolio management (Ni et al., 2023) and recommendation systems (Louëdec et al., 2015) among others. In combinatorial semi-bandits the learner picks an action $\boldsymbol{a}_t \in \mathcal{A}$ at each timestep t. The actionset $\mathcal{A} \subseteq \{0,1\}^K$ is given as part of the problem formulation. The loss of the learner is defined as $\boldsymbol{a}_t^\top \boldsymbol{\ell}_t$, where $\boldsymbol{\ell}_t \in [-1,1]^K$. In the combinatorial semi-bandit setting 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. A practical example is a path-finding problem. Consider a directed weighted graph, where the weight on the edges corresponds to some cost associated with traversing an edge. The objective of the learner is to reach a goal state while incurring the least loss. In this setting the dimension of the actions is equal to the number of edges on the graph and the actionset $\mathcal{A}$ is the set of all valid paths from the starting state to the goal state. The loss is the cost associated with each edge and the feedback is either the individual weights of the edges traversed for semi-bandits or the entire cost of the path taken in full bandits.

We define the pseudo-regret in this setting as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t\right] \qquad \text{with} \qquad \boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} \boldsymbol{a}^\top \boldsymbol{\ell}_t \,.$$

Algorithm Algorithm 1 is inspired by the algorithm of Audibert et al. (2014). In any given round t, Algorithm 1 first computes $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$, the solution of the FTRL optimization problem of Eq. (1) over the convex hull of the action set, that is with $\mathcal{W} = \mathrm{Conv}(\mathcal{A})$. In this setting we are not skipping rounds and the domain is constant. $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$ can be computed efficiently using standard methods from convex optimisation if $\mathrm{Conv}(\mathcal{A})$ can be described in a polynomial number of linear constraints, see Nemirovski (2004). Then, it constructs a probability distribution $\boldsymbol{p}_t$ over $\mathcal{A}$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$. How to construct $\boldsymbol{p}_t$ and if it can be sampled from efficiently depends on the actionset and for many commonly used actionsets, like m-sets and spanning trees, there exist efficient algorithms. The path finding problem outlined above can also be solved efficiently by relaxing the convex hull of paths in the directed graph to so called unit flows, leading to a runtime of $O(n^4)$ where n is the number of nodes in the path finding problem (Koolen et al., 2010). For a more complete discussion on the computational efficiency of FTRL style combinatorial bandit algorithms and for which actionsets $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$ and $\boldsymbol{p}_t$ can be obtained efficiently we refer to Koolen et al. (2010), Cesa-Bianchi and Lugosi (2012), and Audibert et al. (2014). The estimator of loss is

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

$$R_t(\boldsymbol{v}) = \sum_{i=1}^{K} \Big(\frac{1}{\eta_t} \boldsymbol{v}(i) \log(\boldsymbol{v}(i)) - \frac{1}{\gamma} \log(\boldsymbol{v}(i)) \Big), \tag{11}$$

where $\eta_t > 0$ and $\gamma > 0$ are hyperparameters.

Main Result and Discussion The main result of this section is Theorem 8.

Theorem 8 *Suppose that* $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_1 \leq B$. *Algorithm 1 with*

$$\eta_t = \min\left\{ \sqrt{\frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{16(B\sum_{\tau=1}^{t}|m_t| + Kt)}}, \frac{B^2\left(1+\log\left(\frac{K}{B}\right)\right)}{128K(Bd_{\max}+K)} \right\}, \qquad \gamma = \frac{1}{128\sqrt{B}d_{\max}},$$

guarantees that

$$\mathcal{R}_T \leq 12\sqrt{B\left(1+\log\left(\frac{K}{B}\right)\right)(KT+BD)} + 128K^2 d_{\max} + 128\sqrt{B}d_{\max}K\log(T) \ .$$

The result is based on Corollary 4. After confirming the conditions on the regularizer R_t, the proof finds $\alpha_t = \sqrt{\eta_t B}$ and $\beta_t^2 = \eta_t K$. The last thing to do is to bound the size of the regularizer $R_T(\boldsymbol{u})$ on the comparator $\boldsymbol{u}$, which is a term that also arises from Corollary 4. As R_t tends to infinity on parts of the boundary of $\mathcal{W}$ we have to choose a $\boldsymbol{u} \neq \boldsymbol{a}^*$ and we pick $\boldsymbol{u}$ as the best point in hindsight in a slightly shrunken actionset. That allows us to bound $R_T(\boldsymbol{u})$ in exchange for a small additive bias term. The full proof can be found in Appendix A.

Theorem 9 shows that our results are optimal up to log-factors.

Theorem 9 *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*

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

The proof for Theorem 9 can be found in Appendix A. When using an action-set constructed of basis vectors, we recover the delayed multi-armed bandit setting, in which we match the optimal upper bound for delayed adversarial bandits due to Zimmert and Seldin (2020) up to constants and log-factors. In the non-delayed setting, we have $D = 0$ and we recover a bound of $O(\sqrt{B\left(1+\log\left(\frac{K}{B}\right)\right)KT})$, which also matches the optimal upper bound of order by Audibert et al. (2014) up to constants.

5. Linear Bandits

In this section, we show how to apply our analysis of FTRL to linear bandits with delayed feedback, which is an instance of our general setting for $\boldsymbol{\ell}_t \in \mathbb{R}^K$ 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 the domain is constant with $\mathcal{W} \subseteq \mathcal{B}(B)$, where $\mathcal{B}(B)$ is an Euclidean ball with radius B. We are not skipping rounds in this setting.

Algorithm 2: Delayed FTRL for linear bandits

Input: ν-self concordant barrier Ψ for $\mathcal{W}$, hyperparameters $\{\eta_t, \gamma_t\}_{t\geq 1}$.
Initialize: $\widetilde{\Psi}(\cdot) = \Psi(\cdot) - \min_{\boldsymbol{v}\in\mathcal{W}} \Psi(\boldsymbol{v})$ and $R_t(\cdot) = \frac{1}{\eta_t}\|\cdot\|_2^2 + \frac{1}{\gamma_t}\widetilde{\Psi}(\cdot)$ for $t \geq 1$.
for $t = 1, ..., T$ **do**
  Observe $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$ for $\tau \in o_t \setminus o_{t-1}$.
  Compute $\widehat{\boldsymbol{\ell}}_\tau = K\boldsymbol{\ell}_\tau^\top \boldsymbol{a}_\tau \big(\nabla^2\Psi(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau))\big)^{1/2}\boldsymbol{v}_\tau$ for new observations $\tau \in o_t \setminus o_{t-1}$.
  Compute $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v}\in\mathcal{W}} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$.
  Sample $\boldsymbol{v}_t$ uniformly from the unit sphere.
  Play $\boldsymbol{a}_t = \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) + \big(\nabla^2\Psi(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{-1/2}\boldsymbol{v}_t$.
end for

Algorithm Our algorithm for the linear bandit setting is inspired by Abernethy et al. (2008), who provide an algorithm with 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 $R_t(\boldsymbol{v}) = \frac{1}{\eta_t}\|\boldsymbol{v}\|_2^2 + \frac{1}{\gamma_t}\widetilde{\Psi}(\boldsymbol{v})$, where $\widetilde{\Psi}(\boldsymbol{v}) = \Psi(\boldsymbol{v}) - \min_{\boldsymbol{v}'\in\mathcal{W}} \Psi(\boldsymbol{v}')$ for a ν-self-concordant barrier function Ψ. For a thorough introduction to self-concordant barriers, we refer the reader to (Nesterov and Nemirovskii, 1994). In Appendix B, we recall the most important properties, which can be found in (Nemirovski and Todd, 2008, Section 2). The main reason for using self-concordant barriers is to adhere to Assumptions 2(a) and 2(b). As detailed in Appendix B, these are standard properties of self-concordant barriers.

Specific examples of self-concordant barriers are $f(x) = -\log(x)$, which is 1-self-concordant for the non-negative reals, $f(\boldsymbol{x}) = -\log(1 - \|\boldsymbol{x}\|_2^2)$, which is 1-self-concordant for the unit ball, the 1-self concordant barrier $f(\boldsymbol{x}) = -\log(b - \boldsymbol{a}^\top \boldsymbol{x})$ for linear constraints $\boldsymbol{a}^\top \boldsymbol{x} \leq b$, and the entropic barrier, which is defined as

$$f(\boldsymbol{x}) = \sup_{\theta\in\mathbb{R}^d} \{\langle \boldsymbol{x}, \theta\rangle - f^\star(\theta)\} \qquad \text{where } \; f^\star(\theta) = \ln\Big(\int_{\mathcal{W}} \exp(\langle \boldsymbol{x}, \theta\rangle)\,\mathrm{d}\boldsymbol{x}\Big)\,,$$

which is a d-self-concordant barrier for any $\mathcal{W}$. Unfortunately, even though the entropic barrier is a self-concordant barrier for all domains, it can not always be efficiently computed.

Finally, we turn to the way we choose the action $\boldsymbol{a}_t \in \mathcal{A}$ and the construction of the estimator. We use $\boldsymbol{a}_t = \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) + \big(\nabla^2\Psi(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{-1/2}\boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is sampled i.i.d. from the uniform distribution over the unit sphere. To see that $\boldsymbol{a}_t \in \mathcal{A}$, note that $\mathcal{D}_\Psi(\boldsymbol{w}, 1) \subseteq \mathcal{W} = \mathcal{A}$ for any $\boldsymbol{w} \in \mathcal{W}$ (see Appendix B). Since $\|\boldsymbol{a}_t - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)\|_{\Psi,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} = 1$, we have that $\boldsymbol{a}_t \in \mathcal{W}$. As for the loss estimate, we use $\widehat{\boldsymbol{\ell}}_t = K\boldsymbol{\ell}_t^\top \boldsymbol{a}_t\big(\nabla^2\Psi(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{1/2}\boldsymbol{v}_t$, which can be seen to an unbiased estimator for $\boldsymbol{\ell}_t$ after observing that $\mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_t^\top|\mathcal{F}_t] = \frac{1}{K}\boldsymbol{I}$.

Given that $\Psi(\cdot), \nabla\Psi(\cdot)$, and $\nabla^2\Psi(\cdot)$ can be efficiently computed there are two computationally demanding steps in Algorithm 2: the computation of $w_t(\widehat{\boldsymbol{L}}_t)$ and the computation of $\big(\nabla^2\Psi(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{1/2}$ and its inverse. $\big(\nabla^2\Psi(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\big)^{1/2}$ and its inverse can be computed through an eigenvalue decomposition, which can be done in $O(K^3)$. Abernethy et al. (2008) show that an approximation of $w_t(\widehat{\boldsymbol{L}}_t)$ can be computed in $O(K^2)$ per round by using the damped Newton method. This approximation maintains the same same regret bound up to constants. The implementation as well as an overview of the analysis can be found in Appendix B.1.

Main Result and Discussion We arrive at the main result of this section.

Theorem 10 *Suppose that $T > 100$ and $B \geq 1$. Algorithm 2, run with a ν-self-concordant barrier Ψ and with*

$$\gamma_t = \min\left\{\frac{1}{256BKd_{\max}}, \sqrt{\frac{\nu\log\left(1+\sqrt{T}\right)}{16B^2K^2t}}\right\}$$
$$\eta_t = \min\left\{\frac{B}{256d_{\max}}, \sqrt{\frac{B^2}{16\sum_{\tau=1}^{t}|m_t|}}\right\},$$

guarantees that, for any $\boldsymbol{u} \in \mathcal{W}$,

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq 12BK\sqrt{\nu T\log(1+\sqrt{T})} + 12B\sqrt{D} + 2B\sqrt{T} + 512BKd_{\max}\nu\log(1+\sqrt{T})\,.$$

The proof of Theorem 10 can be found in Appendix B. It follows from an application of Corollary 4 and carefully tuning the learning rates.

Let us put Theorem 10 in perspective. For arbitrary $\mathcal{W}$ we can use the entropic barrier as the regularizer, which means $\nu = d$ and thus algorithm 2 has a $\widetilde{O}(K^{2/3}\sqrt{T} + \sqrt{D})$ regret bound. For constant delay, i.e. $d_t = d$, Ito et al. (2020a) show that continuous exponential weights obtains a $\widetilde{O}(K\sqrt{T}+\sqrt{dT})$ regret bound. Even though this algorithm can be computed in $\mathrm{poly}(K, T, B)$ time, the algorithm is far from practical. In contrast, (an approximation of) algorithm 2 can be computed in $O(K^3)$ time, with only a slightly worse regret bound. Huang et al. (2023) provide an algorithm with simular computational complexity as algorithm 2, but their regret bound is $\widetilde{O}(K^{2/3}\sqrt{T} + K^2\sqrt{D})$, which contains an unnecessary dependence on the dimension K in the delay term of the regret bound. However, it seems that the regret bound of Huang et al. (2023) can be improved to $\widetilde{O}(K\sqrt{\nu T} + K\sqrt{\nu D})$. In their terminology: Banker-BOLO is $(O(\nu\log(T), K^2))$-stable, which together with Theorem 4.6 of Huang et al. (2023) leads to a $\widetilde{O}(K\sqrt{\nu T} + K\sqrt{\nu D})$ regret bound. Still, the unnecessary dependence on the dimension K in the delay term of the regret bound remains.

6. Adversarial Markov Decision Processes (MDPs)

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 start with a presentation of the model and regret minimization framework.

A finite-horizon episodic adversarial MDP is defined by $\mathcal{M} = (\mathcal{S}, \mathfrak{A}, H, p, \{\boldsymbol{\ell}_t\}_{t=1}^T, s_{\text{init}})$, where $\mathcal{S}$ and $\mathfrak{A}$ are finite state and action spaces of sizes S and A, respectively, H is the horizon, T is the number of episodes, and $s_{\text{init}} \in \mathcal{S}$ is the initial state. The transition function is $p : [H] \times \mathcal{S} \times \mathfrak{A} \to \Delta_{\mathcal{S}}$, where $\Delta_{\mathcal{S}}$ is the simplex over the states and $p(s' \mid h, s, a)$ is the probability of moving to s' when taking action a in state s at time h. The learner interacts with the environment over T episodes of length H each. At the beginning of episode t, the learner picks a policy $\pi_t = [H] \times \mathcal{S} \to \Delta_{\mathfrak{A}}$ and starts in the initial state $s_{t,1} = s_{\text{init}}$. For each $h \in [H]$, the learner observes the current state $s_{t,h} \in \mathcal{S}$, draws an action from the policy $a_{t,h} \sim \pi_t(\cdot \mid h, s_{t,h})$, and transitions

to the next state $s_{t,h+1} \sim p(\cdot \mid h, s_{t,h}, a_{t,h})$. The cost functions $\boldsymbol{\ell}_t \in [0,1]^{HSA}$ are chosen by an oblivious adversary, and the feedback of episode t contains the elements of the cost function corresponding to 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 $V_t^{\pi}(h,s) = \mathbb{E}\left[\sum_{h'=h}^{H} \boldsymbol{\ell}_t(h', s_{h'}, a_{h'}) \mid s_h = s, \pi, p\right]$ 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

$$\mathcal{R}_T = \sum_{t=1}^{T} V_t^{\pi_t}(1, s_{\text{init}}) - \min_{\pi \in \Pi} \sum_{t=1}^{T} V_t^{\pi}(1, s_{\text{init}}) \,,$$

where Π is the set of all policies admitted by $\mathcal{M}$.

Given a policy π and a transition function p', the occupancy measure $\boldsymbol{q}^{\pi,p'} \in [0,1]^{HS^2A}$ is a vector, where $\boldsymbol{q}^{\pi,p'}(h,s,a,s')$ is the probability to visit state s at time h, take action a and transition to state s'. We also denote

$$\boldsymbol{q}^{\pi,p'}(h,s,a) = \sum_{s'} \boldsymbol{q}^{\pi,p'}(h,s,a,s') \qquad \text{and} \qquad \boldsymbol{q}^{\pi,p'}(h,s) = \sum_{a} \boldsymbol{q}^{\pi,p'}(h,s,a).$$

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

$$\pi(a \mid h, s) = \frac{\boldsymbol{q}^{\pi,p'}(h,s,a)}{\boldsymbol{q}^{\pi,p'}(h,s)} \qquad \text{and} \qquad p'(s' \mid h, s, a) = \frac{\boldsymbol{q}^{\pi,p'}(h,s,a,s')}{\boldsymbol{q}^{\pi,p'}(h,s,a)}.$$

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 $\Pi = \{\pi : [H] \times \mathcal{S} \to \Delta_{\mathcal{A}}\}$. 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., $\langle \boldsymbol{q}^{\pi,p'}, \boldsymbol{\ell} \rangle = \sum_{h,s,a} \boldsymbol{q}^{\pi,p'}(h,s,a)\boldsymbol{\ell}(h,s,a)$. Thus, the regret becomes

$$\mathcal{R}_T = \sum_{t=1}^{T} \langle \boldsymbol{q}^{\pi_t,p}, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{q} \in \Delta(\mathcal{M})} \sum_{t=1}^{T} \langle \boldsymbol{q}, \boldsymbol{\ell}_t \rangle \,.$$

Whenever p' is omitted from the notation $\boldsymbol{q}^{\pi,p'}$, it is understood to be the true transition function p.

With that in hand, the adversarial MDP setting is an instance of the online learning framework where $\boldsymbol{\ell}_t \in [0,1]^{HSA}$, $\mathcal{A} = \Delta(\mathcal{M})$ as the set of all occupancy measures 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. Note that in this context, $\boldsymbol{w}_t(\boldsymbol{L})$ is a vector of dimension HS^2A—we will denote by $\boldsymbol{w}_t(\boldsymbol{L}, h, s, a, s')$ the (h, s, a, s') element of it and also define $\boldsymbol{w}_t(\boldsymbol{L}, h, s, a) = \sum_{s'} \boldsymbol{w}_t(\boldsymbol{L}, h, s, a, s')$.

Algorithm Algorithm 3 is based on the general framework presented in Section 3. To satisfy the stability conditions required for Lemma 3, we employ a hybrid regularization of negative entropy and log-barrier just like in the combinatorial bandit case:

$$R_t(\boldsymbol{v}) = \frac{1}{\eta_t} \sum_{h,s,a,s'} \boldsymbol{v}(h,s,a,s') \log \boldsymbol{v}(h,s,a,s') - \frac{1}{\gamma} \sum_{h,s,a,s'} \log \boldsymbol{v}(h,s,a,s'). \tag{12}$$

Algorithm 3: Delayed FTRL for adversarial MDPs

Input: Regularizers $\{R_t\}_{t\geq 1}$ defined in (12).
for $t = 1, ..., T$ **do**
  Observe feedback $\boldsymbol{\ell}_t(h, s_{\tau,h}, a_{\tau,h})$ for $h \in [H], \tau \in o_t \setminus o_{t-1}$.
  Compute upper occupancy bounds $\boldsymbol{q}_\tau^{\max}(h,s,a) = \max_{\widehat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi_\tau,\widehat{p}}(h,s,a)$.
  Compute $\widehat{\boldsymbol{\ell}}_\tau(h,s,a) = \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}\boldsymbol{\ell}_\tau(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)}$ for $\tau \in o_t \setminus o_{t-1}$.
  Compute $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v}\in\mathcal{W}} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$ and policy $\pi_t(a \mid h, s) = \frac{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)}{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s)}$.
  Play episode t with policy π_t
end for

The main difference is that some of the elements of the occupancy measures may be 0 regardless of the chosen policy (if $p(s' \mid h, s, a) = 0$ then $\boldsymbol{q}^\pi(h, s, a, s') = 0$), in which case the regularization is not well-defined. To avoid that, we augment the set of occupancy measures to include occupancy measures for which the associated transition probability differs a little bit from the true transition probabilities

$$\Delta(\mathcal{P}) = \{\boldsymbol{q}^{\pi,\widehat{p}} : \pi \in \Pi, \widehat{p} \in \mathcal{P}\} \qquad \text{where} \qquad \mathcal{P} = \left\{\widehat{p} :\ \|\widehat{p} - p\|_\infty \leq \frac{1}{THSA}\right\} .$$

To complete the presentation of adversarial MDPs as an instance of our online learning framework, we define the constant domain as $\mathcal{W} = \Delta(\mathcal{P})$. Also, we are not skipping any rounds. This construction allows us to establish the following properties of $\mathcal{W}$:

Lemma 11 *$\mathcal{W}$ satisfies the following:*

1. *For any $\boldsymbol{q} \in \Delta(\mathcal{M})$, there exists $\widetilde{\boldsymbol{q}} \in \mathcal{W}$ such that $\min_{h,s,a,s'} \widetilde{\boldsymbol{q}}(h,s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$ and $\|\boldsymbol{q} - \widetilde{\boldsymbol{q}}\|_1 \leq \frac{2H}{T}$.*
2. *Given $\boldsymbol{v} \in \mathcal{W}$, let π be defined by $\pi(a \mid h, s) = \frac{\boldsymbol{v}(h,s,a)}{\boldsymbol{v}(h,s)}$ and $\boldsymbol{q}^{max}(h,s,a) = \max_{\widehat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi,\widehat{p}}(h,s,a)$. Then, $\|\boldsymbol{q}^\pi - \boldsymbol{v}\|_1 \leq \frac{2H}{T}$ and $\|\boldsymbol{q}^{max} - \boldsymbol{v}\|_1 \leq \frac{4H^2S}{T}$.*

The proof can be found in Appendix C. The importance-weighted loss estimator for Algorithm 3 is inspired by Jin et al. (2020),

$$\widehat{\boldsymbol{\ell}}_\tau(h,s,a) = \frac{\mathbb{I}\{s_{\tau,h} = s, a_{\tau,h} = a\}\boldsymbol{\ell}_\tau(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)} ,$$

where $\boldsymbol{q}_\tau^{\max}(h,s,a) = \max_{\widehat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi_\tau,\widehat{p}}(h,s,a)$ is an upper bound on the occupancy measure for each state h, s, a when following policy π_τ. That means that $\widehat{\boldsymbol{\ell}}_\tau$ is underestimating the actual losses and is a slightly biased estimator.

Note that $\mathcal{W}$ is a convex set defined by $O(HS^2A)$ linear equality and inequality constrains. In practice, we can eliminate the equality constraints through a simple re-parameterization, ensuring the variables lie within the linear subspace that satisfies the constraints (Boyd and Vandenberghe, 2004), thereby making the interior of the decision set non-empty. Using that, we can

apply the interior-point method to approximate the solution to the FTRL step with running time $O(\text{poly}(H,S,A)\log T)$—Nemirovski (2004); see also Abernethy et al. (2012)—with an error up to $1/T$ (which affects the regret only by a constant). In addition, $\boldsymbol{q}_t^{\max}$ can be computed efficiently as well using dynamic programming (Jin et al., 2020). We note that, while this approach is technically efficient, it becomes impractical when the number of states is significantly large.

Main Result and Discussion The main result of this section is Theorem12.

Theorem 12 *Suppose that $T \geq H$. Algorithm 3 with*

$$\gamma = \frac{1}{128Hd_{\max}} \qquad \eta_t = \min\left\{\frac{\log(SA)}{96HSA\sqrt{SAd_{\max}+d_{\max}^2}}, \frac{\sqrt{\log(SA)}}{\sqrt{SAt+\sum_{\tau=1}^{t}|m_t|}}\right\}$$

guarantees

$$\mathbb{E}[\mathcal{R}_T] \leq 72H\sqrt{\log(SA)(TSA+D)} + 1338d_{\max}H^2S^2A^2\log\left(HSAT\right) .$$

The proof relies on yet another regret decomposition given by

$$\mathcal{R}_T = \sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle = \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle}_{\text{ERROR}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle}_{\text{REG}} + \underbrace{\sum_{t=1}^{T}\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle}_{\text{SHIFT-PENALTY}} .$$

$\Delta(\mathcal{P})$ is only slightly larger than $\Delta(\mathcal{M})$, and we can easily bound ERROR using the first property in Lemma 11. Since $R_T(\boldsymbol{u})$ can be arbitrarily large near the boundary of the domain, we slightly shift $\boldsymbol{u}$ to $\widetilde{\boldsymbol{u}}$ using the first property in Lemma 11 to ensure that (i) $R_T(\widetilde{\boldsymbol{u}}) \leq \widetilde{O}\big(\frac{HS^2A}{\gamma}\big)$, and (ii) SHIFT-PENALTY is bounded by $2H$. We can not apply Corollary 4 to bound REG because of the bias in our estimator. We apply Lemma 3 instead. By Lemma 25 we can show that R_t satisfies Assumption 2 and that $R_T(\widetilde{\boldsymbol{u}}) - \min_{\boldsymbol{v}\in\mathcal{W}_1} R_1(\boldsymbol{v})$ is bounded by $\widetilde{O}\big(\frac{H}{\eta_T} + \frac{HS^2A}{\gamma}\big)$.

The fact that $\boldsymbol{q}_t^{\max}(h,s,a)$ upper bounds both $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)$ and $\boldsymbol{q}^{\pi_t}(h,s,a)$ allows us to keep local norms related to α_t and β_t small. In addition, using the second property in Lemma 11, we can also show that the estimator's bias is only of order $1/T$ (ignoring S,H factors). The main part of the remaining of the proof deals with the term $\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$. This term, which arises because we are using biased estimators, is not present when applying Corollary 4. The full proof can be found in Appendix C.

The algorithm is optimal, matching the lower bound of Lancewicki et al. (2022a) up to log-factors and improves on previous state-of-the-art regret bounds $\widetilde{O}\big(H^2S\sqrt{AT} + H(HSA)^{1/4}\sqrt{D}\big)$ by Jin et al. (2022) and $\widetilde{O}\big(H^2\sqrt{SAT} + H^3\sqrt{D}\big)$ by Lancewicki et al. (2023).

7. Adversarial MDPs with Unknown Transitions

In this section, we apply our FTRL approach to adversarial Markov Decision Processes (MDPs) setting detailed in Section 6, for the case that the transition function is unknown to the learner in advance. We show that it yields the first algorithm that handles delay asymptotically optimal in this setting, up to sub-optimality gaps that already exist in the non-delayed setting.

Algorithm 4: Delayed FTRL for adversarial MDPs with unknown transitions

Initialize $j = 1$, $\widehat{\mathcal{P}}_1$ as the set of all transition functions, $\mathcal{W}_0 = \Delta(\widehat{\mathcal{P}}_1)$.
For all $(h, s, a, s') \in [H] \times \mathcal{S} \times \mathfrak{A} \times \mathcal{S}$ set $N_0(s'|h, s, a) = N_1(s'|h, s, a) = 0$.
for $t = 1, ..., T$ **do**
`/* Transition estimation and epochs */`
Observe trajectories $(s_{\tau,h}, a_{\tau,h})$ for $h \in [H], \tau \in \widetilde{o}_t \setminus \widetilde{o}_{t-1}$.
Update counters: $N_j(s_{\tau,h+1}|h, s_{\tau,h}, a_{\tau,h})$ += 1 for $h \in [H], \tau \in \widetilde{o}_t \setminus \widetilde{o}_{t-1}$.
if $\exists h$ such that $N_j(h, s_{\tau,h}, a_{\tau,h}) \geq \max\{1, 2N_{j-1}(h, s_{\tau,h}, a_{\tau,h})\}$ **then**
j += 1
For all $(h, s, a, s') \in \mathcal{S} \times \mathfrak{A} \times \mathcal{S}$, set $N_j(s'|h, s, a) = N_{j-1}(s'|h, s, a)$.
Update set $\widehat{\mathcal{P}}_j$ as in equation (13).
Set $\mathcal{W}_t = \bigcap_{j'=1}^{j} \Delta(\widehat{\mathcal{P}}_{j'})$. If $\mathcal{W}_t = \emptyset$ then set $\mathcal{W}_t = \Delta(\widehat{\mathcal{P}}_j)$.
Skip all rounds that are missing by adding all elements in $\widetilde{m}_t$ to Λ.
end if
`/* Loss estimation and episode execution */`
If $\mathcal{W}_t$ is not defined by an epoch change, set $\mathcal{W}_t = \mathcal{W}_{t-1}$.
Observe feedback $\boldsymbol{\ell}_t(h, s_{\tau,h}, a_{\tau,h})$ for $h \in [H], \tau \in o_t \setminus o_{t-1}$.
Compute upper occupancy bounds $\boldsymbol{q}_\tau^{\max}(h, s, a) = \max_{\widehat{p} \in \widehat{\mathcal{P}}_j} \boldsymbol{q}^{\pi_\tau, \widehat{p}}(h, s, a)$.
Compute $\widehat{\boldsymbol{\ell}}_\tau(h, s, a) = \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}\boldsymbol{\ell}_\tau(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)+\xi}$ for new observations $\tau \in o_t \setminus o_{t-1}$.
Compute $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v} \in \mathcal{W}_t} \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$ and policy $\pi_t(a \mid h, s) = \frac{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a)}{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s)}$.
Play episode t with policy π_t
end for

Algorithm Algorithm 4 is very similar to the one presented in Section 6 for the known transitions case, with one main differences: In order to estimate the transition function we use a delayed version of the confidence set for the transition function of Jin et al. (2020). The confidence sets are updated in epochs. Specifically, the algorithm maintains counters $N_j(s' \mid h, s, a)$ to track the number of visits to state-action pair (s, a) and transitioning to state s' at time h. Mirroring the notation used for occupancy measures we also define $N_j(h, s, a) = \sum_{s'} N_j(s' \mid h, s, a)$ as the number of visits to state-action pair (s, a) at time h. In each epoch j, if the counter $N_j(h, s, a)$ doubles compared to $N_{j-1}(h, s, a)$ for some triplet (h, s, a), a new epoch (epoch $j + 1$) is initiated. The confidence set in epoch j is defined as

$$\widehat{\mathcal{P}}_j = \left\{\widehat{p} : \left|\widehat{p}(s'|h, s, a) - \overline{p}_j(s'|h, s, a)\right| \leq \epsilon_j(s'|h, s, a)\,,\ \forall (h, s', a, s) \in [H] \times \mathcal{S} \times \mathfrak{A} \times \mathcal{S}\right\}, \tag{13}$$

where

$$\epsilon_j(s'|h, s, a) = 2\sqrt{\frac{\overline{p}_j(s'|h, s, a)\log(HSAT^3)}{\max\{1, N_j(h, s, a) - 1\}}} + \frac{14\log(HSAT^3)}{3\max\{1, N_j(h, s, a) - 1\}},$$

for $\overline{p}_j(s'|h, s, a) = \frac{N_j(s'|h,s,a)}{N_j(h,s,a)}$ being the empirical transition, calculated using the visit counts $N_j(s'|h, s, a)$ at the beginning of the epoch. The domain is constant in each epoch and is computed as the intersection over all previous $\Delta(\widehat{\mathcal{P}})$. That is, if round t is in epoch j, then $\mathcal{W}_t =$

$\bigcap_{j'=1}^{j} \Delta(\widehat{\mathcal{P}}_{j'})$. Lemma 13 below shows that the true transition function is contained in our confidence set with high probability.

Lemma 13 *With probability at least $1 - 4/T^2$, we have $p \in \widehat{\mathcal{P}}_j$ for all j.*

Proof The proof is a straightforward modification of the proof of Lemma 2 of Jin et al. (2020). ■

As a consequence of Lemma 13, the occupancy measure of the benchmark policy is contained in each domain. The reason that we define $\mathcal{W}_t$ as the intersection of $\Delta(\mathcal{P}_{j'})$ up to the current epoch is to ensure that $\mathcal{W}_t \subseteq \mathcal{W}_{t+1}$. This will later be crucial in the analysis to apply the be-the-leader lemma (Lemma 18 in Appendix F). In order to ensure that Assumption 1, specifically the fact that $\mathcal{W}_t = \mathcal{W}_\tau$ for all outstanding observations $\tau \in m_t$, is met, we skip all outstanding rounds at the beginning of a new epoch. The loss estimator is an importance-weighted estimator with $\boldsymbol{q}_\tau^{\max}(h, s, a)$ being an upper confidence estimate for $\boldsymbol{q}^{\pi,p}(h, s, a)$. In addition we add a small bias ξ, so that the estimator is also bounded under the bad event.

In terms of implementation, we can take the same approach as in the known transition case (Section 6), with the main difference being that the decision set is defined by $O(H^2S^3A^2 \log T)$ linear constraints due to the number of epochs being at most $HSA \log T$. Thus, the FTRL solution can be $1/T$-approximated with a running time of $O(\text{poly}(H, S, A, \log T))$. As noted before, while this approach is technically efficient, it becomes impractical when the number of states is large.

Main Result and Discussion The main result of this section is Theorem 14. To slightly simplify the analysis, we choose $\eta_t = \eta$. However, a decreasing learning rate is also possible, as shown for MDPs with known transitions in section 6.

Theorem 14 *Algorithm 4 with $\gamma = \frac{1}{128\sqrt{H} d_{\max}}$, $\eta = \frac{\sqrt{\log(SA)}}{\sqrt{SAT+D}}$, $\xi = \frac{1}{T}$ and $T \geq 4$ guarantees,*

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] \lesssim\ & H^2 S \sqrt{AT \log(HSAT)} + H \sqrt{D \log(SA)} \\ & + H^3 S^2 A \log(HSAT)\, d_{\max} + H^3 S^3 A \log^2(HSAT).\end{aligned}$$

The proof relies on the same regret decomposition as in Section 6.

$$\mathcal{R}_T = \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t \rangle}_{\text{ERROR}} + \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t \rangle}_{\text{REG}} + \underbrace{\sum_{t=1}^{T} \langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle}_{\text{SHIFT-PENALTY}} \right].$$

The ERROR and SHIFT-PENALTY terms are bounded using standard arguments (see Lemma 29 in the appendix). To bound the REG term we will apply Lemma 3 just as in the previous sections. Since we now have a changing domain we need to ensure that all loss estimators that we observe in round t, that is where $\tau \in o_t \setminus o_{t-1}$, use the same domain as round t does. Since our domains are constant within any epoch, we will simply skip outstanding observations at the start of each epoch, skipping at most $d_{\max}$ rounds whenever we change epochs. From here applying Lemma 3, bounding the bias of the estimator and bounding the total regret of skipped rounds by $H \cdot d_{\max} \cdot HSA \log(T)$ yields the desired result. The detailed proof can be found in Appendix D.

The first term in the regret matches the best known regret for adversarial MDPs even without delayed feedback (Jin et al., 2020). The second term matches the lower bound of Lancewicki et al. (2022a) up to logarithmic terms. This improves over the previous state-of-the-art regret bounds $\widetilde{O}\big(H\sqrt{SAT} + H(HSA)^{1/4}\sqrt{D}\big)$ by Jin et al. (2022) and $\widetilde{O}\big(H^3S\sqrt{AT} + H^3\sqrt{D}\big)$ by Lancewicki et al. (2023).

8. Experiments

In this section we are evaluating the performance of Algorithm 1 and Algorithm 2 on synthetic experiments. The full code for the experiments can be found here[1]

8.1 Experiments for combinatorial bandits

For the combinatorial bandit setting we split the time horizon of $T = 10000$ rounds into b blocks of length J and the algorithm only receives the feedback for all rounds in a block at the end of that block. As actions we use m-sets with $m = 3$ and $K = 10$, the losses of in dimension i are either fixed or oscillating. The fixed arms are always 0, the oscillating arms have a constant loss of -1 in block j if j is even and 0.9 otherwise. In other words, the oscillating arms are good arms and the constants are the bad arms. We use $m = 3$ oscillating and 7 fixed arms. As mentioned earlier, Algorithm 1 is the first algorithm for delayed adversarial combinatorial bandits. We will therefore compare Algorithm 1 to a standard algorithm not adapted to delay. Namely, an FTRL based version of the algorithm presented in Figure 3 of Audibert et al. (2014), which is the same as running Algorithm 1 with regularizer $R_t(\boldsymbol{v}) = \sum_{i=1}^{K} \frac{1}{\eta_t} \boldsymbol{v}(i) \log(\boldsymbol{v}(i))$ and learning rate $\eta_t = \sqrt{\frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{16Kt}}$.

The results of experiments for varying block sizes can be found in Figure 1. Dropping all other dependencies, and assuming a constant delay of $d_t = d$, our analysis finds that $\mathcal{R}_T \lesssim \frac{1}{\eta} + \eta T + \eta d T$. The delay unaware algorithms tunes $\eta \approx \frac{1}{\sqrt{T}}$, leading to a regret bound of $\mathcal{R}_T \lesssim \sqrt{T} + d\sqrt{T}$, which matches the roughly linear dependency on delay which we observe for the delay unaware algorithm. When the block size is $b = 1$, there is no delay present and the delay unaware method outperforms our algorithm slightly as we are over-regularising. But even for small delays, the delay aware tuning outperforms the non-delayed tuning significantly.

8.2 Experiments for linear bandits

In this section we present synthetic experiments for the linear bandit setting. The losses are generated using the same block structure as for the experiments for combinatorial semi-bandits, where the algorithm only observes feedback at the end of the block. There are $T = 10000$ rounds split into blocks, where the block size is $J \in \{30, 60, 90, 120, 150\}$. In each block the losses are either $(1/\sqrt{K}, \dots, 1/\sqrt{K})$ or $(-1/\sqrt{K}, \dots, -1/\sqrt{K})$. As in the combinatorial semi-bandit setting, the sign of the losses alternates between blocks. The dimension is varied between experiments, with $K \in [10, 20, 40]$. We implemented Algorithm 2, Algorithm 5, and Banker-BOLO Huang et al. (2023) with $-\log(1 - \|\boldsymbol{x}\|_2^2)$ as the $1-$self-concordant barrier for the unit ball. We also implemented a version of Banker-BOLO with what we believe to be improved tuning as described in Section 5. This version of Banker-BOLO is denoted by Banker-BOLO-V2. A fifth possible algorithm to compare with is the algorithm of Ito et al. (2020a). This is an instance of continuous exponential weigths, which means its computational complexity is $O(\text{poly}(K, T))$. However, the degree of this polynomial is high, which means that running this algorithm is infeasible for us.

The results can be found in Figure 2 in the main body, and Figures 3 and 4 in Appendix G. As predicted by theory, the regret grows with the square root of the block size for all algorithms.

1. `https://github.com/LukasZierahn/A-Unified-Analysis-of-Nonstochastic-Delayed-Feedback.git`

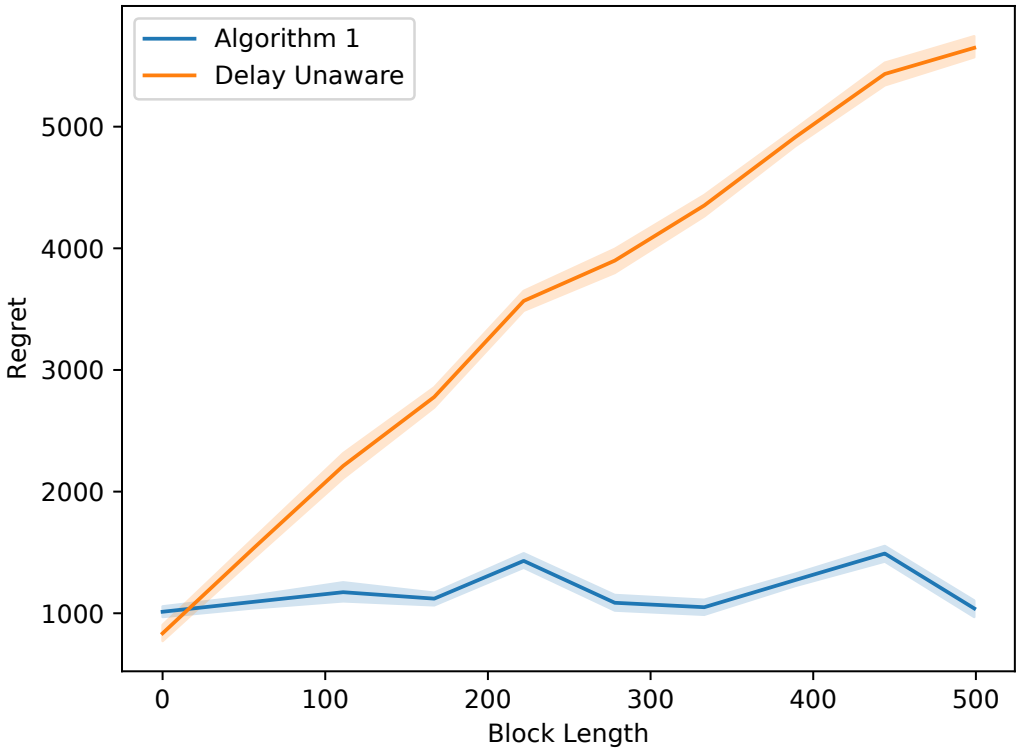

Figure 1: The results for our combinatorial semi-bandits experiments. The solid line is the mean regret in the 20 repetitions over $T = 10000$ rounds. The shaded areas are 95% confidence intervals.

However, it seems that the $K\sqrt{\nu}\sqrt{D}$ and $K^{3/2}\sqrt{D}$ terms in the regret bounds of Banker-BOLO-V2 and Banker-BOLO could possibly improved, as we do not see the difference in the regret between our algorithms and the Banker-BOLO algorithm increase as the dimension increases. This is to be expected, as one can probably derive a similar decomposition of the regret for OMD, upon which Banker-BOLO is based, as we did for FTRL. As with FTRL, this would most likely lead to $\sqrt{D}$ term in the regret bound for the cost of delay, given that the algorithm is appropriately tuned. We see that our algorithms consistently outperform both versions of Banker-BOLO. However, we believe that with the right tuning a version of OMD will perform similarly to our algorithms. It can be seen that Banker-BOLO-V2 has better performance than Banker-BOLO, which is predicted by theory. We also see that the performance of Algorithms 2 and 5 hardly differs, which is also predicted by theory.

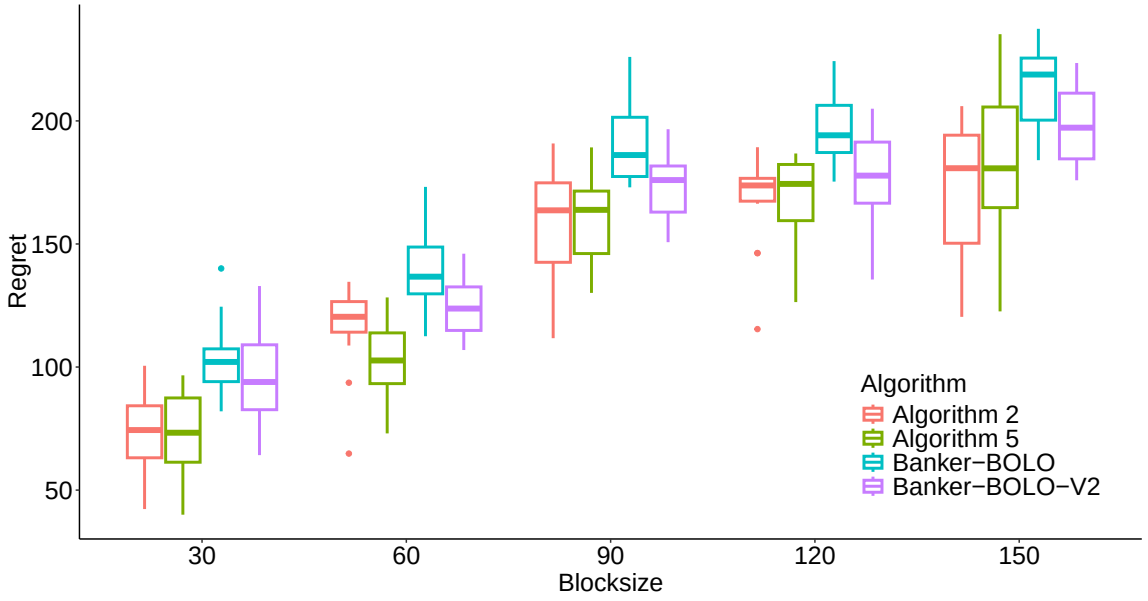

Figure 2: Boxplot of the regret in the linear bandit experiments with 20 repetitions over $T = 10000$ rounds with $K = 20$.

9. Conclusion

In sections 4, 6, and 7 we have shown that FTRL leads to optimal regret bounds under delayed feedback in combinatorial semi-bandits, MDPs with known transitions. For MDPs with unknown transitions we provide state-of-the-art results. Furthermore, in section 5 we have provided an efficient algorithm with nearly optimal regret for linear bandits. In section 8 we have shown that Algorithm 1 and Algorithm 2 outperform delay-unaware and previous algorithms respectively on our synthetic datasets.

Acknowledgments

The authors would like to thank András György for an insightful discussion about the regret decomposition and the suggestion to simplify the decomposition compared to the conference version of this paper. This work was done while LZ was at the Università degli Studi di Milano, Italy and partially done while DvdH was at the University of Milan supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and partially done while DvdH was at the University of Amsterdam supported by Netherlands Organization for Scientific Research (NWO), grant number VI.Vidi.192.095. NCB and LZ acknowledge the financial support from: the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), the FAIR project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme, the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS. 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).

Appendix A. Combinatorial Bandits

In this appendix we proof the main results of Section 4.

Theorem 8 (RESTATED) *Suppose that* $\max_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{a}\|_1 \le B$. *Algorithm 1 with*

$$\eta_t = \min\left\{\sqrt{\frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{16(B\sum_{\tau=1}^{t}|m_t|+Kt)}}, \frac{B^2\left(1+\log\left(\frac{K}{B}\right)\right)}{128K(Bd_{\max}+K)}\right\}, \qquad \gamma = \frac{1}{128\sqrt{B}d_{\max}},$$

guarantees that

$$\mathcal{R}_T \le 12\sqrt{B\left(1+\log\left(\frac{K}{B}\right)\right)(KT+BD)} + 128K^2 d_{\max} + 128\sqrt{B}d_{\max}K\log(T)\ .$$

Proof We start by verifying the conditions of Corollary 4 for R_t. Because we are not skipping rounds and have a constant actionset of $\mathcal{W} = \mathrm{Conv}(\mathcal{A})$, we have that Assumption 1 holds. Next, note that R_t as specified in (11) does not satisfy Assumption 2(c) because $R_t(\boldsymbol{v}) \le R_{t-1}(\boldsymbol{v})$. However, by using regularizer $\widetilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_t(\boldsymbol{v})$, and running the algorithm with this regularizer instead, we can see that Assumption 2(c) is satisfied and, crucially, the algorithm produces the same iterates. Note also that the gradients and Hessians of R_t and $\widetilde{R}_t$ are equivalent. We continue the analysis as if the algorithm is run with regularizer $\widetilde{R}_t$. By Lemma 26

$$\sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right) \le \frac{\sqrt{B}d_{\max}}{8\sqrt{K}}\,.$$

Thus, using Lemma 25 and plugging in γ gives,

$$\left(\nabla Rt(\boldsymbol{v}) - \nabla Rt + \delta(\boldsymbol{v})\right)^\top \boldsymbol{y} \le \sqrt{\gamma\frac{\sqrt{B}d_{\max}}{8}}\sqrt{\boldsymbol{y}^\top\nabla^2 Rt + \delta(\boldsymbol{v})\boldsymbol{y}} \le \frac{1}{32}\sqrt{\boldsymbol{y}^\top\nabla^2 Rt + \delta(\boldsymbol{v})\boldsymbol{y}}\,,$$

for all t, $\delta \in [d_{\max}]$, $\boldsymbol{v} \in \mathcal{W}$ and all $\boldsymbol{y} \in \mathbb{R}^K$, which verifies Assumption 2(b) for $\kappa = \gamma$. The fact that $4\nabla^2 R_t(\boldsymbol{v}) \succeq \nabla^2 R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathcal{W}$, $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$, and all t is also shown in Lemma 25, showing that Assumptions 2(a) holds for $\kappa = \gamma$.

As the next step, we bound $\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$. We use $\sum_{i=1}^K \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, i) \le B$, $\|\boldsymbol{\ell}\|_\infty \le 1$, and $\eta_t + \gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, i) \ge \gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, i)$ to show that

$$\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} = \sqrt{\sum_{i=1}^K \boldsymbol{\ell}_\tau(i)^2\frac{\eta_t\gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, i)^2}{\eta_t + \gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, i)}} \le \underbrace{\sqrt{\eta_t B}}_{\alpha_t}\,. \tag{14}$$

Next we bound $\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right]$. By the tower rule we have

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right] = \mathbb{E}_{\mathcal{F}_\tau}\left[\mathbb{E}_{\boldsymbol{a}_\tau}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}|\mathcal{F}_\tau\right]\right]$$

where $\mathcal{F}_\tau$ is a filtration over all random events observed by the learner as defined in Section 2. Let us consider $\mathbb{E}_{\boldsymbol{a}_\tau}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}|\mathcal{F}_\tau\right]$ in isolation:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}_\tau}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}|\mathcal{F}_\tau\right] &= \mathbb{E}_{\boldsymbol{a}_\tau}\left[\sum_{i=1}^K\left(\frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)}\right)^2\left(\nabla^2 R_t(\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau))\right)^{-1}(i,i)|\mathcal{F}_\tau\right]\\ &= \sum_{i=1}^K\frac{\boldsymbol{\ell}_\tau(i)^2}{\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)}\frac{\eta_t\gamma\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)^2}{\eta_t+\gamma\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)} \le \underbrace{\eta_t K}_{\beta_t^2}\,, \end{aligned} \tag{15}$$

where we used that $\boldsymbol{a}_\tau^2=\boldsymbol{a}_\tau$, $\mathbb{E}_{\boldsymbol{a}_\tau}[\boldsymbol{a}_\tau]=\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)$, and $\eta_t+\gamma\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)\ge\gamma\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)$. We now bound $\kappa\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$:

$$\kappa\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{L}_t)} = \kappa\sqrt{\sum_{i=1}^K\left(\frac{\boldsymbol{a}_t(i)\boldsymbol{\ell}_t(i)}{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,i)}\right)^2\frac{\eta_{t'}\gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,i)^2}{\eta_{t'}+\gamma\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,i)}} \le \sqrt{\kappa\gamma B} \le \frac{1}{128 d_{\max}}\,,$$

where we used that $\kappa=\gamma=\frac{1}{128\sqrt{B}d_{\max}}$.

The last requirement is to 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 $\widehat{\boldsymbol{\ell}}_\tau(i)=\frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,i)}$, 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 have been used in round τ to compute $\boldsymbol{p}_\tau$ (and vice versa) we have that $\widehat{\boldsymbol{\ell}}_{\tau'}$ and $\widehat{\boldsymbol{\ell}}_\tau$ are independent. We conclude that

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

where we used that $\widehat{\boldsymbol{\ell}}_{\tau'}$ is and unbiased estimator of $\boldsymbol{\ell}_{\tau'}$.

We are now in a position to apply Corollary 4. By using α_t from equation (14) and β_t^2 from equation (15) it follows that for any $\boldsymbol{u}\in\mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^T(\boldsymbol{w}_t-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right] \le R_T(\boldsymbol{u})+16\sum_{t=1}^T\beta_t^2+16\sum_{t=1}^T\alpha_t^2|m_t|\,. \tag{16}$$

Next we want to bound $R_T(\boldsymbol{u})$, however the negative logarithm component 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 $\mathcal{W}$ and instead we will compare to $\boldsymbol{u}=\arg\min_{\boldsymbol{v}\in\widetilde{\mathcal{W}}}\sum_{t=1}^T\boldsymbol{\ell}_t^\top\boldsymbol{v}$ where $\widetilde{\mathcal{W}}=\mathcal{W}\cap\{\boldsymbol{x}\in\mathbb{R}_+:\forall i\in[K]\ \boldsymbol{x}(i)\ge\theta\}$ is a shrunken actionset. 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. More specifically we can write $\boldsymbol{a}^*=\boldsymbol{u}+\theta\xi$, for some ξ with $\|\xi\|_\infty\le 1$.

By Lemma 25, we have

$$R_T(\boldsymbol{u}) \le \frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{\eta_T}+\frac{K\log\left(\frac{1}{\theta}\right)}{\gamma}\,. \tag{17}$$

To finish the proof, we start from the regret

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

where we bound $\xi^\top \boldsymbol{\ell}_t \leq K$ in the inequality. We continue by using (16)

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] &\leq R_T(\boldsymbol{u}) + 16\sum_{t=1}^{T}\kappa\beta_t^2 + 16\sum_{t=1}^{T}\kappa\alpha_t^2|m_t| \\ &\leq \frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{\eta_T} + \frac{K\log(T)}{\gamma} + 16K\sum_{t=1}^{T}\eta_t + 16B\sum_{t=1}^{T}\eta_t|m_t| \\ &\leq \frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{\eta_T} + \frac{K\log(T)}{\gamma} + \\ &\quad + 8\sqrt{B\left(1+\log\left(\frac{K}{B}\right)\right)(KT+BD)}\,,\end{aligned}$$

where we first substituted in (17), with $\theta = \frac{1}{T}$, $\alpha_t^2 = \eta_t B$ and $\beta_t^2 = \eta_t K$, and applied Lemma 24. Using the last two inequalities and substituting

$$\begin{aligned}\eta_t &= \min\left\{\sqrt{\frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{16(B\sum_{\tau=1}^{t}|m_t| + Kt)}}, \frac{B^2\left(1+\log\left(\frac{K}{B}\right)\right)}{128K(Bd_{\max}+K)}\right\} \\ \gamma &= \frac{1}{128\sqrt{B}d_{\max}}\end{aligned}$$

and doing some simplifications yields

$$\mathcal{R}_T \leq 12\sqrt{B\left(1+\log\left(\frac{K}{B}\right)\right)(KT+BD)} + 128K^2 d_{\max} + 128\sqrt{B}d_{\max}K\log(T)$$

concluding the proof. ■

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 8 is optimal. The proof of our lower bound follows from standard arguments in the delayed bandit feedback literature.

Theorem 9 (RESTATED) *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*

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

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 B. Linear Bandits

Recall that a thrice-differentiable function Ψ is called self-concordant if it is convex and satisfies $|\nabla^3\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h},\boldsymbol{h}]| \le 2\big(\nabla^2\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h}]\big)^{3/2}$, where $\nabla^3\Psi(\boldsymbol{v})[\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3] = \frac{\partial^3}{\partial t_1\partial t_2\partial t_3}|_{t_1=t_2=t_3=0}\Psi(\boldsymbol{v}+t_1\boldsymbol{h}_1+t_2\boldsymbol{h}_2+t_3\boldsymbol{h}_3)$. A self-concordant function Ψ is a ν-self-concordant barrier if $|\nabla\Psi(\boldsymbol{v})[\boldsymbol{h}]| \le \sqrt{\nu\nabla^2\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h}]}$. The following property allows us to satisfy the stability condition of the Hessian in Assumption 2(a): for $\boldsymbol{v},\boldsymbol{v}' \in \mathcal{W}$, if $\|\boldsymbol{v}-\boldsymbol{v}'\|^{\star}_{\Psi,\boldsymbol{v}} < 1$, then

$$\big(1-\|\boldsymbol{v}-\boldsymbol{v}'\|^{\star}_{\Psi,\boldsymbol{v}}\big)^2\nabla^2\Psi(\boldsymbol{v}) \preceq \Psi(\boldsymbol{v}') \preceq \big(1-\|\boldsymbol{v}-\boldsymbol{v}'\|^{\star}_{\Psi,\boldsymbol{v}}\big)^{-2}\nabla^2\Psi(\boldsymbol{v})\,. \tag{18}$$

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

$$\Psi(\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathcal{W}}\Psi(\boldsymbol{v}) \le \nu\ln\big((1+\delta)\delta^{-1}\big)\,. \tag{19}$$

This property allows us to show that for any benchmark point $\widetilde{\boldsymbol{u}} \in \mathcal{W}_\delta$, $R_T(\boldsymbol{u})$ and is nicely bounded.

Theorem 10 (RESTATED) *Suppose that $T > 100$ and $B \ge 1$. Algorithm 2, run with a ν-self-concordant barrier Ψ and with*

$$\gamma_t = \min\left\{\frac{1}{256BKd_{\max}}, \sqrt{\frac{\nu\log\big(1+\sqrt{T}\big)}{16B^2K^2t}}\right\}$$
$$\eta_t = \min\left\{\frac{B}{256d_{\max}}, \sqrt{\frac{B^2}{16\sum_{\tau=1}^{t}|m_t|}}\right\},$$

guarantees that, for any $\boldsymbol{u} \in \mathcal{W}$,

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right] \le 12BK\sqrt{\nu T\log(1+\sqrt{T})} + 12B\sqrt{D} + 2B\sqrt{T} + 512BKd_{\max}\nu\log(1+\sqrt{T})\,.$$

Proof We start by verifying the assumptions of Corollary 4. Because we are not skipping rounds and have a constant actionset of $\mathcal{W} = \operatorname{Conv}(\mathcal{A})$, we have that Assumption 1 holds. Using $\mathbb{E}[\boldsymbol{v}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{v}_t \boldsymbol{v}_t^\top] = \frac{1}{K}\boldsymbol{I}$ we see that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t] = \boldsymbol{\ell}_t$. For $\tau \le t$, 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,

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

where we used that $\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}|\mathcal{F}_t] = \mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t, \widehat{\boldsymbol{\ell}}_\tau] = \boldsymbol{\ell}_t$.

We now turn to verifying that Assumption 2 holds. Assumption 2(c) holds by definition of η_t and γ_t. Because $\widetilde{\Psi}$ is a self-concordant, if we choose $\kappa = \frac{1}{256BKd_{\max}}$, $\frac{\kappa}{\gamma_t}\widetilde{\Psi}$ is also self-concordant as self-concordance is preserved by scaling of factors exceeding one. Since $c\|\boldsymbol{v}\|_2^2$ is self-concordant on $\mathbb{R}^d$ for any $c > 0$ and adding two self-concordant barriers yields a self-concordant barrier, κR_t is also a self-concordant barrier. If $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$ it implies that $\boldsymbol{v}' \in \mathcal{D}_{\kappa R_t}(\boldsymbol{v}, \frac{1}{2})$. By equation (18), for $\boldsymbol{v}' \in \mathcal{D}_{\kappa R_t}(\boldsymbol{v}, \frac{1}{2})$, we have $4\nabla^2 \kappa R_t(\boldsymbol{v}) \succeq \nabla^2 \kappa R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 \kappa R_t(\boldsymbol{v})$ or equivalently, for all $\boldsymbol{v} \in \mathcal{W}$ and $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\kappa}})$ we have $4\nabla^2 R_t(\boldsymbol{v}) \succeq \nabla^2 R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{v})$, which verifies Assumption 2(a).

The final condition to check is that $\kappa\left(\nabla R_t(\boldsymbol{v}) - \nabla R_{t+\delta}(\boldsymbol{v})\right)^\top y \le \frac{1}{32}\sqrt{\kappa y^\top \nabla^2 R_{t+\delta}(\boldsymbol{v}) y}$. Let $\boldsymbol{v} \in \mathcal{W}$, $\boldsymbol{y} \in \mathbb{R}^K$, and $\delta \in [d_{\max}]$, then

$$\begin{aligned}
&\left(\nabla R_t(\boldsymbol{v}) - \nabla R_{t+\delta}(\boldsymbol{v})\right)^\top \boldsymbol{y} \\
&\quad = \kappa\left(\frac{2}{\eta_t}\boldsymbol{v} + \frac{1}{\gamma_t}\nabla\Psi(\boldsymbol{v}) - \frac{2}{\eta_{t+\delta}}\boldsymbol{v} - \frac{1}{\gamma_{t+\delta}}\nabla\Psi(\boldsymbol{v})\right)^\top \boldsymbol{y} \\
&\quad = 2\sum_{i=1}^K \kappa\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right)\boldsymbol{v}(i)\boldsymbol{y}(i) + \kappa\left(\frac{1}{\gamma_{t+\delta}} - \frac{1}{\gamma_t}\right)\left(\nabla\Psi(\boldsymbol{v})\right)^\top \boldsymbol{y}\,.
\end{aligned}$$

By using the Cauchy-Schwarz inequality and the fact that $\mathcal{W} \subseteq \mathcal{B}(B)$ we can see that

$$\begin{aligned}
\sum_{i=1}^K \kappa\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right)\boldsymbol{v}(i)\boldsymbol{y}(i) &\le \kappa\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right) B\sqrt{\sum_{i=1}^K \boldsymbol{y}(i)^2} \\
&\le \kappa\sqrt{16\sum_{\tau=t}^{t+\delta}|m_\tau|}\sqrt{\sum_{i=1}^K \boldsymbol{y}(i)^2} \\
&\le 4\kappa d_{\max}\sqrt{\sum_{i=1}^K \boldsymbol{y}(i)^2} \\
&= \sqrt{\kappa\eta_{t+\delta}}4d_{\max}\sqrt{\frac{\kappa}{\eta_{t+\delta}}\sum_{i=1}^K \boldsymbol{y}(i)^2} \\
&\le \frac{1}{64}\sqrt{\frac{\kappa}{\eta_{t+\delta}}\sum_{i=1}^K \boldsymbol{y}(i)^2}\,.
\end{aligned}$$

Similarly, since Ψ is a ν-self-concordant barrier and using that $\log(1+\sqrt{T}) > 1$ by assumption on T,

$$\begin{aligned}\kappa\left(\frac{1}{\gamma_{t+\delta}}-\frac{1}{\gamma_t}\right)(\nabla\Psi(\boldsymbol{v}))^\top \boldsymbol{y} &\le \kappa\sqrt{\nu}\left(\frac{1}{\gamma_{t+\delta}}-\frac{1}{\gamma_t}\right)\sqrt{\boldsymbol{y}^\top\nabla^2\Psi(\boldsymbol{v})\boldsymbol{y}}\\ &\le \kappa\sqrt{16B^2K^2d_{\max}}\sqrt{\boldsymbol{y}^\top\nabla^2\Psi(\boldsymbol{v})\boldsymbol{y}}\\ &= \sqrt{\kappa\gamma_{t+\delta}16B^2K^2d_{\max}}\sqrt{\frac{\kappa}{\gamma_{t+\delta}}\boldsymbol{y}^\top\nabla^2\Psi(\boldsymbol{v})\boldsymbol{y}}\\ &\le \frac{1}{32\sqrt{2}}\sqrt{\frac{\kappa}{\gamma_{t+\delta}}\boldsymbol{y}^\top\nabla^2\Psi(\boldsymbol{v})\boldsymbol{y}}\,.\end{aligned}$$

By using the above two inequalities and $\sqrt{a}+\sqrt{b}\le\sqrt{2}\sqrt{a+b}$ we can see that

$$\begin{aligned}\left(\nabla R_t(\boldsymbol{v})-\nabla R_{t+\delta}(\boldsymbol{v})\right)^\top\boldsymbol{y} &\le \frac{1}{32}\sqrt{\frac{\kappa}{\eta_{t+\delta}}2\sum_{i=1}^K\boldsymbol{y}(i)^2+\frac{\kappa}{\gamma_{t+\delta}}\boldsymbol{y}^\top\nabla^2\Psi(\boldsymbol{v})\boldsymbol{y}}\\ &= \frac{1}{32}\sqrt{\kappa y^\top\nabla^2 R_{t+\delta}(\boldsymbol{v})y}\end{aligned}$$

Next, pick any $t'\in[T]$ and observe that because $\nabla^2 R_{t'}(\boldsymbol{v})\succeq\frac{1}{\gamma_{t'}}\nabla^2\Psi(\boldsymbol{v})$ we have that

$$\kappa\|\widehat{\boldsymbol{\ell}}_t\|^2_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\le\kappa\gamma_{t'}K^2(\boldsymbol{\ell}_t^\top\boldsymbol{a}_t)^2\boldsymbol{v}_t^\top\boldsymbol{v}_t=\kappa\gamma_{t'}K^2(\boldsymbol{\ell}_t^\top\boldsymbol{a}_t)^2\,.$$

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

$$\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\le\sqrt{\kappa}\underbrace{\sqrt{\gamma_{t'}B^2K^2}}_{\beta_{t'}}\le\frac{1}{128d_{\max}}\,,$$

where the last inequality is because $\gamma_{t'}\le\frac{1}{128BKd_{\max}}$. Let $\tau\in m_t\cup\{t\}$, because $\nabla^2R_t(\boldsymbol{v})\succeq\frac{\kappa}{\eta_t}\boldsymbol{I}$ and $\|\boldsymbol{\ell}_\tau\|_2\le 1$ we have that

$$\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\le\underbrace{\sqrt{\eta_t}}_{\alpha_t}\,.$$

We have fulfilled all requirements for Corollary 4.

Let $\widetilde{\boldsymbol{u}}=\frac{\boldsymbol{u}-\boldsymbol{v}^+}{1+\frac{1}{\sqrt{T}}}+\boldsymbol{v}^+\in\mathcal{W}_{\frac{1}{\sqrt{T}}}$, with $\boldsymbol{v}^+=\arg\min_{\boldsymbol{v}\in\mathcal{W}}R_1(\boldsymbol{v})$. By equation 19 and $\mathcal{W}\subseteq\mathcal{B}(B)$ we have that

$$\begin{aligned}R_T(\widetilde{\boldsymbol{u}})-R_1(\boldsymbol{v}^+) &\le \frac{B^2}{\eta_T}+\frac{1}{\gamma_T}\widetilde{\Psi}(\widetilde{\boldsymbol{u}})-\frac{1}{\gamma_1}\widetilde{\Psi}(\boldsymbol{v}^+)\\ &\le \frac{B^2}{\eta_T}+\frac{\nu}{\gamma_T}\log\left(\frac{1+\xi}{\xi}\right),\end{aligned}\tag{20}$$

where we used that $\widetilde{\Psi}(\boldsymbol{v}^+)\ge 0$. Furthermore, by using that $\mathcal{W}\in\mathcal{B}(B)$ and $\|\boldsymbol{\ell}_t\|_2\le 1$, we have that

$$\sum_{t=1}^T(\widetilde{\boldsymbol{u}}-\boldsymbol{u})^\top\boldsymbol{\ell}_t=\sum_{t=1}^T\left(1-\frac{1}{1+\frac{1}{\sqrt{T}}}\right)(\widetilde{\boldsymbol{u}}-\boldsymbol{v}^+)^\top\boldsymbol{\ell}_t\le 2TB\left(\frac{\frac{1}{\sqrt{T}}}{1+\frac{1}{\sqrt{T}}}\right)\le 2B\sqrt{T}\,.$$

Algorithm 5: Efficient implementation of delayed FTRL for linear bandits

Input: ν-self concordant barrier Ψ for $\mathcal{W}$, hyperparameters η, γ.
Set $\boldsymbol{z}_1 = \arg\min_{\boldsymbol{v}} \Psi_1(\boldsymbol{v})$
for $t = 1, ..., T$ **do**
  Observe $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$ for $\tau \in o_t \setminus o_{t-1}$.
  Find loss estimators $\widehat{\boldsymbol{\ell}}_\tau = K \boldsymbol{\ell}_\tau^\top \boldsymbol{a}_\tau \big(\nabla^2 \Psi(\boldsymbol{z}_\tau)\big)^{1/2} \boldsymbol{v}_\tau$ for new observations $\tau \in o_t \setminus o_{t-1}$.
  Compute $\boldsymbol{z}_t = DN(\Psi_{t-1}, \boldsymbol{z}_{t-1})$
  Play $\boldsymbol{a}_t = \boldsymbol{z}_t + \big(\nabla^2 \Psi(\boldsymbol{z}_t)\big)^{-1/2} \boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is uniformly sampled from the unit sphere.
end for

Thus, by Corollary 4 with $\alpha_t^2 = \eta_t$ and $\beta_t^2 = \gamma_t B^2 K^2$ we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] &\le \mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{a}_t - \widetilde{\boldsymbol{u}})^\top \boldsymbol{\ell}_t\right] + 2B\sqrt{T} \\
&\le R_T(\boldsymbol{u}) - R_1(\boldsymbol{v}^+) + 16B^2K^2 \sum_{t=1}^T \gamma_t + 16 \sum_{t=1}^T \eta_t |m_t| + 2B\sqrt{T} \\
&\le \frac{B^2}{\eta_T} + \frac{\nu}{\gamma_T} \log(1+\sqrt{T}) + 8BK\sqrt{\nu T \log(1+\sqrt{T})} + 8B\sqrt{D} + 2B\sqrt{T} \\
&\le 512 BK d_{\max} \nu \log(1+\sqrt{T}) + 12BK\sqrt{\nu T \log(1+\sqrt{T})} + 12B\sqrt{D} + 2B\sqrt{T}\,,
\end{aligned}$$

where in the second inequality we used $\sum_{t=1}^T \gamma_t \le \frac{\sqrt{\nu T \log(1+\sqrt{T})}}{2BK}$ and $\sum_{t=1}^T \eta_t |m_t| \le \frac{1}{2} B\sqrt{D}$, both of which follow from Lemma 24, and the last inequality follows from simplifications. ■

B.1 Efficient Implementation

In this section we will use fixed learning rates $\eta_t = \eta$ and $\gamma_t = \gamma$ for simplicity. Define

$$\begin{aligned}
\Phi_t(\boldsymbol{v}) &= \gamma \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + \gamma R(\boldsymbol{v}) = \gamma \widehat{\boldsymbol{L}}_t^\top \boldsymbol{v} + \frac{\gamma}{\eta} \|\boldsymbol{v}\|_2^2 + \widetilde{\Phi}(\boldsymbol{v}) \\
e(\Phi_t, \boldsymbol{v}) &= -\big(\nabla^2 \Phi_t(\boldsymbol{v})\big)^{-1} \nabla \Phi_t(\boldsymbol{v}) \\
\lambda(\Phi_t, \boldsymbol{v}) &= \sqrt{\nabla \Phi_t(\boldsymbol{v})^\top \big(\nabla^2 \Phi_t(\boldsymbol{v})\big)^{-1} \nabla \Phi_t(\boldsymbol{v})} \\
DN(\Phi_t, \boldsymbol{v}) &= \boldsymbol{v} - \frac{1}{1+\lambda(\Phi_t, \boldsymbol{v})} e(\Phi_t, \boldsymbol{v}) \\
\boldsymbol{z}_t^\star &= \arg\min_{\boldsymbol{v}} \Phi_{t-1}(\boldsymbol{v})\,.
\end{aligned}$$

The following facts can be found in Nemirovski and Todd (2008)

$$\lambda(\Phi_t, DN(\Phi_t, \boldsymbol{v})) \le 2\lambda(\Phi_t, \boldsymbol{v})^2 \tag{21}$$

$$\|\boldsymbol{v} - \boldsymbol{z}_t^\star\|_{\Phi_{t-1}, \boldsymbol{z}_t^\star} \le \frac{\lambda(\Phi_{t-1}, \boldsymbol{v})}{1 - 2\lambda(\Phi_{t-1}, \boldsymbol{v})} \qquad \text{if } \lambda(\Phi_{t-1}, \boldsymbol{v}) < \tfrac{1}{2} \tag{22}$$

Algorithm 5 is a simple modification of Algorithm 2 in section 9 of Abernethy et al. (2008). Abernethy et al. (2008) show that in the non-delayed setting, given the previous iterate, it takes essentially one iteration of the damped Newton method to compute $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$. If an easily computed self-concordant barrier is available, the computational complexity of the damped Newton method is $O(K^2)$. Since we can compute $(\nabla^2 R_t(\boldsymbol{z}_t))^{1/2}$ and its inverse in $O(K^3)$ time by means of an eigenvalue decomposition, the total runtime is $O(K^3)$. In what follows we provide a modification of Lemma 7 by Abernethy et al. (2008) to the delayed setting. In what follows we will show that $\boldsymbol{z}_t^\star$ is close to $\boldsymbol{z}_t$ as measured in local distance. While this may seem arbitrary, we have that $\boldsymbol{z}_t^\star = \arg\min_{\boldsymbol{v}} \Phi_t(\boldsymbol{v}) = \arg\min_{\boldsymbol{v}} \frac{1}{\gamma}\Phi_t(\boldsymbol{v})$, which in turn is the FTRL objective we have been working with throughout this paper. Thus, showing that $\boldsymbol{z}_t^\star$ is close to $\boldsymbol{z}_t$ implies that $\boldsymbol{z}_t$ will have a similar regret bound as we would obtain from $\boldsymbol{z}_t^\star$, as argued by Abernethy et al. (2008). With Lemma 15 in hand, one can follow the steps provided by Abernethy et al. (2008) to see that the regret of Algorithm 5 is of the same order as that of Algorithm 2.

Lemma 15 *Suppose that $\eta_t = \eta > 0$ and $\gamma_t = \gamma \leq \frac{1}{162K^2 d_{\max}}$ It holds that for all t*

$$\lambda(\Phi_t, \boldsymbol{z}_t)^2 \leq 9\gamma^2 K^2 d_{\max} \qquad \textit{and} \qquad \|\boldsymbol{z}_t - \boldsymbol{z}_t^\star\|_{\Phi_{t-1}, \boldsymbol{z}_t^\star} \leq 648\gamma^2 K^2 d_{\max}\,.$$

Proof The proof is by induction on t. The base case holds by definition. Suppose the statement holds for $t-1$. Using $(\boldsymbol{x}+\boldsymbol{y})^\top A(\boldsymbol{x}+\boldsymbol{y}) \leq 2\boldsymbol{x}^\top A\boldsymbol{x} + 2\boldsymbol{y}^\top A\boldsymbol{y}$ we get that

$$\begin{aligned}
\lambda(\Phi_t, \boldsymbol{z}_t)^2 &= \nabla\Phi_t(\boldsymbol{z}_t)^\top \left(\nabla^2\Phi_t(\boldsymbol{z}_t)\right)^{-1} \nabla\Phi_t(\boldsymbol{z}_t) \\
&= \left(\nabla\Phi_{t-1}(\boldsymbol{z}_t) + \gamma \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau\right)^\top \left(\frac{\gamma}{\eta}\boldsymbol{I} + \nabla^2\Psi(\boldsymbol{z}_t)\right)^{-1} \left(\nabla\Phi_{t-1}(\boldsymbol{z}_t) + \gamma \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau\right) \\
&\leq 2\gamma^2 \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau^\top \left(\frac{\gamma}{\eta}\boldsymbol{I} + \nabla^2\Psi(\boldsymbol{z}_t)\right)^{-1} \widehat{\boldsymbol{\ell}}_\tau \\
&\quad + 2\underbrace{\nabla\Phi_{t-1}(\boldsymbol{z}_t)^\top \left(\nabla^2\Psi_{t-1}(\boldsymbol{z}_t)\right)^{-1} \nabla\Phi_{t-1}(\boldsymbol{z}_t)}_{\lambda(\Phi_{t-1}, \boldsymbol{z}_t)^2}\,.
\end{aligned}$$

Now, with a minor modification of Lemma 7 we can see that $\boldsymbol{z}_t^\star \in \mathcal{D}_{R_t}(\boldsymbol{z}_{t-\delta}^\star, \frac{1}{2})$ and $\boldsymbol{z}_t^\star \in \mathcal{D}_{R_t}(\boldsymbol{z}_{t-\delta}, \frac{1}{2})$ for all $\delta \in [\min\{d_{\max}, T-t\}]$. In turn, this implies that

$$\begin{aligned}
&2\gamma^2 \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau^\top \left(\frac{\gamma}{\eta}\boldsymbol{I} + \nabla^2\Psi(\boldsymbol{z}_t)\right)^{-1} \widehat{\boldsymbol{\ell}}_\tau \\
&\leq 2\gamma^2 \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau^\top \left(\nabla^2\Psi(\boldsymbol{z}_t)\right)^{-1} \widehat{\boldsymbol{\ell}}_\tau \\
&\leq 8\gamma^2 \sum_{\tau\in m_t} \widehat{\boldsymbol{\ell}}_\tau^\top \left(\nabla^2\Psi(\boldsymbol{z}_\tau)\right)^{-1} \widehat{\boldsymbol{\ell}}_\tau && \text{(equation (18))} \\
&\leq 8\gamma^2 K^2 |m_t| \leq 8\gamma^2 K^2 d_{\max}\,. && \text{(by def. of } \widehat{\boldsymbol{\ell}}_\tau)
\end{aligned}$$

By equation (21) and the induction assumption we have that

$$\lambda(\Phi_{t-1}, \boldsymbol{z}_t)^2 \leq 2\lambda(\Phi_{t-1}, \boldsymbol{z}_{t-1})^4 \leq 162\gamma^4 K^4 d_{\max}^2\,. \tag{23}$$

Thus, we can apply the assumption $\gamma^2 \leq \frac{1}{162K^2 d_{\max}}$ to find that

$$\lambda(\Phi_t, \boldsymbol{z}_t)^2 \leq 8\gamma^2 K^2 d_{\max} + 162\gamma^4 K^4 d_{\max}^2 \leq 9\gamma^2 K d_{\max} ,$$

after which we have proven the induction step for the first claim. For the second claim, we start with equation (22) and then the fact that $\lambda(\Phi_{t-1}, \boldsymbol{z}_t)^2 \leq \frac{1}{16}$ which follows by equation (23) and the assumption that $\gamma^2 \leq \frac{1}{162K^2 d_{\max}}$, then using equation (21) and finally applying the first claim yields

$$\|\boldsymbol{z}_t - \boldsymbol{z}_t^\star\|_{\Phi_{t-1}, \boldsymbol{z}_t^\star} \leq \frac{\lambda(\Phi_{t-1}, \boldsymbol{z}_t)}{1 - 2\lambda(\Phi_{t-1}, \boldsymbol{z}_t)} \leq 2\lambda(\Phi_{t-1}, \boldsymbol{z}_t) \leq 4\lambda(\Phi_{t-1}, \boldsymbol{z}_{t-1})^2 \leq 648\gamma^2 K^2 d_{\max} .$$

■

Appendix C. Adversarial Markov Decision Processes (MDPs)

Lemma 11 (RESTATED) *$\mathcal{W}$ satisfies the following:*

1. *For any $\boldsymbol{q} \in \Delta(\mathcal{M})$, there exists $\widetilde{\boldsymbol{q}} \in \mathcal{W}$ such that $\min_{h,s,a,s'} \widetilde{\boldsymbol{q}}(h,s,a,s') \geq \frac{1}{T^3 H^2 S^4 A^2}$ and $\|\boldsymbol{q} - \widetilde{\boldsymbol{q}}\|_1 \leq \frac{2H}{T}$.*
2. *Given $\boldsymbol{v} \in \mathcal{W}$, let π be defined by $\pi(a \mid h, s) = \frac{\boldsymbol{v}(h,s,a)}{\boldsymbol{v}(h,s)}$ and $\boldsymbol{q}^{max}(h,s,a) = \max_{\widehat{p} \in \mathcal{P}} \boldsymbol{q}^{\pi, \widehat{p}}(h,s,a)$. Then, $\|\boldsymbol{q}^\pi - \boldsymbol{v}\|_1 \leq \frac{2H}{T}$ and $\|\boldsymbol{q}^{max} - \boldsymbol{v}\|_1 \leq \frac{4H^2 S}{T}$.*

Proof We start by proving the first claim. Define $\widetilde{p} : [H] \times \mathcal{S} \times \mathfrak{A} \to \Delta_{\mathcal{S}}$ by $\widetilde{p}(s' \mid h,s,a) = (1 - \frac{1}{THSA})p(s' \mid h,s,a) + \frac{1}{THS^2 A}$ and notice that $\widetilde{p} \in \mathcal{P}$ since $|p(s' \mid h,s,a) - \widetilde{p}(s' \mid h,s,a)| \leq \frac{1}{THSA}$. Next, let π_u be the uniformly random policy, and define $\widetilde{\boldsymbol{q}} = (1 - \frac{1}{T})\boldsymbol{q} + \frac{1}{T}\boldsymbol{q}^{\pi_u, \widetilde{p}}$. It holds that $\widetilde{\boldsymbol{q}} \in \mathcal{W}$ because $\mathcal{W}$ is a convex set. Moreover, notice that $\boldsymbol{q}^{\pi_u, \widetilde{p}}(h,s,a,s') \geq \frac{1}{(THS^2 A)^2 A}$ which implies that $\widetilde{\boldsymbol{q}}(h,s,a,s') \geq \frac{1}{T^3 H^2 S^4 A^2}$. Finally,

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

Now we prove the second claim. Define loss function $\widetilde{\boldsymbol{\ell}}(h,s,a) = \operatorname{sign}\big(\boldsymbol{q}^\pi(h,s,a) - \boldsymbol{v}(h,s,a)\big)$ and note that $\|\boldsymbol{q}^\pi - \boldsymbol{v}\|_1 = V^{\pi, p, \widetilde{\boldsymbol{\ell}}}(1, s_{\text{init}}) - V^{\pi, \widehat{p}, \widetilde{\boldsymbol{\ell}}}(1, s_{\text{init}})$ for some $\widehat{p} \in \mathcal{P}$. Combining the value difference lemma (see, e.g, Shani et al. (2020)) with $\|p - \widehat{p}\|_\infty \leq \frac{1}{THSA}$ proves that $\|\boldsymbol{q}^\pi - \boldsymbol{v}\|_1 \leq \frac{2H}{T}$. 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, $\|\boldsymbol{u} - \boldsymbol{v}\|_1 \leq \sum_{h,s} \|\boldsymbol{q}^{\pi, \widehat{p}^{h,s}} - \boldsymbol{v}\|_1 \leq \frac{4H^2 S}{T}$. ■

Theorem 12 (RESTATED) *Suppose that $T \geq H$. Algorithm 3 with*

$$\gamma = \frac{1}{128 H d_{\max}} \qquad \eta_t = \min\left\{ \frac{\log(SA)}{96HSA\sqrt{SAd_{\max} + d_{\max}^2}}, \frac{\sqrt{\log(SA)}}{\sqrt{SAt + \sum_{\tau=1}^{t} |m_t|}} \right\}$$

guarantees

$$\mathbb{E}[\mathcal{R}_T] \leq 72H\sqrt{\log(SA)(TSA + D)} + 1338 d_{\max} H^2 S^2 A^2 \log\left(HSAT\right) .$$

Proof As in the proof of Theorem 8, the regularizer R_t as specified in (12) does not satisfy Assumption 2(c) because we can have $R_t(\boldsymbol{v}) \leq R_{t-1}(\boldsymbol{v})$, but, as argued in the proof of Theorem 8, we can overcome this issue in a relative straightforward manner via the regularizer $\widetilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}' \in \mathcal{W}} R_t(\boldsymbol{v}')$, which has no impact on the iterates. We continue by decomposing the regret as

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

where, by the second property in Lemma 11, ERROR is bounded by $2H$; and by the first property $\widetilde{\boldsymbol{u}} \in \mathcal{W}$ exists such that both SHIFT-PENALTY is bounded by $2H$ and $\min_{h,s,a,s'} \widetilde{\boldsymbol{u}}(h,s,a,s') \geq \frac{1}{T^3 H^2 S^4 A^2}$, which we will choose as our comparator to ensure that the regularisation is always bounded. For REG we use Lemma 3 with $\kappa = \gamma$. The fact that $4\nabla^2 R_t(\boldsymbol{v}) \succeq \nabla^2 R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathcal{W}$ and $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\gamma}})$ and all t comes directly from Lemma 25.

For $\delta \in d_{\max}$ we have that

$$\sqrt{\eta_{t+\delta}}\Big(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\Big) \leq \sqrt{\eta_{t+\delta}} \frac{\sqrt{SAd_{\max} + d_{\max}^2}}{\sqrt{\log(SA)}} \leq \frac{1}{\sqrt{32HSA}} (SAd_{\max} + d_{\max}^2)^{1/4} .$$

Thus, by definition of γ, Lemma 25 also implies that that

$$\gamma \left(\nabla R_t(\boldsymbol{v}) - \nabla R_{t+\delta}(\boldsymbol{v})\right)^\top y \leq \frac{1}{32}\sqrt{\gamma y^\top \nabla^2 R_{t+\delta}(\boldsymbol{v}) y} .$$

Next pick any $t \in [T], \tau \in m_t$, then

$$\|\boldsymbol{\ell}_\tau\|_{R_t, \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \leq \sqrt{\eta_t \sum_{h,s,a} \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a) \boldsymbol{\ell}(h,s,a)^2} \leq \sqrt{\eta_t \sum_{h,s,a} \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a)} = \underbrace{\sqrt{\eta_t H}}_{\alpha_t} .$$

Since $\boldsymbol{q}_\tau^{\max}(h,s,a) = \max_{\widehat{p} \in \mathcal{P}} \boldsymbol{q}^{\pi_\tau, \widehat{p}}(h,s,a) \geq \max\{\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau, h, s, a), \boldsymbol{q}^{\pi_\tau}(h,s,a)\}$ we have that

$$\begin{aligned}
\mathbb{E}[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t, \boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}] &= \eta_t \, \mathbb{E}\left[\sum_{h,s,a} \boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau, h, s, a) \widehat{\boldsymbol{\ell}}_\tau(h,s,a)^2\right] \\
&\leq \eta_t \, \mathbb{E}\left[\sum_{h,s,a} \frac{\mathbb{E}\left[\mathbb{I}\{s_{t,h} = s, a_{t,h} = a\} \mid \mathcal{F}_\tau\right]}{\boldsymbol{q}_\tau^{\max}(h,s,a)}\right] \\
&= \eta_t \, \mathbb{E}\left[\sum_{h,s,a} \frac{\boldsymbol{q}^{\pi_\tau(h,s,a)}}{\boldsymbol{q}_\tau^{\max}(h,s,a)}\right] \leq \underbrace{\eta_t HSA}_{\beta_t^2} .
\end{aligned}$$

Finally, pick any t, t',

$$\begin{aligned}\sqrt{\gamma}\|\widehat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} &\le \gamma\sqrt{\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)^2\widehat{\boldsymbol{\ell}}_t(h,s,a)^2} \le \gamma\sqrt{\sum_{h,s,a}\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}\\ &= \gamma\sqrt{H} \le \frac{1}{128 d_{\max}}\,,\end{aligned}$$

where the second inequality is due to the fact that $\boldsymbol{q}_t^{\max}(h,s,a) \ge \boldsymbol{q}^{\pi_t}(h,s,a)$ and that $\gamma = \kappa$ and the last inequality is by definition of γ. Thus, applying Lemma 3 with $\boldsymbol{b}_t = \mathbb{E}[\widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t \mid \mathcal{F}_t]$, we get

$$\begin{aligned}\textsc{Reg} \le \underbrace{R_T(\widetilde{\boldsymbol{u}}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v})}_{\textsc{Penalty}} + 8HSA\sum_{t=1}^T\eta_t + 8H\sum_{t=1}^T\eta_t|m_t| + \underbrace{\sum_{t=1}^T\mathbb{E}[\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)^\top(\boldsymbol{\ell}_t-\widehat{\boldsymbol{\ell}}_t)]}_{\textsc{Bias}_1}\\ + \underbrace{\sum_{t=1}^T\mathbb{E}[\widetilde{\widetilde{\boldsymbol{u}}}^\top(\widehat{\boldsymbol{\ell}}_t-\boldsymbol{\ell}_t)]}_{\textsc{Bias}_2} + 8\sqrt{H}\underbrace{\sum_{t=1}^T\sqrt{\eta_t}\,\mathbb{E}[\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}]}_{\textsc{Drift}}.\end{aligned}$$

Recall that $\min_{h,s,a,s'}\widetilde{\boldsymbol{u}}(h,s,a,s') \ge \frac{1}{T^3H^2S^4A^2}$. Thus, using the third fact of Lemma 25 with $b = \frac{1}{T^3H^2S^4A^2}$, $K = HS^2A$ and $B = H$, we conclude

$$\begin{aligned}\textsc{Penalty} &\le \frac{H\left(1+\log\left(S^2A\right)\right)}{\eta_T} + \frac{HS^2A\log\left(T^3H^2S^4A^2\right)}{\gamma}\\ &\le \frac{4H\log\left(SA\right)}{\eta_T} + \frac{4HS^2A\log\left(HSAT\right)}{\gamma}.\end{aligned}$$

Now we deal with the primary term that makes up $\textsc{Drift}$, for each t

$$\begin{aligned}&\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} = \left(\sum_{\tau\in m_t}\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau\right)\nabla^{-2}R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))\left(\sum_{\tau\in m_t}\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau\right)\\ &= \sum_{\tau\in m_t}\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} + \sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))(\boldsymbol{\ell}_{\tau'}-\widehat{\boldsymbol{\ell}}_{\tau'})\,. \qquad (24)\end{aligned}$$

We continue by bounding the first term on the right hands side in expectation:

$$\begin{aligned}\sum_{\tau\in m_t}\mathbb{E}\left[\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right] &\le 4\sum_{\tau\in m_t}\mathbb{E}\left[\|(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right]\\ &\le 4\sum_{\tau\in m_t}\mathbb{E}\left[\|\boldsymbol{\ell}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right] + 4\sum_{\tau\in m_t}\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right]\\ &\le 4|m_t|(\alpha_t^2+\beta_t^2) \le 8\eta_t HSA|m_t|.\end{aligned}$$

We bound the second term on the right hand side of (24) by first applying the law of total expectation,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau\in m_t}\sum_{\tau'\in m_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R_t(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t))(\boldsymbol{\ell}_{\tau'}-\widehat{\boldsymbol{\ell}}_{\tau'})\right]\\
&=\mathbb{E}\Bigg[\eta_t\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)(\boldsymbol{\ell}_\tau(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau(h,s,a)\mid\mathcal{F}_t])\big(\boldsymbol{\ell}_{\tau'}(h,s,a)\\
&\quad-\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}(h,s,a)\mid\mathcal{F}_t]\big)\Bigg]\\
&+\mathbb{E}\Bigg[\gamma\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)^2(\boldsymbol{\ell}_\tau(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau(h,s,a)\mid\mathcal{F}_t])\big(\boldsymbol{\ell}_{\tau'}(h,s,a)\\
&\quad-\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}(h,s,a)\mid\mathcal{F}_t]\big)\Bigg].
\end{aligned}$$

Then, since $\boldsymbol{\ell}_{\tau'}(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}(h,s,a)\mid\mathcal{F}_t]\in[0,1]$, we can bound the first term on the right-hand-side above by

$$\begin{aligned}
&\mathbb{E}\Big[\eta_t\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)(\boldsymbol{\ell}_\tau(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau(h,s,a)\mid\mathcal{F}_t])\Big]\\
&\quad\le\mathbb{E}\left[\eta_t|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)\boldsymbol{\ell}_\tau(h,s,a)\frac{\boldsymbol{q}_\tau^{\max}(h,s,a)-\boldsymbol{q}^{\pi_\tau}(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)}\right]\\
&\quad\le 2\,\mathbb{E}\left[\eta_t|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,h,s,a)\frac{\boldsymbol{q}_\tau^{\max}(h,s,a)-\boldsymbol{q}^{\pi_\tau}(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)}\right]\\
&\quad\le 2\,\mathbb{E}\left[\eta_t|m_t|\sum_{\tau\in m_t}\sum_{h,s,a}\boldsymbol{q}_\tau^{\max}(h,s,a)-\boldsymbol{q}^{\pi_\tau}(h,s,a)\right]\\
&\quad\le 2\,\mathbb{E}\left[\eta_t|m_t|\sum_{\tau\in m_t}\|\boldsymbol{q}_\tau^{\max}-\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)\|_1+\|\boldsymbol{q}^{\pi_\tau}-\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)\|_1\right]\le 12\eta_t|m_t|^2\frac{H^2S}{T}\,,
\end{aligned}$$

where the third inequality is by equation (37), the fourth inequality is since $\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,h,s,a)\le\boldsymbol{q}_\tau^{\max}(h,s,a)$, and the fifth inequality is by Lemma 11. Likewise we can see that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{\tau\in m_t}\sum_{\tau'\in m_t/\{\tau\}}\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)^2(\boldsymbol{\ell}_\tau(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau(h,s,a)\mid\mathcal{F}_t])\big(\boldsymbol{\ell}_{\tau'}(h,s,a)\\
&\quad-\mathbb{E}[\widehat{\boldsymbol{\ell}}_{\tau'}(h,s,a)\mid\mathcal{F}_t]\big)\Bigg]\le 12|m_t|^2\frac{H^2S}{T}\,.
\end{aligned}$$

These inequalities combined with Jensen's inequality gives

$$\begin{aligned}
8\sqrt{H}\cdot \textsc{Drift} &\le 32H\sum_{t=1}^{T}\eta_t\sqrt{SA|m_t|}+\sum_{t=1}^{T}96H(\eta_t+\sqrt{\eta_t\gamma})|m_t|\sqrt{\frac{H^2S}{T}}\\
&\le 32H\sum_{t=1}^{T}\eta_t(SA+|m_t|)+H\sqrt{ST}\\
&\le 65H\sqrt{\log(SA)(TSA+D)}\,,
\end{aligned}$$

where we used that $\eta,\gamma\le\frac{1}{96Hd_{\max}}$. Next, we bound $\textsc{Bias}_1$. Let $\mathcal{G}_t$ be the history of all episodes in $[t-1]$, and note that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)$, $\boldsymbol{q}_t^{\max}$ and $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)$ are all determined by $\mathcal{G}_t$. Therefore,

$$\begin{aligned}
\textsc{Bias}_1 &= \mathbb{E}\left[\sum_{t,h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m,h,s,a)(\boldsymbol{\ell}_t(h,s,a)-\mathbb{E}[\widehat{\boldsymbol{\ell}}_t(h,s,a)\mid\mathcal{G}_t])\right]\\
&= \mathbb{E}\left[\sum_{t,h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m,h,s,a)\boldsymbol{\ell}_t(h,s,a)\left(1-\frac{\boldsymbol{q}^{\pi_t}(h,s,a)}{\boldsymbol{q}_{t,h}^{\max}(s,a)}\right)\right]\\
&\le \mathbb{E}\left[\sum_{t,h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m,h,s,a)\frac{|\boldsymbol{q}_{t,h}^{\max}(s,a)-\boldsymbol{q}^{\pi_t}(h,s,a)|}{\boldsymbol{q}_t^{\max}(h,s,a)}\right].
\end{aligned}$$

Now, as in the proof of Lemma 3, $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)\in\mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t),\frac{1}{2\sqrt{\gamma}})$. Thus, by equation (37), $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m,h,s,a)\le 2\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)\le 2\boldsymbol{q}_t^{\max}(h,s,a)$. Therefore,

$$\textsc{Bias}_1\le 2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{q}_t^{\max}-\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)\|_1]\le 2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{q}_t^{\max}-\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)\|_1+\|\boldsymbol{q}^{\pi_t}-\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)\|_1]\le 12H^2S.$$

where the last is by article 2 in Lemma 11.

Recall that by definition $\boldsymbol{q}_t^{\max}(h,s,a)\ge\boldsymbol{w}^{\pi_t}(h,s,a)$. Thus, $\mathbb{E}[\widehat{\boldsymbol{\ell}}_t(h,s,a)\mid\mathcal{F}_t]\le\boldsymbol{\ell}_t$ and $\textsc{Bias}_2\le 0$. Putting everything together gives

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &\le 16H^2S+71H\sqrt{\log(SA)(TSA+D)}+\frac{4H\log(SA)}{\eta_T}+\frac{4HS^2A\log(HSAT)}{\gamma}\\
&\le 16H^2S+72H\sqrt{\log(SA)(TSA+D)}+512d_{\max}H^2S^2A\log(HSAT)\\
&\quad+800d_{\max}H^2S^2A^2\\
&\le 72H\sqrt{\log(SA)(TSA+D)}+1338d_{\max}H^2S^2A^2\log(HSAT)\,.
\end{aligned}$$

■

Appendix D. Adversarial MDPs with Unknown Transitions

Theorem 14 (RESTATED) *Algorithm 4 with* $\gamma = \frac{1}{128\sqrt{Hd_{\max}}}$, $\eta = \frac{\sqrt{\log(SA)}}{\sqrt{SAT+D}}$, $\xi = \frac{1}{T}$ *and* $T \geq 4$ *guarantees,*

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] \lesssim\ & H^2 S\sqrt{AT\log\left(HSAT\right)} + H\sqrt{D\log\left(SA\right)} \\ & + H^3S^2A\log\left(HSAT\right)d_{\max} + H^3S^3A\log^2\left(HSAT\right).\end{aligned}$$

Proof We introduce ς, the good event, where $p \in \mathcal{P}_{j'}$ for all $j' \leq j_t$ and the compliment of the good event ς^c. By Lemma 13 we have that the good event holds with probability of at least $1 - 4/T^2$. We then start by decomposing the regret in the same way as in the known transition setting, let $\boldsymbol{u}$ be any comparator

$$\begin{aligned}\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \\ &= \mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^{T}\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle}_{\textsc{Shift-Penalty}}\right]. \qquad (25)\end{aligned}$$

Under the good event and by Lemma 11 there exists an $\widetilde{\boldsymbol{u}} \in \mathcal{W}_T$ such that SHIFT-PENALTY is bounded by $2H$. That allows us to bound

$$\textsc{Shift-Penalty} = \mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^{T}\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle + \mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \leq 2H + 4\frac{HSA}{T}\,, \qquad (26)$$

where we also used that $\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle \leq 1$ as well as the probability of the bad event being greater or equal to $4/T^2$. The ERROR can be bound using standard tools in the analysis of MDPs with unknown transitions (Lemma 29) by

$$\textsc{Error} \lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H^3S^3A\log^2(HSAT^3) + H^3S^2Ad_{max}\,. \qquad (27)$$

Bounding the REG will be the main challenge of this proof as we now estimate the transition function and consequently have a changing domain $\mathcal{W}_t$. We are looking to apply Lemma 3, our analysis is structured around epochs and to make sure that $\mathcal{W}_t = \mathcal{W}_\tau$ whenever $\tau \in m_t$, as is required by Assumption 1, we will only change our $\mathcal{W}_t$ once each epoch and not use any delayed information from any previous epoch. We define $\mathcal{E} = \{t : j_t \neq j_{t-1}\}$ be the set of rounds in which a new epoch starts and if we are changing epoch in round $t \in \mathcal{E}$, then we skip all oustanding observations, $m_t \subseteq \Lambda$.

$\mathcal{W}_t \subseteq \mathcal{W}_{t-1}$ holds by the construction of $\mathcal{W}_t$ and $\mathcal{W}_T$ is non-empty under the good event, fulfilling the rest of Assumption 1. Our regularizer fulfils Assumptions 2 under the same caveates as in the previous section. We split REG into the good and bad event

$$\textsc{Reg} = \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] = \mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] + \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right]$$

and bound the bad event first. By Lemma 13 the good event ς happens with a probability of at least $1 - \frac{4}{T^2}$ and we have that

$$\mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}})^\top \widehat{\boldsymbol{\ell}}_t\right] \le \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\ell}}_t\|_1\right] \le \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}4HT^2\right] \le 4H\,, \tag{28}$$

where we used Hölders inequality on $(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}})^\top \widehat{\boldsymbol{\ell}}_t$, upper bounded $\|\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}\|_\infty \le 1$ and finally upper bound all H non-zero elements of $\widehat{\boldsymbol{\ell}}_t$ with $\widehat{\boldsymbol{\ell}}_t(h,s,a) \le \frac{1}{\xi} = T$. Under the good event we already showed that Assumption 1 and Assuming 2 are satisfied, both of which are required for Lemma 3. Next we bound $\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}$ for $\kappa = \gamma$. We have that

$$\begin{aligned}\sqrt{\kappa}\|\widehat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} &\le \sqrt{\kappa}\sqrt{\gamma\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)^2\widehat{\boldsymbol{\ell}}_t(h,s,a)^2} \le \sqrt{\kappa\gamma}\sqrt{\sum_{h,s,a}\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}\\ &= \gamma\sqrt{H} = \frac{1}{128 d_{\max}}\,,\end{aligned}$$

where the second inequality is due to the fact that $\boldsymbol{q}_t^{\max}(h,s,a) \ge \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)$ and the last inequality is by definition of γ. For all $\tau \in m_t \cup \{t\}$, we have that

$$\|\boldsymbol{\ell}_\tau\|^2_{R,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} \le \eta\sum_{h,s,a}\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t,h,s,a)\boldsymbol{\ell}_t(h,s,a)^2 \le \underbrace{\eta H}_{\alpha^2}\,. \tag{29}$$

We re-define the filtration over all past events observed by the learner to include state information $\mathcal{F}_t = \big\{\big(\tau, s_{\tau,h}, a_{\tau,h}, h, \boldsymbol{\ell}_\tau(h, s_{\tau,h}, a_{\tau,h})\big) \,:\, \tau + d_\tau < t, h \in [H]\big\}$. Then using that on the good event ς, $\boldsymbol{q}_\tau^{\max}(h,s,a) \ge \boldsymbol{q}^{\pi_\tau}(h,s,a)$, we can bound

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)} \,|\, \mathcal{F}_t, \varsigma] &\le \mathbb{E}\left[\eta\sum_{h,s,a}\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau,h,s,a)\widehat{\boldsymbol{\ell}}_\tau(h,s,a)^2\,\middle|\,\mathcal{F}_t,\varsigma\right]\\ &\le \mathbb{E}\left[\eta\sum_{h,s,a}\frac{\mathbb{I}\{s_{\tau,h}=s,a_{\tau,h}=a\}}{\boldsymbol{q}^{\pi_\tau}(h,s,a)}\,\middle|\,\mathcal{F}_\tau,\varsigma\right] = \underbrace{\eta HSA}_{\beta^2}\,.\end{aligned} \tag{30}$$

As in the proof of Theorem 8, the regularizer R_t as specified in (12) does not satisfy Assumption 2(c) because we can have $R_t(\boldsymbol{v}) \le R_{t-1}(\boldsymbol{v})$, but, as argued in the proof of Theorem 8, we can overcome this issue in a relative straightforward manner via the regularizer $\widetilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}'\in\mathcal{W}} R_t(\boldsymbol{v}')$, which has no impact on the iterates. This already puts us in a position to apply Lemma 3 to find that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\,\middle|\,\varsigma\right]\\ &\le \underbrace{\mathbb{E}\left[\sum_{t\in\Lambda}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}})^\top\boldsymbol{\ell}_t\,\middle|\,\varsigma\right]}_{\text{SKIPPED ROUNDS}} + \underbrace{R_T(\widetilde{\boldsymbol{u}}) - \min_{v\in\mathcal{W}_1}R_1(\boldsymbol{v})}_{\text{PENALTY}} + \sum_{t\in\overline{\Lambda}}8\beta_t^2\\ &\quad\underbrace{-\sum_{t\in\overline{\Lambda}}\mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \widetilde{\boldsymbol{u}})^\top\boldsymbol{b}_t\,\middle|\,\varsigma\right]}_{\text{BIAS}} + \sum_{t\in\overline{\Lambda}}\left(8\alpha_t^2|m_t| + 8\alpha_t\underbrace{\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\,\middle|\,\varsigma\right]}_{\text{MISSING ESTIMATES}}\right)\end{aligned} \tag{31}$$

Since we only start a new episode when any counter N_j doubles, we only start an new episode logarithmically often. This implies that $|\Lambda| \leq d_{\max} HSA \log(T)$, which means that the cost for the SKIPPED ROUNDS is upper bounded by:

$$\sum_{t\in\Lambda} (\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}})^\top \boldsymbol{\ell}_t \leq \sum_{t\in\Lambda} \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)^\top \boldsymbol{\ell}_t - \widetilde{\boldsymbol{u}}^\top \boldsymbol{\ell}_t \leq d_{\max} H^2 SA \log(T)\,. \tag{32}$$

where we used that $\boldsymbol{\ell}_t \in [0,1]$ per assumption.

We now bound the BIAS term. We know that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\gamma}})$ when $t \in \overline{\Lambda}$, which is also shown in the proof of Lemma 3. By Lemma 19 we have that $\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m, h, s, a) \leq 2\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a)$. By definition $\boldsymbol{q}_t^{\max}(h,s,a) \geq \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a)$ and thus,

$$\begin{aligned}
&- \mathbb{E}\left[\sum_{t\in\overline{\Lambda}} \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m)^\top \boldsymbol{b}_t\right] \\
&= \mathbb{E}\left[\sum_{t\in\overline{\Lambda}}\sum_{h,s,a} \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m, h, s, a)\boldsymbol{\ell}_t(h,s,a)\Big(1 - \frac{\boldsymbol{q}^{\pi_t}(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\Big)\right] \\
&= \mathbb{E}\left[\sum_{t\in\overline{\Lambda}}\sum_{h,s,a} \frac{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m, h, s, a)\boldsymbol{\ell}_t(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\big(\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a) + \xi\big)\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{t\in\overline{\Lambda}}\sum_{h,s,a} \frac{\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t, h, s, a)\boldsymbol{\ell}_t(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\big(|\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a)| + \xi\big)\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{t\in\overline{\Lambda}}\Big(\sum_{h,s,a} |\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a)| + \xi HSA\Big)\right] \\
&\lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H^3S^3A\log^2(HSAT^3) + H^3S^2Ad_{max} + HSA\,,
\end{aligned} \tag{33}$$

where the last inequality is due to $\xi = 1/T$ and Lemma 29, where we take an expectation over the event that the inequality in Lemma 29 holds similar to how we have been treating the good event and its complementary in other equations. For the second term in the BIAS, note that $\mathbb{E}[\widetilde{\boldsymbol{u}}^\top \boldsymbol{b}_t \mid \varsigma] \leq 0$, since under the good event we have $\boldsymbol{q}_\tau^{\max}(h,s,a) \geq \boldsymbol{q}^{\pi_\tau}(h,s,a)$ and thus, $\boldsymbol{b}_t(h,s,a) \leq 0$.

Next is the MISSING ESTIMATES term. We can not simply use the same argument as in Corollary 4 out of the box because $\widehat{\boldsymbol{\ell}}_t$ is a biased estimator, so we add and subtract the bias $\boldsymbol{b}_t$ and use the triangle inequality to find

$$\mathbb{E}\left[\Big\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\Big\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\,\varsigma\right] \leq \mathbb{E}\left[\Big\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau + \boldsymbol{b}_\tau - \widehat{\boldsymbol{\ell}}_\tau)\Big\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)} + \Big\|\sum_{\tau\in m_t}\boldsymbol{b}_\tau\Big\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\,\varsigma\right]$$

Now we recognise that we are not using any information from rounds that we have not seen yet and thus $\widehat{\boldsymbol{\ell}}_\tau$ and $\widehat{\boldsymbol{\ell}}_{\tau'}$ are independent if $\tau, \tau' \in m_t$. Furthermore, $\widehat{\boldsymbol{\ell}}_t$ is an unbiased estimator of $\boldsymbol{\ell}_t + \boldsymbol{b}_t$,

which allows us to use the exact same arguments as in Corollary 4 for the first term to find

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\varsigma\right] &\le \sqrt{\sum_{\tau\in m_t}\left(\mathbb{E}\left[\|\widehat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]-\mathbb{E}\left[\|\boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau\|^2_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\right]\right)}\\ &\le \sqrt{4|m_t|\beta_t^2}\end{aligned}$$

For the second term we start with the triangle inequality and Lemma 19,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau\in m_t}\boldsymbol{b}_\tau\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\varsigma\right] &\le \mathbb{E}\left[\sum_{\tau\in m_t}\|\boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\varsigma\right]\\ &\le 2\,\mathbb{E}\left[\sum_{\tau\in m_t}\|\boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\Big|\varsigma\right]\\ &\overset{\text{(a)}}{\le} 2\,\mathbb{E}\left[\sum_{\tau\in m_t}\left(\|\boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}+\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right)\Big|\varsigma\right]\\ &\overset{\text{(b)}}{\le} 2\,\mathbb{E}\left[\sum_{\tau\in m_t}\left(2\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_\tau(\widehat{\boldsymbol{L}}_\tau)}\right)\Big|\varsigma\right]\le 4|m_t|\alpha_t\,,\end{aligned}$$

where we added and subtracted $\boldsymbol{\ell}_\tau$ and used the triangle inequality again in inequality (a). Inequality (b) holds as $\boldsymbol{b}_\tau \le 0$, which holds as $\boldsymbol{q}_\tau^{\max}(h,s,a) \ge \boldsymbol{q}^{\pi_\tau}(h,s,a)$ under the good event for all (h,s,a). At the same time we have that $\widehat{\boldsymbol{\ell}}_\tau \ge 0$ by the construction of $\widehat{\boldsymbol{\ell}}_\tau$, showing that $\boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau = \mathbb{E}[\widehat{\boldsymbol{\ell}}_\tau]$ is non-negative. Together both of those facts allow us to conclude that $|\boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau| = \boldsymbol{\ell}_\tau+\boldsymbol{b}_\tau \le \boldsymbol{\ell}_\tau$, which we use in inequality (b).

Putting the last two equations together lets us bound the MISSING ESTIMATES term

$$\mathbb{E}\left[\left\|\sum_{\tau\in m_t}(\boldsymbol{\ell}_\tau-\widehat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t)}\Big|\varsigma\right] \le \sqrt{4|m_t|\beta_t^2}+4|m_t|\alpha_t\,. \tag{34}$$

The last thing to bound is the PENALTY term. Using the third fact of Lemma 25 with $b=\frac{1}{T^3H^2S^4A^2}$, $K=HS^2A$ and $B=H$, we conclude

$$\begin{aligned}\text{PENALTY} &\le \frac{H\left(1+\log\left(S^2A\right)\right)}{\eta}+\frac{HS^2A\log\left(T^3H^2S^4A^2\right)}{\gamma}\\ &\le \frac{4H\log\left(SA\right)}{\eta}+\frac{4HS^2A\log\left(HSAT\right)}{\gamma}.\end{aligned} \tag{35}$$

Putting things together for the REG term gives

$$
\begin{aligned}
\textsc{Reg} &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] \\
&= \mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] + \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] \\
&\le \underbrace{\frac{4H\log(SA)}{\eta} + \frac{4HS^2A\log(HSAT)}{\gamma}}_{\textsc{Penalty}} + \sum_{t\in\overline{\Lambda}} 8\beta_t^2 \\
&\quad + \sum_{t\in\overline{\Lambda}}\left(8\alpha_t^2|m_t| + 8\alpha_t\Big(\underbrace{\sqrt{4|m_t|\beta_t^2} + 4|m_t|\alpha_t}_{\textsc{missing estimates}}\Big)\right) + \underbrace{d_{\max}H^2SA\log(T)}_{\textsc{skipped rounds}} \\
&\quad \underbrace{- \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + 4H}_{\textsc{bias}} \\
&\le \frac{4H\log(SA)}{\eta} + \frac{4HS^2A\log(HSAT)}{\gamma} + 136\eta HSAT \\
&\quad + 148\eta HD + d_{\max}H^2SA\log(T) \\
&\quad - \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + 4H\,,
\end{aligned}
$$

where we used equations (28), (31), (32), (34), (35) in the first inequality and plugged in the values of α and β we found in equations (29) and (30) and also used that $\sqrt{ab} \le \frac{1}{2}(a+b)$ for $a, b > 0$. We plug in the learning rates to find

$$
\begin{aligned}
\textsc{Reg} &\lesssim \frac{H\log(SA)}{\eta} + \frac{HS^2A\log(HSAT)}{\gamma} + \eta HSAT \\
&\quad + \eta HD + d_{\max}H^2SA\log(T) \\
&\quad - \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + H \\
&\lesssim H\sqrt{SAT\log(SA)} + H\sqrt{D\log(SA)} \\
&\quad + d_{\max}\sqrt{H}S^2A\log(HSAT) + d_{\max}H^2SA\log(T) \\
&\quad - \sum_{t\in\overline{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t^m) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + H\,. \qquad (36)
\end{aligned}
$$

We can finally put everything together, starting from the regret again

$$\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle + \sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle + \sum_{t=1}^{T}\langle \widetilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] && \text{(Eqn (25))} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\widehat{\boldsymbol{L}}_t) - \widetilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] + \sqrt{H^4S^2AT\log(HSAT^3)} \\
&\quad + H^3S^3A\log^2(HSAT^3) + d_{\max}H^3S^2A && \text{(Eqns (26) and (27))} \\
&\lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H\sqrt{D\log(SA)} + d_{\max}H^3S^2A && \text{(Eqns (36) and (33))} \\
&\quad + H^3S^3A\log^2(HSAT^3) + d_{\max}\sqrt{H}S^2A\log(HSAT) + d_{\max}H^2SA\log(T)\,,
\end{aligned}$$

which concludes the proof. ■

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. (2022a)).

Algorithm 6: 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

Theorem 16 *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 6 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 6, $M \leq H$).

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,

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

■

Appendix F. Auxiliary Lemmas

Lemma 17 *Let $t \in [T]$ and suppose that $4\nabla^2 R_t(\boldsymbol{u}) \succeq \nabla^2 R_t(\boldsymbol{u}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{W}_t$ and $\boldsymbol{u}' \in \mathcal{D}_{R_t}(\boldsymbol{u}, \frac{1}{2})$. Let $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2})$ or $\boldsymbol{v} \in \mathcal{D}_{R_t}(\boldsymbol{v}', \frac{1}{2})$, then*

$$\|x\|_{R_t, \boldsymbol{v}'} \leq 2\|x\|_{R_t, \boldsymbol{v}} \ ,$$

for all $x \in \mathbb{R}^K$.

Proof First consider that if $\boldsymbol{v}' \in \mathcal{D}_{R_t}(\boldsymbol{v}, \frac{1}{2})$ then $\nabla^2 R_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2 R_t(\boldsymbol{v})$ and thus

$$\left(\nabla^2 R_t(\boldsymbol{v}')\right)^{-1} \preceq 4 \left(\nabla^2 R_t(\boldsymbol{v})\right)^{-1} \ .$$

We can arrive to the same inequality if $\boldsymbol{v} \in \mathcal{D}_{R_t}(\boldsymbol{v}', \frac{1}{2})$, by using $4\nabla^2 R(\boldsymbol{v}') \succeq \nabla^2 R(\boldsymbol{v})$. We can then follow directly

$$\|x\|_{R_t, \boldsymbol{v}'} = \sqrt{x^\top \left(\nabla^2 R(\boldsymbol{v}')\right)^{-1} \boldsymbol{x}} \leq 2\sqrt{x^\top \left(\nabla^2 R(\boldsymbol{v})\right)^{-1} \boldsymbol{x}} = 2\|x\|_{R_t, \boldsymbol{v}} \ .$$

■

Lemma 18 (Be-The-Leader Lemma) *Let $\widetilde{\boldsymbol{w}}_t(\widehat{\boldsymbol{L}}_t^\star) = \arg\min_{\boldsymbol{w} \in \mathcal{W}_t} \boldsymbol{w}^\top \widehat{\boldsymbol{L}}_t^\star + R_t(\boldsymbol{w})$. Suppose that $R_t(\boldsymbol{v}) \leq R_{t+1}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathcal{W}_t$ and all $t \in [T]$ and that $\mathcal{W}_t \subseteq \mathcal{W}_{t-1}$ is a non-empty compact convex set. Then, for any fixed $\boldsymbol{u} \in \mathcal{W}_T$, we have that*

$$\sum_{t \in \overline{\Lambda}} \widehat{\boldsymbol{\ell}}_t^\top (\widetilde{\boldsymbol{w}}_t(\widehat{\boldsymbol{L}}_t^\star) - \boldsymbol{u}) \leq R_T(\boldsymbol{u}) - \min_{\boldsymbol{v} \in \mathcal{W}_1} R_1(\boldsymbol{v})$$

Proof We will prove the statement by induction on T. For the induction step, assume that

$$\sum_{t \in \overline{\Lambda} \cap [T-1]} \widehat{\boldsymbol{\ell}}_t^\top \widetilde{\boldsymbol{w}}_t(\widehat{\boldsymbol{L}}_t^\star) + R_1(\boldsymbol{w}_1(\widehat{\boldsymbol{L}}_1)) \leq \sum_{t \in \overline{\Lambda} \cap [T-1]} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{v} + R_{T-1}(\boldsymbol{v})$$

for any $\boldsymbol{v} \in \mathcal{W}_T$. If $\overline{\Lambda} \cap [T-1] = \overline{\Lambda} \cap [T]$ the induction step holds. Otherwise $T \in \overline{\Lambda}$ and adding $\widehat{\boldsymbol{\ell}}_T^\top \widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star)$ to both sides of the above inequality and setting $\boldsymbol{v} = \widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star)$ on the right-hand side of the above inequality we find

$$\begin{aligned}
\sum_{t \in \overline{\Lambda} \cap [T]} \widehat{\boldsymbol{\ell}}_t^\top \widetilde{\boldsymbol{w}}_t(\widehat{\boldsymbol{L}}_t^\star) + R_1(\boldsymbol{w}_1(\widehat{\boldsymbol{L}}_1)) &\leq \sum_{t \in \overline{\Lambda} \cap [T]} \widehat{\boldsymbol{\ell}}_t^\top \widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star) + R_{T-1}(\widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star)) \\
&\leq \sum_{t \in \overline{\Lambda} \cap [T]} \widehat{\boldsymbol{\ell}}_t^\top \widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star) + R_T(\widetilde{\boldsymbol{w}}_T(\widehat{\boldsymbol{L}}_T^\star)) \\
&\leq \sum_{t \in \overline{\Lambda} \cap [T]} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{u} + R_T(\boldsymbol{u}),
\end{aligned}$$

which proves the induction step after reordering and observing that the base case holds by definition of $\boldsymbol{w}_1(\widehat{\boldsymbol{L}}_1)$. The statment is proven after applying $R_T \geq R_\tau$, which holds for all $\tau \in \overline{\Lambda}$, once. ■

Lemma 19 *Let $\mathcal{V} \subseteq \{\boldsymbol{x} \in \mathbb{R}^n : \forall i \in [n],\ \boldsymbol{x}(i) > 0\}$. Let $R : \mathcal{V} \to \mathbb{R}$ be some twice-differentiable convex function, and let $\phi(\boldsymbol{v}) = -\frac{1}{\gamma}\sum_{i=1}^n \log \boldsymbol{v}(i)$ be the log barrier with $\gamma \in (0,1)$. Assume that for any $\boldsymbol{v} \in \mathcal{V}$, $\nabla^2 R(\boldsymbol{v}) \succeq \nabla^2 \phi(\boldsymbol{v})$. Then for any $\boldsymbol{v}' \in \mathcal{D}_R(\boldsymbol{v}, \frac{1}{2\sqrt{\gamma}})$ and all $i \in [n]$,*

$$\tfrac{1}{2}\boldsymbol{v}(i) \leq \boldsymbol{v}'(i) \leq 2\boldsymbol{v}(i)\ .$$

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

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

On the other hand,

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

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

Lemma 20 *Let $a, b \in \mathbb{R}_+$ such that $a \geq b$, then*

$$\sqrt{a} - \sqrt{b} \leq \sqrt{a-b}\ .$$

Proof We show directly

$$\sqrt{a} - \sqrt{b} = \sqrt{\left(\sqrt{a} - \sqrt{b}\right)^2} = \sqrt{a + b - 2\sqrt{ab}} \leq \sqrt{a-b}\ .$$

■

Lemma 21 $\log(x)^2 \leq \frac{1}{x}$ *for all* $0 < x \leq 1$.

Proof Note that since $\log(x) \leq 0 \leq 1/x$ for $0 < x \leq 1$, $\log(x)^2 \leq \frac{1}{x}$ is equivalent to $-\log(x) \leq \frac{1}{\sqrt{x}}$ which we rearange to $-\sqrt{x}\log(x) \leq 1$. We maximise the function on the lefthandside on $x \in (0,1]$, taking a derivative yields

$$\frac{\partial - \sqrt{x}\log(x)}{\partial x} = -\frac{1}{2\sqrt{x}}\log(x) - \frac{1}{\sqrt{x}}\ .$$

Setting the derivative to 0 gives $x = e^{-2}$ as a possible maximum and $-\sqrt{e^{-2}}\log(e^{-2}) = \frac{2}{e} < 1$. The supremum of $-\sqrt{x}\log(x)$ may also lie on the boundary of $(0,1]$ but $-\sqrt{1}\log(1) = 0 < 1$ and

$$\lim_{x\to 0_+} -\sqrt{x}\log(x) = \lim_{x\to 0_+} \frac{-\log(x)}{\frac{1}{\sqrt{x}}} = \lim_{x\to 0_+} \frac{x^{-\frac{1}{2}}}{2x^{-\frac{3}{2}}} = 0 < 1\ ,$$

where we also used L'Hôpital's rule. We conclude that $-\sqrt{x}\log(x) \leq 1$ for all $0 < x \leq 1$. ■

Lemma 22 *Let $a, b, c \in \mathbb{R}$. Let $b \geq c$, then*

$$\max\{a, b\} - \max\{a, c\} \leq b - c$$

Proof If $a \geq b$, then $\max\{a, b\} - \max\{a, c\} = a - a = 0 \leq b - c$.
If $b \geq a \geq c$, then $\max\{a, b\} - \max\{a, c\} = b - a \leq b - c$.
If $b, c \geq a$, then $\max\{a, b\} - \max\{a, c\} = b - c$.
■

Lemma 23 (Part of Lemma 14 from Gaillard et al. (2014)) *Let $a_1, \ldots, a_m \in \mathbb{R}_+$ and call $s_i = a_1 + \ldots + a_i$. Let $f : (0, \infty) \to [0, \infty]$ be a non-increasing function. Then*

$$\sum_{i=1}^{m} a_i f(s_i) \leq \int_{a_1}^{s_m} f(x)\,\mathrm{d}x$$

Proof

$$\sum_{i=1}^{m} a_i f(s_i) = \sum_{i=1}^{m} \int_{s_{i-1}}^{s_i} f(s_i)\,\mathrm{d}x \leq \sum_{i=1}^{m} \int_{s_{i-1}}^{s_i} f(x)\,\mathrm{d}x = \int_{a_1}^{s_m} f(x)\,\mathrm{d}x\,,$$

where we used a telescoping sum in the first equality, the fact that f is non-increasing in the inequality and another telescoping sum in the last equality. ■

Lemma 24

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}\,, \qquad \sum_{t=1}^{T} \frac{|m_t|}{\sqrt{\sum_{\tau=1}^{t} |m_\tau|}} \leq 2\sqrt{D}\,, \qquad \sum_{t=1}^{T} t^{-\frac{1}{4}} \leq \frac{4}{3} T^{\frac{3}{4}}\,.$$

Proof By Lemma 23 with $a_1, \ldots, a_T = 1$ and $f(x) = \frac{1}{\sqrt{x}}$

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \int_{1}^{T} \frac{1}{\sqrt{x}}\,\mathrm{d}x \leq 2\sqrt{T}\,.$$

We replace m_t by $\widetilde{m}_t$, where we used that $|m_t| \leq |\widetilde{m}_t|$ which holds with probability one, then by Lemma 23 with $a_i = |\widetilde{m}_i|$ and $f(x) = \frac{1}{\sqrt{x}}$

$$\sum_{t=1}^{T} \frac{|m_t|}{\sqrt{\sum_{\tau=1}^{t} |m_\tau|}} \leq \sum_{t=1}^{T} \frac{|\widetilde{m}_t|}{\sqrt{\sum_{\tau=1}^{t} |\widetilde{m}_\tau|}} \leq \int_{1}^{D} \frac{1}{\sqrt{x}}\,\mathrm{d}x \leq 2\sqrt{D}\,.$$

One last time by Lemma 23 with $a_1, \dots, a_T = 1$ and $f(x) = x^{-\frac{1}{4}}$

$$\sum_{t=1}^{T} t^{-\frac{1}{4}} \le \int_1^T x^{-\frac{1}{4}} \, \mathrm{d}x \le \frac{4}{3} T^{\frac{3}{4}} \, .$$

■

Lemma 25 *Let* $\mathcal{V}(b) \subseteq \{\boldsymbol{x} : 0 \le b \le \boldsymbol{x}(i) \le 1\}$ *and let* $\Gamma_t(\boldsymbol{v}) = \sum_{i=1}^{K} \left(\frac{1}{\eta_t} \boldsymbol{v}(i) \log(\boldsymbol{v}(i)) - \frac{1}{\gamma} \log(\boldsymbol{v}(i)) \right)$ *for some* $\gamma, \eta_t > 0$ *and* $\eta_t \ge \eta_{t+1}$*, then*

$$4 \nabla^2 \Gamma_t(\boldsymbol{v}) \succeq \nabla^2 \Gamma_t(\boldsymbol{v}') \succeq \frac{1}{4} \nabla^2 \Gamma_t(\boldsymbol{v}) \, ,$$

for all $\boldsymbol{v}', \boldsymbol{v} \in \mathcal{V}(b)$*,* $\boldsymbol{v}' \in \mathcal{D}_{\Gamma_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\gamma}})$*, and all* t*. Furthermore, if there exists an* $\lambda > 0$ *and* η_t *is such that* $\sqrt{\eta_{t+\delta}} \left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t} \right) \le \frac{\sqrt{\lambda}}{\sqrt{K}}$ *for a given* t *and* $\delta \in [d_{\max}]$*, then*

$$\left(\nabla \Gamma_t(\boldsymbol{v}) - \nabla \Gamma_{t+\delta}(\boldsymbol{v}) \right)^\top y \le \sqrt{\lambda} \sqrt{y^\top \nabla^2 \Gamma_{t+\delta}(\boldsymbol{v}) y} \, ,$$

for all $\boldsymbol{v} \in \mathcal{V}(b)$ *and all* $\boldsymbol{y} \in \mathbb{R}^K$*. Finally, if* $b' \ge 0$*,* $b > 0$*, and* $\|v\|_1 \le B$ *for some* $B > 0$ *and all* $\boldsymbol{v}(i) \in \mathcal{V}(b')$*, then for all* $\boldsymbol{u} \in \mathcal{V}(b)$

$$\Gamma_T(\boldsymbol{u}) - \min_{\boldsymbol{v} \in \mathcal{V}(b')} \Gamma_T(\boldsymbol{v}) \le \frac{B \left(1 + \log\left(\frac{K}{B}\right)\right)}{\eta_T} + \frac{K \log\left(\frac{1}{b}\right)}{\gamma} \, .$$

Proof We start with the first statement and we state the derivatives of Γ_t

$$\begin{aligned}
\Gamma_t(\boldsymbol{v}) &= \sum_{i=1}^{K} \left(\frac{1}{\eta_t} \boldsymbol{v}(i) \log(\boldsymbol{v}(i)) - \frac{1}{\gamma} \log(\boldsymbol{v}(i)) \right) \\
\left(\nabla \Gamma_t(\boldsymbol{v}) \right)(i) &= \frac{1}{\eta_t} \log(\boldsymbol{v}(i)) - \frac{1}{\gamma \boldsymbol{v}(i)} \\
\left(\nabla^2 \Gamma_t(\boldsymbol{v}) \right)(i,i) &= \frac{1}{\eta_t \boldsymbol{v}(i)} + \frac{1}{\gamma \boldsymbol{v}^2(i)} \, ,
\end{aligned}$$

where $\left(\nabla^2 \Gamma_t(\boldsymbol{v}) \right)(j,i) = 0$ if $j \ne i$. Now, we have that

$$\frac{1}{2\sqrt{\gamma}} \ge \|\boldsymbol{v} - \boldsymbol{v}'\|_{\Gamma_t, \boldsymbol{v}} \ge \frac{|\boldsymbol{v}(i) - \boldsymbol{v}'(i)|}{\sqrt{\gamma} \boldsymbol{v}(i)} = \frac{1}{\sqrt{\gamma}} \left| 1 - \frac{\boldsymbol{v}'(i)}{\boldsymbol{v}(i)} \right| \, ,$$

or equivalently, $\left| 1 - \frac{\boldsymbol{v}'(i)}{\boldsymbol{v}(i)} \right| \le \frac{1}{2}$. If $\boldsymbol{v}(i) \ge \boldsymbol{v}'(i)$ then $\left| 1 - \frac{\boldsymbol{v}'(i)}{\boldsymbol{v}(i)} \right| = 1 - \frac{\boldsymbol{v}'(i)}{\boldsymbol{v}(i)}$ and we re-arrange to find $\boldsymbol{v}'(i) \ge \frac{1}{2} \boldsymbol{v}(i)$. Likewise, if $\boldsymbol{v}(i) \le \boldsymbol{v}'(i)$ then we can see that $\boldsymbol{v}'(i) \le \frac{3}{2} \boldsymbol{v}(i)$. Thus, we can conclude that for $\boldsymbol{v}' \in \mathcal{D}_{\Gamma_t}(\boldsymbol{v}, \frac{1}{2\sqrt{\gamma}})$

$$\frac{1}{2} \boldsymbol{v}(i) \le \boldsymbol{v}'(i) \le \frac{3}{2} \boldsymbol{v}'(i) \, . \tag{37}$$

Using these properties and the second derivative of Γ_t as written above we can verify the first statement as

$$4\nabla^2\Gamma_t(\boldsymbol{v}) = 4\,\mathrm{diag}\left(\frac{\gamma}{\eta_t\boldsymbol{v}} + \frac{1}{\boldsymbol{v}^2}\right) \succeq \mathrm{diag}\left(\frac{\gamma}{\eta_t\boldsymbol{v}'} + \frac{1}{\boldsymbol{v}'^2}\right) = \nabla^2\Gamma_t(\boldsymbol{v}')\,,$$

where the division is meant elementwise and $\nabla^2\Gamma_t(\boldsymbol{v}') \succeq \frac{1}{4}\nabla^2\Gamma_t(\boldsymbol{v})$ goes through analogously. Next for the second statement, we first pick any $\boldsymbol{y} \in \mathbb{R}^K$ and then establish that

$$\sum_{i=1}^{K} -|y(i)|\log(\boldsymbol{v}(i)) \leq \sqrt{K}\sqrt{\sum_{i=1}^{K} y(i)^2\log(\boldsymbol{v}(i))^2} \leq \sqrt{K}\sqrt{\sum_{i=1}^{K} y(i)^2\frac{1}{\boldsymbol{v}(i)}}$$

using the AM-QM inequality, which holds as all $-|\boldsymbol{y}(i)|\log(\boldsymbol{v}(i))$ are real positive numbers and using the fact that $\log(x)^2 \leq \frac{1}{x}$ for $0 < x \leq 1$ as shown by Lemma 21. Using the above equation gives

$$\begin{aligned}
\left(\nabla\Gamma_t(\boldsymbol{v}) - \nabla\Gamma_{t+\delta}(\boldsymbol{v})\right)^\top\boldsymbol{y} &\leq \left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right)\sum_{i=1}^{K} -|\boldsymbol{y}(i)|\log(\boldsymbol{v}(i)) \\
&\leq \left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right)\sqrt{K}\sqrt{\sum_{i=1}^{K}\boldsymbol{y}(i)^2\frac{1}{\boldsymbol{v}(i)}} \\
&= \sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right)\sqrt{K}\sqrt{\sum_{i=1}^{K}\frac{1}{\eta_{t+\delta}}\boldsymbol{y}(i)^2\frac{1}{\boldsymbol{v}(i)}} \\
&\leq \sqrt{\lambda}\sqrt{\sum_{i=1}^{K}\left(\frac{1}{\eta_{t+\delta}}\boldsymbol{y}(i)^2\frac{1}{\boldsymbol{v}(i)} + \frac{1}{\gamma}\frac{\boldsymbol{y}(i)^2}{\boldsymbol{v}(i)^2}\right)} \\
&= \sqrt{\lambda}\sqrt{\boldsymbol{y}^\top\nabla^2\Gamma_{t+\delta}(\boldsymbol{v})\boldsymbol{y}}\,,
\end{aligned}$$

where we only used $|\boldsymbol{y}(i)| \geq \boldsymbol{y}(i)$ in the first inequality, the above equation in the second inequality, and the assumption on η_t and λ and the fact that $\frac{1}{\gamma}\frac{\boldsymbol{y}(i)^2}{\boldsymbol{v}(i)^2} \geq 0$ in the last inequality.

For the last statement we start with the negative entropy component of Γ_T. Without loss of generality we may assume that $\boldsymbol{v}(i) > 0$ as we may define $-\boldsymbol{v}(i)\log(\boldsymbol{v}(i)) = 0$. We can bound the negative entropy component of Γ_T as

$$\begin{aligned}
-\sum_{i=1}^{K}\boldsymbol{v}(i)\log\boldsymbol{v}(i) &= \|\boldsymbol{v}\|_1\sum_{i=1}^{K}\frac{\boldsymbol{v}(i)}{\|\boldsymbol{v}\|_1}\log\frac{1}{\boldsymbol{v}(i)} \\
&\leq \|\boldsymbol{v}\|_1\log\left(\sum_{i=1}^{K}\frac{\boldsymbol{v}(i)}{\|\boldsymbol{v}\|_1}\frac{1}{\boldsymbol{v}(i)}\right) \\
&\leq \|\boldsymbol{v}\|_1\log\left(\frac{K}{\|\boldsymbol{v}\|_1}\right) + \|\boldsymbol{v}\|_1 \leq B\left(1 + \log\left(\frac{K}{B}\right)\right)\,,
\end{aligned}$$

where we used Jensen's inequality in the second step and the fact that $x\log(\frac{K}{x})+x$ is in increasing on $x\in[1,K]$ in the last inequality. Set $\boldsymbol{v}^+ = \arg\min_{\boldsymbol{v}\in\mathcal{W}(b')}\Gamma_T(\boldsymbol{v})$, then

$$\begin{aligned}\Gamma_T(\boldsymbol{u})-\Gamma_T(\boldsymbol{v}^+) &= \sum_{i=1}^{K}\Big(\frac{\boldsymbol{u}(i)}{\eta_T}\log(\boldsymbol{u}(i))-\frac{1}{\gamma}\log(\boldsymbol{u}(i))\Big)\\ &\quad-\sum_{i=1}^{K}\Big(\frac{\boldsymbol{v}^+(i)}{\eta_T}\log(\boldsymbol{v}^+(i))-\frac{1}{\gamma}\log(\boldsymbol{v}^+(i))\Big)\\ &\le \frac{B\left(1+\log\left(\frac{K}{B}\right)\right)}{\eta_T}+\frac{K\log\left(\frac{1}{b}\right)}{\gamma}\end{aligned}$$

where we used the fact that $b\le\boldsymbol{u}(i)\le 1$ since $\boldsymbol{u}(i)\in\mathcal{V}(b)$, $\frac{1}{\eta_1}\le\frac{1}{\eta_T}$, and the fact that $-\log(x)$ is a decreasing function and non-negative for $x\in(0,1]$. ■

Lemma 26 *Let* $\eta_t = \min\left\{a, \frac{1}{\sqrt{bt+c\sum_{\tau=1}^{t}|m_\tau|}}\right\}$ *for some* $a,b,c\in\mathbb{R}_+$. *If* $a\le\frac{d}{bd_{\max}+cd_{\max}^2}$ *for some* $d\in\mathbb{R}$, *then*

$$\sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}}-\frac{1}{\eta_t}\right)\le\sqrt{d}\,.$$

Proof We start by showing that

$$\begin{aligned}\frac{1}{\eta_{t+\delta}}-\frac{1}{\eta_t} &\le \sqrt{b(t+\delta)+c\sum_{\tau=1}^{t+\delta}|m_\tau|}-\sqrt{bt+c\sum_{\tau=1}^{t}|m_\tau|}\\ &\le \sqrt{b\delta+c\sum_{\tau=t+1}^{t+\delta}|m_\tau|}\\ &\le \sqrt{bd_{\max}+cd_{\max}^2}\,,\end{aligned}$$

where we used our assumption on η_t together with $\max\{x,y\}-\max\{x,z\}\le y-z$ (Lemma 22) in the first inequality, $\sqrt{a}-\sqrt{b}\le\sqrt{a-b}$ (Lemma 20) in the second inequality and the fact that $\delta\le d_{\max}$ and $|m_t|\le d_{\max}$ in the third inequality. And from here we can see that

$$\sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}}-\frac{1}{\eta_t}\right)\le\sqrt{\eta_{t+\delta}}\sqrt{bd_{\max}+cd_{\max}^2}\le\sqrt{d}\,.$$

■

Lemma 27 *Let* $\eta_t = \min\left\{a, \frac{1}{\sqrt{bt+c\sum_{\tau=1}^{t}|m_\tau|}}\right\}$ *for some* $a,b,c\in\mathbb{R}_+$. *If* $a\le\frac{d}{\sqrt{bd_{\max}+cd_{\max}^2}}$ *for some* $d\in\mathbb{R}$, *then*

$$\sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}}-\frac{1}{\eta_t}\right)\le\sqrt{d}(bd_{\max}+cd_{\max}^2)^{1/4}\,.$$

Proof We start by showing that

$$\begin{aligned}\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t} &\leq \sqrt{b(t+\delta) + c\sum_{\tau=1}^{t+\delta}|m_\tau|} - \sqrt{bt + c\sum_{\tau=1}^{t}|m_\tau|} \\ &\leq \sqrt{b\delta + c\sum_{\tau=t+1}^{t+\delta}|m_\tau|} \\ &\leq \sqrt{bd_{\max} + cd_{\max}^2}\,,\end{aligned}$$

where we used our assumption on η_t together with $\max\{x,y\} - \max\{x,z\} \leq y - z$ (Lemma 22) in the first inequality, $\sqrt{a} - \sqrt{b} \leq \sqrt{a-b}$ (Lemma 20) in the second inequality and the fact that $\delta \leq d_{\max}$ and $|m_t| \leq d_{\max}$ in the third inequality. And from here we can see that

$$\sqrt{\eta_{t+\delta}}\left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t}\right) \leq \sqrt{\eta_{t+\delta}}\sqrt{bd_{\max} + cd_{\max}^2} \leq \sqrt{d}\,.$$

■

Lemma 28 (Lemma D.11 of Jin et al. (2022); see also Lemma 4 of Jin et al. (2020)) *With probability $1-\delta$, for any collection of transition functions $\{p^s_{t,h}\}_{s\in\mathcal{S}}$ such that $p^s_t \in \widehat{P}_j$*

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathfrak{A}}\left|q_t^{p_t^s,\pi_t}(h,s,a) - q_t^{\pi_t}(h,s,a)\right| &\lesssim H\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S},a\in\mathfrak{A}}\epsilon_t(h,s,a)q^{\pi_t}(h,s,a) \\ &+ HS\sum_{t=1}^{T}\sum_{1\leq h<\widetilde{h}\leq H}\sum_{s\in\mathcal{S},a\in\mathfrak{A},s'\in\mathcal{S}}\sum_{\widetilde{s}\in\mathcal{S},\widetilde{a}\in\mathfrak{A}}\epsilon_t(s' \mid h,s,a)q^{\pi_t}(h,s,a) \\ &\cdot\min\left\{2, \sum_{\widetilde{s}'\in\mathcal{S}}\epsilon_t(\widetilde{s}' \mid \widetilde{h},\widetilde{s},\widetilde{a})\right\}q^{\pi_t}(\widetilde{h},\widetilde{s},\widetilde{a} \mid s'; h+1) + H^3S^2Ad_{max} \qquad (38)\end{aligned}$$

where $q^{\pi_t}(\widetilde{h},\widetilde{s},\widetilde{a} \mid \widetilde{s}'; h)$ be the probability to visit $(\widetilde{s},\widetilde{a})$ in time $\widetilde{h}$ given that we visited $\widetilde{s}'$ in time h.

Lemma 29 (Lemma D.12 of Jin et al. (2022) adapted to epochs) *With probability $1 - 10/T$, for any collection of transition functions $\{p_t^{h,s}\}_{h\in[H],s\in\mathcal{S}}$ such that $p_t^{h,s} \in \widehat{P}_j$*

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathfrak{A}}\left|q_t^{p_t^{h,s},\pi_t}(h,s,a) - q_t^{\pi_t}(h,s,a)\right| \\ &\qquad\lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H^3S^3A\log^2(HSAT^3) + H^3S^2Ad_{max}.\end{aligned}$$

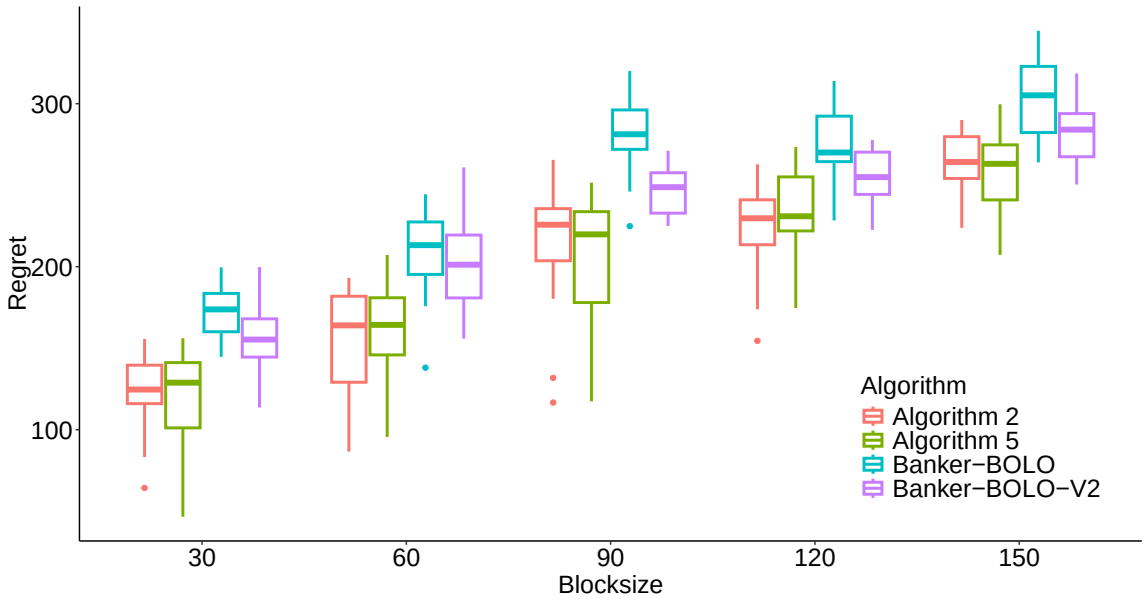

Figure 3: Boxplot of the regret over 20 repetitions over $T = 10000$ rounds with $K = 10$.

Proof Following the exact same steps as in the proof of Lemma E.5 in Lancewicki et al. (2022b), with probability of at least $1 - \delta$, the first term in (38) can be bounded by,

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S},a\in\mathfrak{A}} \epsilon_t(h,s,a)q^{\pi_t}(h,s,a) \lesssim \sqrt{S\log(HSAT^3)}\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{\sqrt{N_{j(t)}(h,s,a)\vee 1}}$$
$$+ S\log(HSAT^3)\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{N_{j(t)}(h,s,a)\vee 1} + HS\log^2(HSAT^3) \tag{39}$$

Similarly, the second summation (38) is bounded by,

$$HS\log^2(HSAT^3)\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{N_{j(t)}(h,s,a)\vee 1} \tag{40}$$

Finally, (39) and (40) are bounded by $O(HS\sqrt{AT\log(HSAT^3)} + HS^2\log^2(HSAT^3))$ and $O(H^2S^2A\log^2(HSAT^3))$ respectively using standard arguments - see for example the proof of Lemma 10 in Jin et al. (2020). ■

Appendix G. Further Results of the Experiments

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.

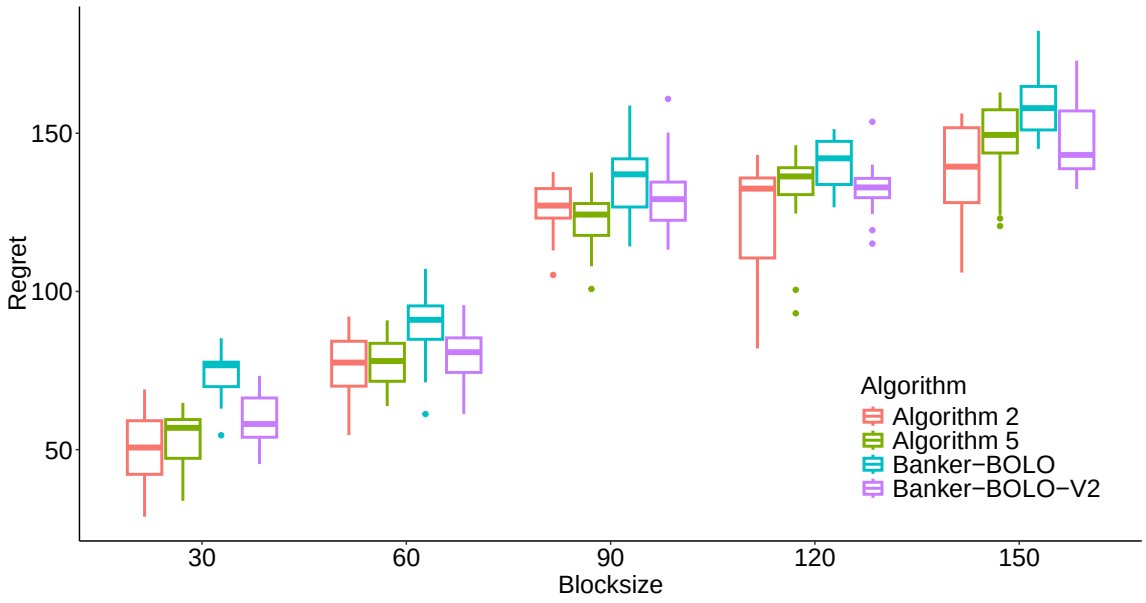

Figure 4: Boxplot of the regret over 20 repetitions over $T = 10000$ rounds with $K = 40$.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

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.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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.

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. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/chen13a.html`.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 11437–11449. Curran Associates, Inc., 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.

Varsha Dani, Sham M Kakade, and Thomas Hayes. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2007.

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.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012. doi: 10.1109/TNET.2011.2181864.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 176–196. PMLR, 13–15 Jun 2014.

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.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(79):2369–2403, 2007. URL `http://jmlr.org/papers/v8/gyoergy07a.html`.

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

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

Dirk van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. In *Conference on Learning Theory*, pages 1285–1321, 2023.

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

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

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

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In *Advances in Neural Information Processing Systems*, 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. In *Advances in Neural Information Processing Systems*, volume 35, pages 33469–33481. Curran Associates, Inc., 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.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA, 09–12 May 2015. PMLR. URL `https://proceedings.mlr.press/v38/kveton15.html`.

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, 2022a.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Cooperative online learning in stochastic and adversarial MDPs. In *International Conference on Machine Learning*, pages 11918–11968, 2022b.

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

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.

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

Jonathan Louëdec, Max Chevalier, Josiane Mothe, Aurélien Garivier, and Sébastien Gerchinovitz. A multiple-play bandit algorithm applied to recommender systems. In *28th International Florida Artificial Intelligence Research Society (FLAIRS 2015)*, pages 67–72, Hollywood, United States, May 2015. URL `https://hal.science/hal-04077707`.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 11752–11762. Curran Associates, Inc., 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.

He Ni, Hao Xu, Dan Ma, and Jun Fan. Contextual combinatorial bandit on portfolio management. *Expert Systems with Applications*, 221:119677, 2023.

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.

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.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1113–1122, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/wen15.html.

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.