Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Artificial Intelligence (37.th cycle)

Adversarial Bandits and which Leader to Follow

Lukas Zierahn

* * * * *

Supervisors
Professor Nicolò Cesa-Bianchi, Supervisor
Professor Gergely Neu, Co-supervisor

Politecnico di Torino
October 13, 2025

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

.....

Lukas Zierahn

Milan, October 13, 2025

Summary

Multi-armed bandits provide a versatile framework for learning in repeated decision-making scenarios. At each discrete timestep, a learner selects an action and observes only the loss associated with that action. The feedback for unchosen actions remains unobserved. The learner's performance is measured by regret, defined as the difference between the learner's cumulative loss and the loss of the best fixed action in hindsight. Bandit algorithms have seen widespread application, including in ad selection, hyperparameter tuning, and pathfinding.

This work explores bandits in four chapters. The first chapter introduces the classical bandit setting, outlining assumptions on actions, losses, and the feedback the learner observes. We discuss importance-weighted loss estimators and highlight the limitations of the Follow-the-Leader (FTL) algorithm, which simply chooses the best action in hindsight at each timestep. We then illustrate Follow-the-Regularized-Leader (FTRL), a well known improvement on FTL that achieves optimal regret bounds in many bandit settings.

In Chapter 2, we consider combinatorial bandits, where the learner plays multiple actions simultaneously. Depending on the feedback model, the learner either observes individual losses (semi-bandit feedback) or their sum (full-bandit feedback). We provide a regret analysis for FTRL in this setting by decomposing the regret to a stability and regularization term, offering a more gentle introduction to the advanced concepts developed later.

Chapter 3 introduces adversarial contextual combinatorial bandits. Contextual bandits extend the bandit setting by allowing the learner to observe a context each timestep. The regret is then augmented to be the difference between the loss of the algorithm and the best context to action mapping in hindsight, allowing for greater granularity. In contextual combinatorial bandits, the learner observes a context before choosing a combinatorial action with the incurred loss being linear in both the chosen action and the context. We introduce novel estimators for both the semi-bandit and full-bandit feedback settings, with the latter requiring the introduction of four-dimensional tensors, that we rigorously define. We prove the first (nearly optimal) regret bounds in this setting and validate our methods empirically on a synthetic dataset.

Chapter 4 addresses delayed feedback, a common challenge in real-world applications where losses are observed only after a delay. We present a new analysis technique that decomposes the stability term of the regret into the standard stability and regularization terms and a novel delay-dependent component, allowing us to isolate the cost of feedback delay. This leads to the first optimal (up to logarithmic factors) regret bounds for combinatorial semi-bandits and Markov Decision Processes (MDPs) with known transitions, and nearly optimal bounds for linear bandits. For MDPs with unknown transitions, our results match the best-known non-delayed regret bounds and achieve optimal delay-dependency. Empirical evaluations in the combinatorial and linear settings further support our theoretical findings.

Acknowledgements

First, I would like to thank my primary advisor, Nicolò Cesa-Bianchi. I have greatly appreciated your vast expertise in bandits and many other fields, your strategic foresight, and your friendly demeanor. But what I am most thankful for is your immense patience and unwavering support. For me, you were a calm voice in any crisis and despite your many responsibilities you always had time for me. I am deeply grateful for your guidance and support throughout my journey.

Next, I want to thank my co-advisor Gergely Neu. I am thankful for all the time you took to disseminate your knowledge to me, be it one-on-one, in reading-groups or lectures; It is abundantly clear that you were personally invested in fostering my knowledge. At the same time I saw you go out of your way to treat everyone as an equal regardless of their career stages or academic backgrounds, which made me feel intensely accepted.

The third person to thank must be my third supervisor in everything but name, Dirk van der Hoeven. Known to my friends as 'Dircules', you have made me feel at home in Milan from the first day. Your influence is palpable in all of my work, in the way I have learned to approach problems, in the tools I use, and in the presentation of my research. I am very thankful for the opportunity to have learned so much from you and throughly enjoyed the time we had.

All three of you made me the researcher, and to an extend also the person I am today, and I could not have wished for better guidance. Thank you.

I would also like to thank my thesis reviewers Pierre Gaillard and Wouter M. Koolen for their great care in reviewing my work and their encouraging and helpful comments.

Finally, I would like to thank all the beautiful people at the LAILA Lab at UniMi, the Artificial Intelligence and Machine Learning Research Group at UPF and at Amazon Web Services in Berlin that I had the pleasure to meet, that challenged me and my perspectives and that have always welcomed me with open arms.

Contents

Chapter 1

Introduction

We usually don't have all the required information to make a decision. Any reader has surely faced everyday dilemmas like: "Should I bring an umbrella?", "What's the fastest way to get to work?", or "Which dish should I order at this restaurant?" And any reader certainly wants to learn that it will always rain, that the bus is always late and that the spaghetti is always too salty at that one place. When we are not privy to perfect information but face a decision repeatedly, we can mitigate the uncertainty by learning from experience. Online learning provides us with powerful tools to make better decisions over time in a principled manner.

The study of online learning has emerged as a central area in machine learning precisely because of this need to make sequential decisions under uncertainty. Unlike traditional batch learning, where a model is trained once on a fixed dataset, online learning algorithms update incrementally as a dataset is revealed step by step. The foundations of this field were laid in the 1990s, with early work focusing on regret minimization: ensuring that the learner's cumulative performance is nearly as good as the best fixed decision in hindsight. Pioneering algorithms such as the Weighted Majority Algorithm (Littlestone and Warmuth, 1994) and Hedge (Freund and Schapire, 1997) enjoy strong provable performance guarantees in an online setting but with the limiting assumption that the outcomes of actions we did not take are always available. Certainly, that is the case in some problems. If one believes that the weather does not depend on their personal choice of which accessories to bring along, then it is easy to observe if it also would have rained if they had ended up bringing the umbrella. However, in many more problems we only receive feedback on the action we actually have chosen, ordering the spaghetti offers little insight on how the lasagne would have tasted after all.

This more restrictive mode of feedback is known as the bandit setting, where the learner observes bandit feedback. Bandits derive their name from a hypothetical gambler playing slot machines (also called one-armed bandits for their proclivity to relieve costumers of their capital, much like a regular two-armed bandit would)

in a casino, where the gambler clearly does not have access to the counterfactual information of "What would have happened if I played the other machine instead?". Despite their namesake, bandits have originally been conceived for clinical trials (Thompson, 1933) and have been applied in recommendation systems (Silva, Werneck, Silva, Pereira, and Rocha, 2022; Li, Chu, Langford, and Wang, 2011), hyperparameter optimization (Li, Jamieson, DeSalvo, Rostamizadeh, and Talwalkar, 2018), automated resource allocation (Delande, Stolf, Feraud, Pierson, and Bottaro, 2021), finance (Shen, Wang, Jiang, and Zha, 2015) and many more settings (Bouneffouf, Rish, and Aggarwal, 2020).

Online settings can be considered a harder problem than offline settings, after all not having the entire dataset available immediately is a clear disadvantage. However, in the bandit setting the online nature offers a unique opportunity. The learner has direct influence over the data they observe and can thus create a higher quality dataset. The potential to strategically explore does come at a cost however, because each time we take a decision, we must suffer the consequences. Carefully balancing the exploration and thus the quality of the data we have access to, with the desire to exploit actions that perform well is a dichotomy at the core of all bandit algorithms.

In this work we will present multiple bandit algorithms and settings, we will demonstrate how we can design successful algorithms, quantify the difficulty of problems, explore which assumptions are reasonable and how they affect how we learn. This rest of this work is structured as follows:

- The remainder of Chapter 1 is dedicated to going through the core assumptions on bandits in detail, gives an introduction to two famous bandit algorithms: Follow the Leader and Follow the Regularized Leader, and doubles as a survey of relevant literature.

- Chapter 2 covers combinatorial bandits, a bandit setting where the learner is able to play multiple actions concurrently and proves a regret bound for Follow the Regularized Leader in this setting, with the goal of providing a soft introduction for the concepts pivotal to the rest of the work.

- Chapter 3 studies a contextual online combinatorial optimization problem, where a learner selects actions from a combinatorial decision space after observing vector-valued contexts, incurring losses from a bilinear function of context and action vectors, with two feedback variants: semi-bandit (component-level losses) and full-bandit (only total loss). We develop computationally efficient algorithms using a novel loss estimator that leverages the problem's structure, achieving near-optimal regret bounds with polynomial scaling in relevant parameters, supported by experimental validation on simulated data. Chapter 3 is based on Zierahn, Hoeven, Cesa-Bianchi, and Neu (2023).

- Chapter 4 presents a new analysis of Follow The Regularized Leader (FTRL) for online learning with delayed bandit feedback that separates delay costs from bandit feedback costs, enabling optimal regret bounds for combinatorial semi-bandits and adversarial Markov decision processes with delay, as well as an efficient near-optimal algorithm for delayed linear bandits. Chapter 4 is based on Van der Hoeven, Zierahn, Lancewicki, Rosenberg, and Cesa-Bianchi (2023) and Zierahn, Hoeven, Lancewicki, Rosenberg, and Cesa-Bianchi (2025).

Notation Throughout this work, great care was taken to unify and standardize the notation. Vectors are bold lower case letters like $\boldsymbol{a}, \boldsymbol{x}$, matrices are bold upper case letters like $\boldsymbol{A}, \boldsymbol{\Sigma}$ and tensors (where they appear) are denoted by $\boldsymbol{\Psi}$ or $\boldsymbol{\Phi}$. Sets and probability distributions are denoted by upper case calligraphic letters such as $\mathcal{O}$ or $\mathcal{D}$.

These rules should be seen for the guidelines and visual aids that they are. To conform with prior intuition and to follow the conventions in the field, those rules are sometimes bent or broken. To name just two examples: $\mathcal{R}_T \in \mathbb{R}$ denotes the regret and is a real number, not a set or probability distribution and a policy π associated with some Markov Decision Process can often be seen as vector and thus should be bold but is interpreted as a function in this work instead.

Furthermore, $f \in O(g)$ is the big-O notation denoting that f is asymptotically at most as large as g, or more precisely that there exists an c and an x_0 such that for all $x \geq x_0$ it follows that $|f(x)| \leq c|g(x)|$. Similarly, we use $f \in \Omega(g)$ to denote that f is at least as large as g. If f and g are of the same size, that is both $f \in O(g)$ and $f \in \Omega(g)$, then we write $f \in \Theta(g)$. We also write $f \in o(g)$ as the little-o, meaning that f is smaller than g, that is if for every $\epsilon \geq 0$, there exists an x_0 such that for all $x \geq x_0$ it follows that $|f(x)| \leq c|g(x)|$. We also define similarly $f \in \omega(g)$ as f being larger than g.

1.1 Bandits

In this section we aim to establish and justify the set of assumptions defining the core bandit from which all other bandit settings branch off. This setting, also referred to as the Multi-armed Bandit (MAB) setting or just the bandit setting is making a number of convenient assumptions to isolate many of the same underlying difficulties that more sophisticated settings face. Throughout this chapter we will refer to a recurring example of a movie streaming service that is interested in recommending movies to their users to illustrate the assumptions made.

In the MAB setting, the learner is provided an action-set $\mathcal{A}$ of size $K = |\mathcal{A}|$ and plays for T timesteps where in each timestep t the learner selects one action $a_t \in \mathcal{A}$ to play. For the movie streaming service, each user visiting their website is

a single timestep and the action-set is made up out of all the movie available on the website. The assumption that the action set is fixed means that any algorithm will have to be restarted when any movie arrives or leaves the catalogue of the website and the algorithm picks one to recommend to the user.

Sleeping bandits are designed to accommodate changing but still finite action-sets by assuming that the available actions are either drawn i.i.d. each round (Neu and Valko, 2014; Cortes, Desalvo, Gentile, Mohri, and Yang, 2019; Saha, Gaillard, and Valko, 2020) or chosen adversarially (Kleinberg, Niculescu-Mizil, and Sharma, 2010; Kanade and Steinke, 2014; Kale, Lee, and Pal, 2016; Saha and Gaillard, 2021). The linear bandit setting makes a strong linearity assumption on the performance of actions but does allow for changing and infinite action sets, as long as the actions are bounded, we will explore linear bandits in more detail later.

After playing an action $a_t \in \mathcal{A}$, the learner observes the loss $\boldsymbol{\ell}_t(a_t) \in \mathbb{R}$. The goal of the learner is to keep the losses of their actions as small as possible, that is large $\boldsymbol{\ell}_t(a_t)$ are undesirable for the learner. This is a matter of convention and many previous works use rewards $\boldsymbol{r}_t(a_t) \in \mathbb{R}$ instead, where the goal of the learner is to maximize the reward instead of minimizing a loss. Naturally, there are different ways to convert between losses and rewards at runtime but many algorithms are not able to accommodate negative losses or rewards. Together with the fact that upper and lower bounds on losses and rewards are not always available a priori, the choice of dealing with losses or rewards can surprisingly be a meaningful distinction. In this work we will stick to losses, though all algorithms presented here work with negative losses just fine and thus the loss-to-reward conversion is always straightforward.

1.1.1 Regret

The sum of losses the algorithm incurs over the T timesteps is called the cumulative loss and spelled out as

$$\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t) \ .$$

In our movie example, each loss is associated to the quality of the recommendation for that specific user, keeping the cumulative loss small means giving good suggestions to the user on average. It may be tempting to evaluate the performance of any algorithm using the cumulative loss alone. That is celebrating an algorithm as 'good' whenever it can guarantee a 'low' cumulative loss. The issue with that approach is that does not take the overall difficulty of the underlying problem into account. We have not made any assumptions on the losses at all at this point but imagine a setting where only a single action is available i.e., there is only a single movie on your website, which, for a lack of choice, you now have to recommend to everyone. If that action incurs a large loss at each timestep, then the cumulative loss will be large but also entirely unavoidable and thus we should not be too quick

to dismiss any algorithm that suffers large losses in this example as performing poorly. A solution is to normalize the cumulative loss by a baseline performance that we reasonably hope to achieve. Picking the single best action in hindsight as the baseline gives the so called regret

$$\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \boldsymbol{\ell}_t(a) \ .$$

Since the losses might be random, and the learner might randomize, the regret itself is also a random variable and guaranteeing a good performance deterministically, no matter how bad the scenario and how unlucky the learner, is generally not possible. There are two primary ways to deal with regret being random. The first is by guaranteeing that the regret is bounded most of the time, but not always (i.e., it is bounded with high probability) and the second is by adding an expectation over all possible random events in the algorithm and in the generation of the losses. The latter is the expected regret and will be the exclusive focus of this work

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t)\right] - \min_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(a)\right] \ . \tag{1.1}$$

High probability regret has first been introduced by Auer, Cesa-Bianchi, Freund, and Schapire (2002), the modern approach to obtaining high probability bounds has been pioneered by Kocák, Neu, Valko, and Munos (2014) and Neu (2015).

It is important to be aware of the implicit assumptions of this regret definition. One implicit assumption is that the sum of losses is something we want to keep small. Presume that we were not interested in recommending the best movie on average to the user but instead to identify the best movie to recommend as quickly as possible. This is the Best Action Identification (BAI) setting, where the learner has to recommend an action $\hat{a}$ after T timesteps. In this setting, the learner has no incentive at all to keep the cumulative loss small and is instead focused on just exploring the actions. One might wonder how closely the goals are aligned. After all, regret minimizing algorithms will also find the best action, which we can see by dividing the definition of the expected regret 1.1 by T on both sides

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t)\right] - \frac{1}{T}\min_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(a)\right] = \frac{\mathcal{R}_T}{T} \ .$$

If the regret of an algorithm is sub-linear in T, that is $\mathcal{R}_T \in o(T)$, then the right-hand side of the equality will tend to 0 as T grows large. This implies that the average loss the algorithm suffers, compared to the average loss of the best action in hindsight, will also tend to 0 and the algorithm will converge to playing the best action more and more often. So one general recipe to identify the best action is to run a regret minimizing algorithms and at some point recommend the action

the algorithm concentrates on as the best action. However, regret minimizing algorithms spend many episodes playing the best action and do not gather the evidence to reject bad actions quickly enough and it can be shown that this approach is not optimal for finding the best action (Degenne, Nedelec, Calauzenes, and Perchet, 2019; Lattimore and Szepesvári, 2020, Chapter 33, Note 2). All this to say that if the goal is not to keep the cumulative loss small, then there might be better options than just bandit algorithms available.

A second implicit assumption of the regret is that the best fixed action in hindsight is an appropriate baseline for the performance of the algorithm. Returning to the movie recommendation example, we would expect the best performing movie to change over time and it would be natural to desire our algorithm to dynamically keep recommending the best movie at each timestep, tracking the best action. By picking the best fixed action in hindsight as our baseline, we only expect our algorithm to be close in performance to the single best movie however, even if it there are better actions available for large time periods. If we would like to be adaptive to non-stationarity, we can replace the best fixed action in hindsight by the best action at each timestep and define the dynamic regret

$$\mathcal{R}_T^{\text{Dyn}} = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t)\right] - \min_{u_1,\dots,u_T \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(u_t)\right] . \tag{1.2}$$

Since we have strictly increased the size of our comparator class, it is perhaps unsurprising that bounding the dynamic regret will require stronger assumptions than the ones sufficient for the expected regret. Common approaches include to assume that losses are not moving too much over time as for example in Zhao, Zhang, Jiang, and Zhou (2020) or to assume an environment that is stationary for longer time periods and then apply changepoint detection methods on top of bandit algorithms, see (Hartland, Gelly, Baskiotis, Teytaud, and Sebag, 2006; Mellor and Shapiro, 2013; Alami, 2023).

We finish this chapter by exploring what it means to be able to learn a bandit problem. If we have a bound on the expected loss for each action a, $|\,\mathbb{E}\,[\boldsymbol{\ell}_t(a)]\,| \leq c$ then both the regret and the dynamic regret will grow at most linearly in T, that is

$$\mathcal{R}_T \leq 2c \cdot T \qquad \text{and} \qquad \mathcal{R}_T^{\text{Dyn}} \leq 2c \cdot T \ ,$$

no matter what any algorithm is doing. That means that even nonsensical algorithms such as always picking one action or picking actions entirely at uniform random achieve a $O(T)$ regret bound. A weak notion of an algorithm that is able to learn is the Hannan consistency (Hannan, 1957), where an algorithm is called Hannan consistent if it achieves

$$\limsup_{T\to\infty} \frac{\mathcal{R}_T}{T} \leq 0 \ ,$$

meaning it has sub-linear regret. Hannan consistency is aligned with the earlier observation that if an algorithm has sub-linear regret in T, then a regret minimization algorithm concentrates around playing the best action in hindsight. In the next chapter we will explore what kind of assumptions on the losses are required to allow algorithms to achieve a sub-linear regret in bandits.

1.1.2 Losses

A critical component of the problem that we have not explored much at all yet are the losses itself, which are critical for designing and applying algorithms. In our movie example the loss might be binary: A 0 if the user watched the movie we recommended and a 1 if they did not. Or perhaps the user can give more fine grained feedback and select a number between 1 and 10 as their score for the movie. Yet another choice for feedback may be the number of minutes spent on the website immediately after the movie was recommended, which can be a real number that may not be bounded. For each of these losses, the assumption that the loss at each timestep is a single sample from an independent and identically distribution might be reasonable for short time frames but since public opinion is sure to change over time, the distribution will not stay identical. And if many users have seen a movie already, they are probably less inclined to watch it a second time, leading to correlated samples for longer timeframes. While the exact properties of the losses can change drastically from setting to setting, the bandit literature differentiates between two broad settings in which learning is possible: stochastic bandits and adversarial bandits.

Adversarial bandits assume that losses are bounded, usually $\boldsymbol{\ell}_t(a) \in [0, 1]$ for all $t \in [T]$ and $a \in \mathcal{A}$ but make almost no further assumptions on how the losses are picked. To show that an algorithm has sub-linear regret in this setting, it must simultaneously guarantee a sub-linear regret on all possible deterministic sequences of losses, which also includes exactly those sequences of losses that happen to be especially challenging for that particular algorithm. In that sense it almost feels like the losses are hand-picked by an adversary to throw the algorithm off on purpose, which is where this setting derives its name from. Depending on the setting, the adversary has to design the entire loss sequence a priori (the oblivious adversary), or can even chose losses adaptively based on the past actions the learner played (the non-oblivious adversary). In either case the adversary has full knowledge of algorithm but it is assumed that the adversary does not posses precognition and cannot predict the future moves of the learner. That means, if the learner randomizes over actions that the adversary has full access to the probability distribution the learner samples from before picking a loss but not the actual action the learner will play.

Boundedness is a critical assumption in this adversarial setting. Without the boundedness the adversary could assign losses of order T to all actions except one in the first round and assign a loss of 0 for all actions afterwards. Since we have no information in the first round, the best we can do is pick any action at random, which incurs an expected regret of order $O(T)$. Though it is possible to get some weaker results with unbounded losses (Allenberg, Auer, Györfi, and Ottucsák, 2006), bounded losses are a common assumption to avoid this kind of scenario. The adversary can still try to make the learner incur a large cumulative loss while hiding a good action but to have a meaningful impact on regret, the adversary has to follow that strategy for a significant fraction of the timesteps, which the learner can then pick up on. If the losses for all actions are available to the learner after playing an action, that is if we are in the so-called full information setting, then regret bounds of order $O(\sqrt{T\log(K)})$ are possible (Littlestone and Warmuth, 1994; Freund and Schapire, 1997). Since there also exist lower bounds of order $\Omega(\sqrt{T\log(K)})$ in this setting, we know we cannot do better and the algorithms achieving $O(\sqrt{T\log(K)})$ are asymptotically optimal.

It is important to keep in mind what those regret guarantees mean. In adversarial bandits the losses are able to change arbitrarily between timesteps, which can lead to the impression that algorithms designed for the adversarial bandit setting are a good fit for a non-stationary environments but that is only partly true. As we have discussed in the previous section, when using the regret, the comparator is the best fixed action in hindsight and if an algorithm ought to track the best action as it changes over time, then a regret measure such as dynamic regret in Equation (1.2) is more appropriate as adversarial bandits are not adaptive to a non-stationarity. However, the weak assumptions on the losses mean that the regret guarantees of adversarial bandits hold for non-stationary environments, and so adversarial bandits are resistant to non-stationarity but are not able to exploit it.

The adversarial setting will be the exclusive focus of this manuscript.

Stochastic bandits are the second big category of bandit problems and assume that the losses of each action are independent and identical distributed (i.i.d.). Because the distribution of the actions does not change, the best action (in expectation) also does not change and thus using the best fixed action in hindsight as the comparator in the regret, like in Equation (1.1) really is the right thing to do here. In addition to the i.i.d. assumption, many algorithms require the loss distributions to be well behaved, some form of sub-gaussian losses are a common assumption (for example Auer, Cesa-Bianchi, and Fischer (2002) and Abbasi-yadkori, Pál, and Szepesvári (2011)) though significant work has also been done on heavy tailed distributions (Bubeck, Cesa-Bianchi, and Lugosi, 2012; Agrawal, Juneja, and Koolen, 2021). Under the sub-gaussian assumption, the Upper Confidence Bound (UCB)

Require: Action set $\mathcal{A}$
1: **for** $t = 1, \dots, T$: **do**
2: Find the best action in hindsight

$$a_t \in \arg\min_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \boldsymbol{\ell}_s(a_s) \ ,$$

breaking ties arbitrarily. At $t = 1$, we define $\sum_{s=1}^{t-1} \boldsymbol{\ell}_s(a_s) = 0$.
3: Play a_t, suffer $\boldsymbol{\ell}_t(a_t)$, observe $\boldsymbol{\ell}_t$
4: **end for**

Algorithm 1: Follow The Leader (FTL)

algorithm achieves an optimal regret bound of $O(\frac{K \log(T)}{\Delta})$ (Auer, Cesa-Bianchi, and Fischer, 2002), where Δ is the expected performance distance between the best and second best action. Bounds independent of Δ are also possible, a modification of UCB achieves an optimal regret bound of $\sqrt{TK}$ (Audibert and Bubeck, 2009), matching the bounds in the adversarial setting. While algorithms designed for adversarial bandits can be applied to stochastic bandits, as long as the loss distributions are bounded, the reverse is not the case and there is no reason to expect an algorithm designed for stochastic bandits to achieve any reasonable performance as soon as the i.i.d. assumption is violated. For that reason stochastic bandits are usually perceived as a stronger assumption than adversarial bandits though stochastic bandits can usually accommodate potentially unbounded losses, which are a pain-point for adversarial bandits.

1.2 Follow The Leader & Follow The Regularized Leader

In this section we discuss what kind of methods can achieve sub-linear regret in the adversarial bandit setting. That is we have bounded losses $\boldsymbol{\ell}_t(a) \in [0, 1]$, which are picked by a non-oblivious adversary and we aim to minimize the regret as given in Equation (1.1). We will also stay in the full information setting, that is we assume that $\boldsymbol{\ell}_t(a)$ is revealed to us for all $a \in \mathcal{A}$ after having played our action a_t at timestep t.

Since our comparator for the regret is the best fixed action in hindsight, a first idea for our own algorithm might be to simply play the best current action at each timestep, that is we pick

$$a_t \in \arg\min_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \boldsymbol{\ell}_s(a_s) \ .$$

This algorithm is called Follow The Leader (FTL) (Algorithm 1), originally due to

Hannan (1957) and is not able to achieve a sub-linear regret in this setting. As briefly touched on in the previous section, the regret guarantee has to simultaneously hold for all possible loss sequences, so we can put ourselves in the shoes of the adversary and ask ourselves "What sequences of losses will this algorithm struggle with the most?". Our construction of losses is inspired by Shalev-Shwartz (2012, Example 2.2) and uses just two actions:

$$\boldsymbol{\ell}_t(a_1) = \begin{cases} \frac{1}{2}, & \text{if } t = 1 \\ 0, & \text{if } t \text{ is even} \\ 1, & \text{otherwise} \end{cases} \qquad \boldsymbol{\ell}_t(a_2) = \begin{cases} 0, & \text{if } t = 1 \\ 1, & \text{if } t \text{ is even} \\ 0, & \text{otherwise} \end{cases} \tag{1.3}$$

The action picked by FTL in the first round doesn't matter. In the second round FTL will have observed a loss of $\frac{1}{2}$ for action a_1 and a loss of 0 for action a_2. The best action in hindsight thus is a_2, which is the action FTL will play in round $t = 2$. The learner now suffers a loss of $\boldsymbol{\ell}_2(a_2) = 1$ but also observes the loss of the other action, $\boldsymbol{\ell}_2(a_1) = 0$. The cumulative loss for action a_1 remains unchanged at $\frac{1}{2}$ but the cumulative loss for action a_2 has increased to 1. The best action to play in hindsight is thus action a_1, which will net another loss of $\boldsymbol{\ell}_3(a_1) = 1$ in the next timestep. This pattern continues and FTL suffers a cumulative loss of at least $T - 1$. Sticking to either action suffers a cumulative loss of $\frac{T-1}{2}$ at most and thus the performance of the comparator is also $\frac{T-1}{2}$. Together that gives that the regret of FTL in this setting after T timesteps is at least

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \boldsymbol{\ell}_t(a)\right] \geq T - 1 - \frac{T-1}{2} = \frac{T-1}{2} \,.$$

Since FTL suffers linear regret in this setting, it is not able to learn and this is reflected in FTL being continuously tricked by the adversary into perpetually collecting the largest loss available. Given this lower bound, there is no fundamental reason why FTL should be preferable to, for example picking actions at random.

This lower bound exploits two critical shortcomings of FTL in the adversarial setting. The first is that FTL is predictable, we know exactly what the algorithm will do given the history, which allows us to construct losses that always happen to be bad for FTL. In fact, any algorithm that achieves a sub-linear regret in adversarial bandits must randomize (Lattimore and Szepesvári, 2020, Chapter 11).

The second shortcoming this lower bound exploits is that FTL can entirely change its mind on which action to play given very little additional data. An algorithm can suggest to play any probability distribution over actions but FTL jumps from playing one action with probability 1, to playing the other action with probability 1 after just observing one more loss. FTL is overfitting to the data. A standard response to overfitting is to introduce some form of regularization, that is what Follow The Regularized Leader (FTRL) (Gordon, 1999; Shalev-Shwartz,

2007; Shalev-Shwartz and Singer, 2007) does. We define

$$\boldsymbol{w}_t \in \arg\min_{\boldsymbol{v}\in\Delta_{\mathcal{A}}} \left(\sum_{s=1}^{t-1} \boldsymbol{\ell}_s\right)^\top \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ , \tag{1.4}$$

and then sample our action as prescribed by $\boldsymbol{w}_t$, which conveniently also makes our algorithm randomize. The added regularization term consists of two components, the (inverse of the) learning rate η and a convex regularization function $R(\boldsymbol{v})$. The learning rate trades off exploiting the information we already gathered with the stability enforced through the regularization that keeps us from overfitting. When $\eta \to \infty$, the regularization becomes less and less important and we recover FTL. Conversely, if $\eta \to 0$ the regularization term dominates and the algorithm becomes maximally stable by ignoring all data and always playing the point minimizing the regularization function (usually the uniform distribution over actions). In many cases, the right η is of order $\eta \approx \frac{1}{\sqrt{T}}$, where T is the total number of rounds played. Since the sum of losses, i.e., the first term in the argmin, can scale with T and the regularization only scales with $\sqrt{T}$, the contribution of the losses for the argmin is able to overcome the regularization. The regularization function itself must be convex and can induce a wide range of different behaviors. Using the negative entropy $R(\boldsymbol{v}) = -\sum_{a\in\mathcal{A}} \boldsymbol{v}_a \log(\boldsymbol{v}_a)$ as the regularization recovers Exponential Weights (Vovk, 1990; Littlestone and Warmuth, 1994) and Exp3 (Auer, Cesa-Bianchi, Freund, and Schapire, 2002), both well known algorithms. Using the log barrier $R(\boldsymbol{v}) = -\sum_{a\in\mathcal{A}} \log(\boldsymbol{v}_a)$ has been used in a reinforcement learning setting to enable the algorithm to deal with negative losses (Dai, Luo, Wei, and Zimmert, 2023) and using an L_2-norm as regularization, that is $R(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2$, recovers (online) gradient descent (Lattimore and Szepesvári, 2020, Example 28.1). FTRL thus can also be seen as a collection of algorithms, where the nature of the algorithm depends on the regularization function.

FTRL is able to achieve the optimal rates for both adversarial and stochastic bandits that were mentioned in the previous section and is the predominant algorithm used in this work. We explore how to prove regret bounds for FTRL in Section 2.1.1.

1.3 Bandit Feedback

So far we've assumed that the entire loss vector $\boldsymbol{\ell}_t$ is always available to the learner after playing an action. That is, the learner always had answers to the question "What would've happened if I played this other action instead?". In many real world scenarios this assumption is violated. Returning to our motivating example, if we recommend one movie to a visitor, we have no idea if they would've clicked on any other movie or not. Only observing the loss of the played action $\boldsymbol{\ell}_t(a_t)$ is called bandit feedback, while observing the entire loss vector $\boldsymbol{\ell}_t$ is the full

information setting and also called prediction with expert advice. Plenty of work has been done in both settings and in the in-between including defining feedback graphs between actions (Mannor and Shamir, 2011; Cohen, Hazan, and Koren, 2016; Esposito, Fusco, Hoeven, and Cesa-Bianchi, 2022) and decoupling the observation and the action played entirely (Avner, Mannor, and Shamir, 2012; Rouyer and Seldin, 2020).

Despite not being able to observe the losses of the actions we did not play, it is still possible to reason about them. The plan is to build an unbiased estimator of the entire loss vector $\hat{\boldsymbol{\ell}}_t$ with just the single observation available to us and then simply replace the losses in Equation (1.4) with their estimates as follows

$$\boldsymbol{w}_t \in \arg\min_{\boldsymbol{v}\in\Delta_{\mathcal{A}}} \left(\sum_{s=1}^{t-1} \hat{\boldsymbol{\ell}}_s\right)^\top \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ , \tag{1.5}$$

We use the importance weighted estimator (Goertzel, U.S. Atomic Energy Commission, and Oak Ridge National Laboratory, 1950; Kahn and Harris, 1951; Kloek and Dijk, 1978), given by

$$\hat{\boldsymbol{\ell}}_t(a) = \frac{\mathbb{I}[a_t = a]\boldsymbol{\ell}_t(a)}{\mathbb{P}(a_t = a \mid \mathcal{F}_t)} \ , \tag{1.6}$$

where $\mathcal{F}_t$ is the filtration over all past randomness up to, but not including timestep t and $\mathbb{I}[\cdot]$ is the indicator function which takes value 1 if the expression in the brackets is true and 0 otherwise. The loss estimate $\hat{\boldsymbol{\ell}}_t(a)$, given all the history before timestep t, is a random variable, however all the randomness comes from a_t as $\boldsymbol{\ell}_t$, a, and $\mathbb{P}(a_t = a \mid \mathcal{F}_t)$ are non-random. $\mathbb{P}(a_t = a \mid \mathcal{F}_t)$ is the probability that the algorithm plays action a at timestep t given all the information that is available at the start of timestep t. When using FTRL we sample our actions using $\boldsymbol{w}_t$ as defined in Equation (1.5) and thus $\mathbb{P}(a_t = a \mid \mathcal{F}_t) = \boldsymbol{w}_t(a)$. First we show mechanically that $\hat{\boldsymbol{\ell}}_t(a)$ is an unbiased estimator and then we will give some additional intuition. We start by taking an expectation over a_t given the history $\mathcal{F}_t$ and then write the expectation as a sum

$$\mathop{\mathbb{E}}_{a_t}[\hat{\boldsymbol{\ell}}_t(a) \mid \mathcal{F}_t] = \sum_{a'\in\mathcal{A}} \mathbb{P}(a_t = a' \mid \mathcal{F}_t) \cdot \frac{\mathbb{I}[a' = a]\boldsymbol{\ell}_t(a)}{\mathbb{P}(a_t = a \mid \mathcal{F}_t)} = \boldsymbol{\ell}_t(a) \ ,$$

where the last equality follows by recognizing that $\mathbb{I}[a' = a]$ is zero for all elements of the sum except for the one where $a' = a$ and the probabilities cancel.

The only loss we have access to is $\boldsymbol{\ell}_t(a_t)$ and the indicator in the numerator of the estimator makes sure that $\boldsymbol{\ell}_t(a)$ for all a that are not a_t are multiplied by 0, meaning that we can compute $\hat{\boldsymbol{\ell}}_t(a)$ without observing those losses and $\hat{\boldsymbol{\ell}}_t(a)$ is a valid estimator in the bandit setting. This also means that $\boldsymbol{\ell}_t(a)$ will be 0 for all $a \neq a_t$, and since $\mathbb{P}(a_t = a \mid \mathcal{F}_t) < 1$, we scale up the one non-zero entry of $\hat{\boldsymbol{\ell}}_t(a)$

to compensate. If we happen to pick an actions that was less likely, we scale up more. This can lead to large variances, especially because $\mathbb{P}(a_t = a \mid \mathcal{F}_t)$ can be arbitrarily close to 0, which in turn means that $\frac{1}{\mathbb{P}(a_t=a|\mathcal{F}_t)}$ can be unbounded. The upshot is that $\mathbb{P}(a_t = a \mid \mathcal{F}_t)$ is fully controlled by the algorithm and we can bound it away from 0, for example by forcing the algorithm to play each action with some probability for an additional cost in the regret. Still, controlling the variance of the importance weighted estimator is usually a major challenge in the regret analysis. In the previous Section 1.2, we argued that an algorithm must be stable to have low regret. If we happen to play an action that has low probability and in turn observe a very large $\hat{\boldsymbol{\ell}}_t(a)$, even FTRL might start to change it's recommendation entirely based on that single large observation. Large variances of the loss estimator are problematic because they force us to to make the algorithm more stable, usually by tuning the learning rate to increase the strength of the regularization.

With the correct regularization FTRL is able to obtain a regret bound of $O(\sqrt{KT})$ (Audibert and Bubeck, 2009; Audibert and Bubeck, 2010) in the adversarial setting with bandit feedback, which matches the lower bound and thus is optimal. Because FTRL sits at the heart of this manuscript, we dedicate the next chapter to giving insights on how to prove a regret bound for FTRL. We use a different regularization, which obtains a slightly worse bound of order $\sqrt{KT\log(K)}$ for MAB and prove the bound for the combinatorial bandit setting, though this regret bound for the adversarial bandit setting laid out here can be derived from those results.

Chapter 2

Combinatorial Bandits and Follow-the-Regularized-Leader

2.1 Combinatorial Bandits

In this chapter we will introduce combinatorial bandits and show how one can proof a regret bound using FTRL in this setting. Combinatorial bandits are an extension to the basic bandit setting, which have been introduced by Cesa-Bianchi and Lugosi (2012a) and which aim to deal with exponentially large action sets by assuming a linear structure on the losses. Combinatorial bandits have been particular impactful on this work and will reappear in Chapters 3 and 4.

Returning to the example of the movie streaming website once more, it seems reasonable that we will recommend a handful of movies at a time, instead of just recommending a single one. Say we can recommend m movies out of all K available movies. In this context we will call each of the $\binom{K}{m}$ possible combinations of movies an action, and each of the K movies the learner can chose to recommend a sub-action. It is possible to simply apply an algorithm for adversarial bandits to the actions, which yields a regret bound of order $O\left(\sqrt{\binom{K}{m}T}\right)$. But since $\binom{K}{m} \geq \frac{K^m}{m^m}$, this introduces an exponential dependency on the number of actions, which can be very large.

Combinatorial bandits assume that the loss of each action is the sum of the losses of active sub-actions, that is for every action $\boldsymbol{a} \in \{0,1\}^K$, the loss is given as

$$\boldsymbol{\ell}_t(\boldsymbol{a}) = \sum_{i=1}^{K} \boldsymbol{\ell}_{t,i}\boldsymbol{a}_i = \boldsymbol{\ell}_t^\top \boldsymbol{a} \ ,$$

where $\boldsymbol{\ell}_{t,i} \in [0,1]$ is the loss of sub-action i and $\boldsymbol{a}_i$ indicates if sub-action i is active in action $\boldsymbol{a}$. Which combinations of sub-actions we can recommend is problem dependent. In the movie example the learner is supposed to recommend m movies

each timestep, and thus the action set is the set of all vectors that have exactly m active sub-actions

$$\mathcal{A}^{\text{movies}} = \{\boldsymbol{a} \in \{0,1\}^K : \|\boldsymbol{a}\|_1 = m\} \ .$$

This specific set of actions is also called m-Sets, though the action set for combinatorial bandits can be any arbitrary set of vectors in $\{0,1\}^K$.

The assumption on the structure of the losses is a strong assumption. It implies independence between sub-actions, that is including or not including sub-actions i may not affect the loss of any other sub-actions. An assumption that is likely violated in the movie recommendation example! After all, if a user decides to watch one movie, they will likely discard all other movies that were recommended. If this assumption of independence does not hold, we have no choice but to treat each of the $\binom{K}{m}$ actions individually and applying an algorithm for adversarial bandits over the actions is optimal.

An example where the independence between sub-actions is more believable is a shortest path problem. Consider a loop-free directed graph, where each edge has a cost of traversal. The source of the graph is the starting position and the sink is the goal, the interpretation of vertices and edges depends on the the problem but vertices might be intersections or servers and the edges might be roads or ethernet and wireless connections, in routing a single car or network traffic respectively. The loss associated with each edge could then be related to driving time on a road and fuel costs or the connection speed between the servers. An action in this setting is then the entire path between source and sink with the edges being the sub-actions. In this case it is plausible that the losses of sub-actions are independent, that is, the loss of the next road does not depend on the path one took to arrive at the last intersection.

The final component of the combinatorial bandit setting is the feedback the learner observes. In the semi-bandit case, the learner is able to observe the losses of all sub-actions the learner played $\boldsymbol{\ell}_t \odot \boldsymbol{a}_t$, where $\odot$ is the element-wise multiplication of two vectors (also called the Hadamard product). The full-bandit setting is more restrictive and the learner is only able to observe the sum of losses the learner suffers $\boldsymbol{\ell}_t^\top \boldsymbol{a}_t$. The difference in the shortest path example is receiving a road by road breakdown of the travel time in the semi-bandit setting or just observing the entire duration of the trip for full-bandits.

Putting all those components together gives the combinatorial bandits learning framework

1. The non-oblivious adversary picks a loss vector $\boldsymbol{\ell}_t \in \mathbb{R}^K$
2. The learner picks an action from the action set $\boldsymbol{a}_t \in \mathcal{A} \subseteq \{0,1\}^K$
3. The learner incurs the loss $\boldsymbol{\ell}_t^\top \boldsymbol{a}_t$
4. The learner observes:
   full-bandit: $\boldsymbol{\ell}_t^\top \boldsymbol{a}_t$
   semi-bandit: $\boldsymbol{\ell}_t \odot \boldsymbol{a}_t$

where $\odot$ is the element-wise multiplication of two vectors.

2.1.1 FTRL for Combinatorial Bandits

In this section we will give strong intuition on how to prove a regret bound for combinatorial bandits using FTRL. As argued in the previous section, simply using FTRL as introduced in Section 1.2 yields a linear dependency on the number of actions, which in turn can be an exponential dependency on the number of sub-actions K, which we are trying to avoid. The optimization problem at the core of FTRL (Equation (1.5)) is on the simplex over actions $\Delta_{\mathcal{A}}$. However, the simplex over actions is a $\mathcal{A} - 1$ dimensional space while the adversary is restricted to the smaller sub-action space which is only K dimensional. There exists a clear relationship between these spaces, which we find by examining the expected regret that will be incurred with a given loss vector $\boldsymbol{\ell}$ when picking a distribution over actions $\boldsymbol{p} \in \Delta_{\mathcal{A}}$

$$\mathop{\mathbb{E}}_{\boldsymbol{a}\sim\boldsymbol{p}}[\boldsymbol{a}^\top \boldsymbol{\ell}] = \left(\mathop{\mathbb{E}}_{\boldsymbol{a}\sim\boldsymbol{p}}[\boldsymbol{a}]\right)^\top \boldsymbol{\ell} = \boldsymbol{x}_{\boldsymbol{p}}^\top \boldsymbol{\ell} \ ,$$

where we defined $\boldsymbol{x}_{\boldsymbol{p}} = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}}[\boldsymbol{a}]$. The vector $\boldsymbol{x}_{\boldsymbol{p}} \in \mathrm{Conv}(\mathcal{A})$ now encodes the chance of playing the sub-actions, that is the element at index $i \in [K]$ of $\boldsymbol{x}_{\boldsymbol{p}} \in [0,1]^K$ gives the chances of playing sub-action i when sampling actions using $\boldsymbol{p} \in \Delta_{\mathcal{A}}$. Using this relationship we can map each point on the simplex over actions $\Delta_{\mathcal{A}}$ to a point in $\mathrm{Conv}(\mathcal{A}) \subseteq [0,1]^K$. Multiple different distribution over actions might map to the same $\boldsymbol{x}$ but it follows by the above equation that all of those distributions over actions will have the same expected loss. With that we can move the optimization problem in Equation (1.5) to the lower dimensional sub-action space

$$\boldsymbol{w}_t \in \mathop{\arg\min}_{\boldsymbol{v}\in\mathrm{Conv}(\mathcal{A})} \left(\sum_{s=1}^{t-1} \hat{\boldsymbol{\ell}}_s\right)^\top \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ , \tag{2.1}$$

where $\hat{\boldsymbol{\ell}}_s$ is now an estimator of the sub-action losses, which we will explore more momentarily.

The algorithm will still need to sample an action each timestep, which means we need to translate the output of the optimization $\boldsymbol{w}_t$, which is also in sub-action space back into action space. For that we move in the opposite direction and find one (of the possibly multiple) distribution over actions $\boldsymbol{p}_t \in \Delta_K$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t$ and then sample our action $\boldsymbol{a}_t \sim \boldsymbol{p}_t$. Often, but not always, this reverse transformation can be implemented efficiently, both the m-Sets from the movie example and the shortest path graphs are examples where an efficient implementation is known. A more thorough discussion on when we can find $\boldsymbol{p}_t$ efficiently can be found in Section 4.4 and the references therein.

Require: Action set $\mathcal{A}$

1: **for** $t = 1, \dots, T$: **do**

2: Find the best regularized assignment of sub-actions in hindsight that is permitted by the actionset

$$\boldsymbol{w}_t = \operatorname*{arg\,min}_{\boldsymbol{v} \in \operatorname{Conv}(\mathcal{A})} \left(\sum_{s=1}^{t-1} \hat{\boldsymbol{\ell}}_s \right)^\top \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ .$$

3: Find distribution $\boldsymbol{p}_t$ over actions that plays $\boldsymbol{w}_t$ in expectation, that is $\mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t$

4: Sample and play $\boldsymbol{a}_t \sim \boldsymbol{p}_t$, suffer $\boldsymbol{\ell}_t^\top \boldsymbol{a}_t$, observe $\boldsymbol{\ell}_t \odot \boldsymbol{a}_t$

5: Compute loss estimate to be used in the next round

$$\hat{\boldsymbol{\ell}}_{t,i} = \frac{\boldsymbol{a}_{t,i} \boldsymbol{\ell}_{t,i}}{\boldsymbol{w}_{t,i}} \ .$$

6: **end for**

Algorithm 2: FTRL for Combinatorial Bandits

The estimators used in the previous section (Equation (1.6)) only estimate the loss of each action but to compute $\boldsymbol{w}_t$ we now need estimators of the losses of each sub-action. Which loss estimator is available depends on the information the learner is privy to. We focus on the semi-bandit setting where the learner observes the losses of each sub-action the algorithm selected. Then we define the importance weighted loss estimator in this setting as

$$\hat{\boldsymbol{\ell}}_{t,i} = \frac{\boldsymbol{a}_{t,i} \boldsymbol{\ell}_{t,i}}{\boldsymbol{w}_{t,i}} \ . \tag{2.2}$$

where the $\boldsymbol{a}_{t,i}$ takes on the role of the indicator function of Equation (1.6) that multiplies all losses that we are not able to observe by 0, making this a valid estimator. Just like the previous importance weighted estimator, this estimator scales the losses of the sub-action by the inverse of the probability of observing that sub-action, which is conveniently given by $\boldsymbol{w}_t$, by the way we constructed our algorithm. We also show that it is unbiased when conditioning on the entire past history $\mathcal{F}_t$ by

$$\mathbb{E}[\hat{\boldsymbol{\ell}}_{t,i} \mid \mathcal{F}_t] = \mathbb{E}\left[\frac{\boldsymbol{a}_{t,i} \boldsymbol{\ell}_{t,i}}{\boldsymbol{w}_{t,i}} \mid \mathcal{F}_t \right] = \frac{\mathbb{E}\left[\boldsymbol{a}_{t,i} \mid \mathcal{F}_t\right] \boldsymbol{\ell}_{t,i}}{\boldsymbol{w}_{t,i}} = \boldsymbol{\ell}_{t,i} \ .$$

Putting these pieces together gives the full algorithm that can be found in Algorithm 2 and we will pick the exact regularization R we will use later.

2.1.2 Regret Analysis

In this section we are looking to give some strong intuition on how to prove a regret bound for FTRL algorithms in general and for combinatorial bandits in particular and thus we will occasionally omit the more technical details of the analysis. The analysis will broadly follow the analysis of Chapter 4 and will refer to the more in-depth explanation of that chapter where appropriate.

We start by restating the regret in our setting

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

where $\boldsymbol{a}^*$ is the best action in hindsight given by

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} \boldsymbol{a}^\top \boldsymbol{\ell}_t \ .$$

For reasons that will become apparent later, we will not be able to compare to $\boldsymbol{a}^*$ directly, instead we use another comparator $\boldsymbol{u} \in \mathrm{Conv}(\mathcal{A})$, which also remains fixed throughout time and which we will specify later. Then we start from a regret like term using $\boldsymbol{u}$ and we apply the tower rule of expectations which allows us to pull the conditional expectation inside of the sum and replace $\boldsymbol{a}_t$ by $\boldsymbol{w}_t$

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{a}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t \mid \mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t - \boldsymbol{u})^\top \hat{\boldsymbol{\ell}}_t\right] ,
\end{aligned}$$

where we also replaced $\boldsymbol{\ell}_t$ by its estimator $\hat{\boldsymbol{\ell}}_t$ in the last step, which we are able to do as $\hat{\boldsymbol{\ell}}_t$ is unbiased and because $\boldsymbol{w}_t$ is independent of $\hat{\boldsymbol{\ell}}_t$ as we are not using $\hat{\boldsymbol{\ell}}_t$ to compute $\boldsymbol{w}_t$. Why we had to replace the loss by its estimate will become apparent momentarily. There is no obvious way of relating $\boldsymbol{w}_t^\top \hat{\boldsymbol{\ell}}_t$ to $\boldsymbol{u}^\top \hat{\boldsymbol{\ell}}_t$ directly. Instead, we introduce a third point $\boldsymbol{w}_{t+1}^\top \hat{\boldsymbol{\ell}}_t$, which can be interpreted as "How much loss would we suffer if we use FTRL but knew the loss of the next round in advance?". Now, $\boldsymbol{w}_{t+1}^\top \hat{\boldsymbol{\ell}}_t$ and $\boldsymbol{a}^{*,\top} \hat{\boldsymbol{\ell}}_t$ will be close because $\boldsymbol{w}_{t+1}$ is cheating and thus can beat the performance of any fixed action in hindsight (up to regularization) and $\boldsymbol{w}_t^\top \hat{\boldsymbol{\ell}}_t$ and $\boldsymbol{w}_{t+1}^\top \hat{\boldsymbol{\ell}}_t$ won't be too far apart either because of the stability of the algorithm, as the point FTRL recommends us to play does not move too much between rounds if we regularize enough. Adding and subtracting $\boldsymbol{w}_{t+1}^\top \hat{\boldsymbol{\ell}}_t$ to the above gives rise to

the following regret decomposition

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_t-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\right]=\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}_t-\boldsymbol{w}_{t+1})^\top\hat{\boldsymbol{\ell}}_t\right]}_{\text{stability term}}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{w}_{t+1}-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\right]}_{\text{cheating regret}}.$$

We do the cheating regret term first and we start with an observation on a term similar to the cheating regret. Let $t\in[T]$ and $\boldsymbol{v}\in\operatorname{Conv}(\mathcal{A})$, then

$$\begin{aligned}\sum_{s=1}^{t-1}(\boldsymbol{w}_t-\boldsymbol{v})^\top\hat{\boldsymbol{\ell}}_s&=\underbrace{\sum_{s=1}^{t-1}\boldsymbol{w}_t^\top\hat{\boldsymbol{\ell}}_s+\frac{1}{\eta}R(\boldsymbol{w}_t)-\left(\sum_{s=1}^{t-1}\boldsymbol{v}^\top\hat{\boldsymbol{\ell}}_s+\frac{1}{\eta}R(\boldsymbol{v})\right)}_{\leq 0}+\frac{1}{\eta}(R(\boldsymbol{v})-R(\boldsymbol{w}_t))\\&\leq\frac{1}{\eta}(R(\boldsymbol{v})-R(\boldsymbol{w}_t))\ ,\end{aligned}$$

where the equality is simply adding and subtracting the regularization terms and the inequality holds due to the optimality of $\boldsymbol{w}_t$ as defined in Equation (2.1). We were only able to use the optimality of $\boldsymbol{w}_t$ in this fashion because we are using the same loss estimators $\hat{\boldsymbol{\ell}}_s$ here that also appear in the definition of $\boldsymbol{w}_t$. That is the reason why we had to replace the losses by their estimates earlier. The above equation alone is not able to bound the cheating regret, however it does illustrate that using $\boldsymbol{w}_{t+1}$ in round t lacks behind the best fixed choice by only a factor related to the regularization. And applying the above inequality repeatedly in a recursive fashion yields that

$$\text{cheating regret}=\sum_{t=1}^{T}(\boldsymbol{w}_{t+1}-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\leq\frac{1}{\eta}(R(\boldsymbol{u})-R(\boldsymbol{w}_1)).\tag{2.3}$$

This is the Be-The-Leader Lemma, a well known result a form of which we will also be proven in Chapter 4 (Lemma 73). We will return to this term after picking our exact regularizer.

The stability term is more involved. Some readers might be familiar with Hölder's inequality, which covers a lot of different spaces and settings, but which also shows that for all $p,q\in[1,\infty]$ such that $\frac{1}{p}+\frac{1}{q}=1$ and for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^K$

$$\boldsymbol{x}^\top\boldsymbol{y}\leq\|\boldsymbol{x}\|_p\|\boldsymbol{y}\|_q\ ,$$

where $\|\boldsymbol{x}\|_p=\sqrt[p]{\sum_{i=1}^{K}\boldsymbol{x}_i^p}$ is an L_p norm. We can extend this concept to matrix norms as follows.

Lemma 1 (Hölder for Matrices)**.** *Let $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^K$ be vectors and $\boldsymbol{A}\in\mathbb{R}^{K\times K}$ be a symmetric and positive definite matrix, then*

$$|\boldsymbol{x}^\top\boldsymbol{y}|\leq\|\boldsymbol{x}\|_{\boldsymbol{A}}\|\boldsymbol{y}\|_{\boldsymbol{A}}^*\ ,$$

where $\|\boldsymbol{x}\|_{\boldsymbol{A}}=\sqrt{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}$ and $\|\boldsymbol{x}\|_{\boldsymbol{A}}^*=\sqrt{\boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x}}$

Proof. The proof follows almost directly by Cauchy-Schwartz

$$|\boldsymbol{x}^\top \boldsymbol{y}| = |\boldsymbol{x}^\top \boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{A}^{\frac{1}{2}} \boldsymbol{y}| \leq \sqrt{\boldsymbol{x}^\top \boldsymbol{A}^{-1} \boldsymbol{x}} \sqrt{\boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y}} = \|\boldsymbol{x}\|_{\boldsymbol{A}} \|\boldsymbol{y}\|_{\boldsymbol{A}}^* \ ,$$

where the positive definiteness ensures that $\boldsymbol{A}^{\frac{1}{2}}$, $\boldsymbol{A}^{-\frac{1}{2}}$, and $\boldsymbol{A}^{-1}$ exist. □

We apply this lemma to a single element of the stability term with $\boldsymbol{A} = \nabla^2 R(\boldsymbol{v})$ for some vector $\boldsymbol{v}$ that we will chose in a moment

$$(\boldsymbol{w}_t - \boldsymbol{w}_{t+1})^\top \hat{\boldsymbol{\ell}}_t \leq \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 R(\boldsymbol{v})} \|\hat{\boldsymbol{\ell}}_t\|_{\nabla^2 R(\boldsymbol{v})}^* \ , \tag{2.4}$$

and we know that $\nabla^2 R(\boldsymbol{v})$ is positive definite because we assume our regularization to be strongly convex. The next step is to look at $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 R(\boldsymbol{v})}^*$ in isolation. For that we define the function

$$f_t(\boldsymbol{v}) = \sum_{s=1}^{t-1} \hat{\boldsymbol{\ell}}_s^\top \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ ,$$

which is also what the argmin of $\boldsymbol{w}_t$ optimizes over in Equation (1.5). By Taylor's theorem there exists a $\boldsymbol{z}_t$ on the line segment between $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$ such that

$$f_t(\boldsymbol{w}_{t+1}) = f_t(\boldsymbol{w}_t) + (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla f_t(\boldsymbol{w}_t) + \frac{1}{2} (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla^2 f_t(\boldsymbol{z}_t) (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)$$

The term $(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla f_t(\boldsymbol{w}_t)$ is the rate of ascent of f_t at the point $\boldsymbol{w}_t$ when traveling in the direction of $\boldsymbol{w}_{t+1}$. Because $\boldsymbol{w}_t$ is optimal, that is precisely because $\boldsymbol{w}_t \in \arg\min_{\boldsymbol{v} \in \mathrm{Conv}(\mathcal{A})} f_t(\boldsymbol{v})$ by its definition in Equation (1.5), the rate of ascent towards any point in $\mathrm{Conv}(\mathcal{A})$ must be positive or 0 and thus

$$(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla f_t(\boldsymbol{w}_t) \geq 0 \ ,$$

which is also called the first-order optimality of $\boldsymbol{w}_{t+1}$. Using this fact, rearranging the Taylor expansion, and plugging in the second derivative of f_t gives that

$$\begin{aligned} f_t(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}_t) &\geq \frac{1}{2} (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla^2 f_t(\boldsymbol{z}_t) (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \\ &= \frac{1}{2\eta} \left(\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 R(\boldsymbol{z}_t)} \right)^2 \ , \end{aligned}$$

where we also used the definition of $\|\cdot\|_{\boldsymbol{A}}$. Having found a lower bound for $f_t(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}_t)$, we can use the optimality of $\boldsymbol{w}_{t+1}$ and Lemma 1 with $A = \nabla^2 R(\boldsymbol{z}_t)$ to find an upper bound as well

$$\begin{aligned} f_t(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}_t) &= \underbrace{\sum_{s=1}^{t} \hat{\boldsymbol{\ell}}_s^\top \boldsymbol{w}_{t+1} + \frac{1}{\eta} R(\boldsymbol{w}_{t+1}) - \sum_{s=1}^{t} \hat{\boldsymbol{\ell}}_s^\top \boldsymbol{w}_t + \frac{1}{\eta} R(\boldsymbol{w}_t)}_{\leq 0} + \hat{\boldsymbol{\ell}}_t (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) \\ &\leq \|\hat{\boldsymbol{\ell}}_t\|_{\nabla^2 R(\boldsymbol{z}_t)}^* \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 R(\boldsymbol{z}_t)} \ . \end{aligned}$$

Putting the lower and upper bound together, we follow that

$$\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 R(\boldsymbol{w}(\boldsymbol{z}_t))} \leq 2\eta \|\hat{\boldsymbol{\ell}}_t\|^*_{\nabla^2 R(\boldsymbol{w}(\boldsymbol{z}_t))} \ .$$

This argument, fleshed out in full detail can be found in Lemma 60. We now apply this inequality to Equation (2.4), where we pick $\boldsymbol{v} = \boldsymbol{z}_t$ and we get a bound each round of the stability term in terms of the estimator and the regularization

$$(\boldsymbol{w}_t - \boldsymbol{w}_{t+1})^\top \hat{\boldsymbol{\ell}}_t \leq 2\eta \left(\|\hat{\boldsymbol{\ell}}_t\|^*_{\nabla^2 R(\boldsymbol{z}_t)} \right)^2 \ . \tag{2.5}$$

Now it is finally time that we pick our regularizer exactly and we will chose a mix of the negative entropy and the log-barrier given by

$$\frac{1}{\eta} R(\boldsymbol{v}) = \frac{1}{\eta} \underbrace{\sum_{i=1}^K \boldsymbol{v}_i \log(\boldsymbol{v}_i)}_{\text{Negative Entropy}} + \frac{1}{\gamma} \underbrace{\sum_{i=1}^K \log(\boldsymbol{v}_i)}_{\text{Log Barrier}} \ , \tag{2.6}$$

where the $\gamma \in \mathbb{R}$ is a constant that determines the influence of the log-barrier. It is a strongly convex function and thus a valid choice. If we had just used the negative entropy then we would have recovered Exp3 but we had to mix in the negative entropy to make sense of the $\boldsymbol{z}_t$ that appears in Equation (2.5). That $\boldsymbol{z}_t$ arrived from a Taylor approximation between $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$ and the log barrier allows us to relate these quantities. Specifically, it is possible to show that

$$\boldsymbol{w}_{t+1} \in [\tfrac{1}{2}\boldsymbol{w}_t, 2\boldsymbol{w}_t] \ , \tag{2.7}$$

if $\gamma \approx \frac{1}{\sqrt{m}}$, the details of which can also be found in Section 4.4 and Chapter 4 generally. With that we also know that $\frac{1}{2}\boldsymbol{w}_t \leq \boldsymbol{z}_{t,i} \leq 2\boldsymbol{w}_t$, which allows us to continue working on Equation (2.5), by finding a lower bound on the inverse of the hessian of the regularizer by focusing on the negative entropy as $(\nabla^2 R(\boldsymbol{x}))^{-1} \succeq \operatorname{diag}(\boldsymbol{x})$

$$\begin{aligned}
2\eta \left(\|\hat{\boldsymbol{\ell}}_t\|^*_{\nabla^2 R(\boldsymbol{z}_t)} \right)^2 &= 2\eta \hat{\boldsymbol{\ell}}_t^\top \left(\nabla^2 R(\boldsymbol{z}_t) \right)^{-1} \hat{\boldsymbol{\ell}}_t \\
&\leq 2\eta \sum_{i=1}^K \hat{\boldsymbol{\ell}}_{t,i}^2 \boldsymbol{z}_{t,i} \\
&\leq 4\eta \sum_{i=1}^K \hat{\boldsymbol{\ell}}_{t,i}^2 \boldsymbol{w}_{t,i} \\
&= 4\eta \sum_{i=1}^K \frac{\boldsymbol{a}_{t,i} \boldsymbol{\ell}_{t,i}^2}{\boldsymbol{w}_{t,i}} \ ,
\end{aligned}$$

and where we also plugged in the definition of $\hat{\boldsymbol{\ell}}_{t,i}$ (Equation (2.2)). From here we are almost done, we bring in the expectation conditioned on the past, that we'll

get from another tower rule and we use $\boldsymbol{\ell}_{t,i} \leq 1$, which holds by assumption to find

$$\mathbb{E}\left[\left(\|\hat{\boldsymbol{\ell}}_t\|^*_{\nabla^2 R(z_t)}\right)^2 \,\middle|\, \mathcal{F}_t\right] = 4\eta \sum_{i=1}^{K} \frac{\mathbb{E}[\boldsymbol{a}_{t,i} \mid \mathcal{F}_t]\boldsymbol{\ell}_{t,i}^2}{\boldsymbol{w}_{t,i}} \leq 4\eta K \ . \tag{2.8}$$

We consolidate equations (2.5) and (2.8) to bound the per round stability term

$$(\boldsymbol{w}_t - \boldsymbol{w}_{t+1})^\top \hat{\boldsymbol{\ell}}_t \leq 4\eta K \ . \tag{2.9}$$

At last we return to the cheating regret term of Equation (2.3) and we pick our $\boldsymbol{u}$ close to $\boldsymbol{a}^*$ (specifically, will need to have that $\|\boldsymbol{u} - \boldsymbol{a}^*\|_\infty \leq \frac{1}{T}$) but in a way such that is not on the edge of $\Delta_{\mathcal{A}}$ where the regularizer is unbounded and then we bound

$$\frac{1}{\eta}(R(\boldsymbol{u}) - R(\boldsymbol{w}_1)) \lesssim \frac{1}{\eta} \underbrace{m \log(K)}_{\text{Negative Entropy}} + \sqrt{m} \underbrace{K \log(T)}_{\text{Log Barrier}} , \tag{2.10}$$

where the log-barrier component scales with $\log(T)$ and will thus not have an impact on the major regret terms. The details of how to chose $\boldsymbol{u}$ can can be found in Section 4.4, how to bound the regularizer is Lemma 80. All is left to assemble the pieces, starting from the regret

$$\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t \mid \mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t \mid \mathcal{F}_t\right]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[(\boldsymbol{u} - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t \mid \mathcal{F}_t\right]\right] \\
&\leq \sum_{t=1}^{T} 4\eta K + \frac{1}{\eta}(R(\boldsymbol{u}) - R(\boldsymbol{w}_1)) + K \\
&\leq 4\eta T K + \frac{1}{\eta} m \log(K) + \sqrt{m} K \log(T) + K \ ,
\end{aligned}$$

where we used $\|\boldsymbol{u} - \boldsymbol{a}^*\|_\infty \leq \frac{1}{T}$ and equations (2.3), (2.9) in the first inequality and Equation 2.10 in the second inequality.

Finally tuning $\eta = \sqrt{\frac{m \log(K)}{4\eta T K}}$ gives that

$$\mathcal{R}_T \leq 2\sqrt{mKT \log(K)} + \sqrt{m} K \log(T) + K \ .$$

Concluding the proof. This bound is optimal for combinatorial bandits, up to the $\sqrt{m} K \log(T)$ factor that arose from mixing in the log-barrier. Indeed, this additional term is not necessary and we can obtain a slightly better regret bound without the log barrier, that is by only using the negative entropy term in Equation (2.6) and then employing a more specialized analysis. We chose to present the proof with the log-barrier here as this proof technique is very flexible and will be relevant for the following Chapters.

Chapter 3

Nonstochastic Combinatorial Contextual Bandits

This chapter is based on

Zierahn, L., van der Hoeven, D., Cesa-Bianchi, N. & Neu, G.. (2023). Nonstochastic Contextual Combinatorial Bandits. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research* 206:8771-8813 Available from https://proceedings.mlr.press/v206/zierahn23a.html.

The author of this dissertation co-derived the theoretical results, co-wrote the paper, and performed the experiments.

3.1 Introduction

One fundamental extensions to the standard multi-armed bandit setup, that we have not discussed so far are *contextual bandits*, which allow taking contextual information into account during decision making. In contextual bandits, a vector of side information $\boldsymbol{x}_t \in \mathbb{R}^K$ is available to the learner. The context can be drawn from some distribution or chosen adversarially, the loss is usually linear in the context. Finally, the definition of the regret changes, the comparator now becomes the best fixed context to action mapping in hindsight, that is for some set of contexts $\mathcal{X}$ and some actionset $\mathcal{A}$, the regret is defined as

$$\mathcal{R} = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}(\boldsymbol{a}_t)\right] - \min_{\pi:\mathcal{X}\to\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}(\pi(\boldsymbol{x}_t))\right] .$$

Both aspects from *contextual bandits* and *combinatorial bandits* are important to handle in the key application areas of bandit algorithms, including online advertising and recommendation systems (Li, Chu, Langford, and Schapire, 2010), and

sequential treatment allocation (Tewari and Murphy, 2017). For instance, recommendation systems often need to produce structured lists of recommendations (providing a combinatorial aspect), while also taking into account the unique preferences of the user (providing a contextual aspect). The two aspects have been successfully addressed in the framework of *contextual combinatorial bandits* in a sequence of works initiated by Qin, Chen, and Zhu (2014). All of these previous works have focused on stochastic losses. As in previous chapters, we study the *nonstochastic* version of the contextual combinatorial bandit problem, where the sequence of losses incurred by the learning agent does not necessarily come from a fixed distribution, but can be possibly influenced by an external (even malicious) force. Since the real world is rarely stationary, this extension is of key practical importance as it significantly broadens the scope of the existing theory.

The setting we consider unifies many previous problem settings, and presents a new level of challenges that have not been encountered in previous work. In particular, handling the nonstochastic setting requires a drastically different set of tools than needed in the i.i.d. case considered in all past work on contextual combinatorial bandits: while all known approaches in this latter scenario are based on the principle of optimism in the face of uncertainty (Auer, 2002; Auer, Cesa-Bianchi, and Fischer, 2002), this idea is known to fail when the losses can be generated by an adversarial external process—see, e.g., Cesa-Bianchi and Lugosi, 2006, Section 4.1. A natural alternative route that we follow in this paper is to adapt the classic Exp3 algorithm of Auer, Cesa-Bianchi, Freund, and Schapire (2002) to deal with the potential nonstationarity of the losses via the use of an importance-weighted loss estimator. This method has been adapted to deal with combinatorial action spaces by Cesa-Bianchi and Lugosi (2012a) and Audibert, Bubeck, and Lugosi (2014), and to deal with contextual information by Neu and Olkhovskaya (2020). Both of these extensions are based on generalizing the standard scalar-valued importance-weighted estimator of Auer, Cesa-Bianchi, Freund, and Schapire (2002) to be able to directly estimate an unknown vector-valued problem parameter. A direct combination of these techniques to tackle our problem is far from straightforward due to the fact that our scenario requires the estimation of a *matrix-valued* parameter. Our main contribution is addressing this challenge via designing a range of new estimation procedures suitable for estimating such parameter matrices based on limited observations, and using them in conjunction with online decision making algorithms.

Following the terminology of Audibert, Bubeck, and Lugosi (2014), we consider two different feedback models: *semi-bandit* feedback, where the learner gets to observe the feedback associated with each component of its combinatorial action, and *full-bandit* feedback, where the learner only observes the total loss associated with its decision (for formal definitions, see Section 3.2). For both of these scenarios, we design new loss estimators based on the geometric resampling method proposed by Neu and Olkhovskaya (2020), which itself is a generalization of the geometric

resampling method of Neu and Bartók (2013) and Neu and Bartók (2016). The most challenging full-bandit scenario requires rather sophisticated treatment: here, our estimators are based on calculating and manipulating certain linear operators over matrices, which we represent by tensors of appropriate size. The estimation procedure used in this setting as well as the control over the resulting estimator is probably the most advanced technical tool we develop in this paper.

The concrete results we achieve in this chapter are the following. We suppose that the context vectors are d-dimensional, that the actions can be represented by K-dimensional binary vectors with at most m components being equal to 1, and that the losses suffered by the learner are linear in both the contexts and the actions, parametrized by a $K \times d$ matrix specifying the loss function. In this setting, we prove regret bounds of order $\sqrt{mKT \max\{d, m/\lambda_{\min}\}}$ in the semi-bandit setting and $m^{3/2}\sqrt{KT \max\{d, m/\lambda_{\min}\}}$ in the full-bandit setting (neglecting minor logarithmic factors). Here, $\lambda_{\min}$ is a lower bound on the smallest eigenvalue of the covariance matrix of the contexts. These bounds are achieved by combining the estimators mentioned in the previous paragraph with appropriately chosen extensions of the classic Exp3 and Follow the Regularzied Leader (FTRL) algorithms adapted to the combinatorial setting (Cesa-Bianchi and Lugosi, 2012a; Audibert, Bubeck, and Lugosi, 2014). The best known results are recovered for both adversarial contextual bandits[1] when $m = 1$ (Neu and Olkhovskaya, 2020) and for combinatorial bandits and semi-bandits when $d = 1$ (Audibert, Bubeck, and Lugosi, 2014). Our algorithms can be implemented with polynomial runtime whenever the decision space allows an efficient implementation of the FTRL/Exp3 variants our methods are based on—more details are given in Sections 3.3 and 3.4 presenting the two methods.

Similar results have been achieved previously for the simpler i.i.d. setting. Qin, Chen, and Zhu (2014) consider a scenario where the loss function is determined by a single d-dimensional parameter vector and the context can be represented by a $K \times d$ matrix. They propose an algorithm based on the principle of optimism in the face of uncertainty and achieve a regret guarantee of order $d\sqrt{mT \log(KT)}$ for this setting[2]. Similar results have been achieved by Li, Wang, Zhang, and Chen (2016) (and a sequence of follow-up works) who consider a slightly different observation model generalizing semi-bandit feedback. On the side of non-stochastic losses,

[1] The bounds stated by Neu and Olkhovskaya (2020) do not explicitly feature the $1/\lambda_{\min}$ factor, although a careful inspection of their proofs reveal that their bounds should indeed increase with this quantity.

[2] Their dependence on K is much milder due to the number of parameters to estimate being only d as opposed to Kd in our setting. The dependence on $\sqrt{m}$ we claim here follows from instantiating their bound with $C = m$ which is required when considering linear losses. Their bounds actually hold with slightly greater generality, allowing generalized linear loss functions.

the only relevant works we are aware of are those of Kale, Reyzin, and Schapire (2010b) and Krishnamurthy, Agarwal, and Dudik (2016), who both consider the semi-bandit setting with loss functions that are potentially non-linear with respect to the contexts, but are restricted to work with a finite policy class that maps contexts to combinatorial actions. A naïve instantiation of their bounds roughly[3] results in a regret bound of order $K\sqrt{dmT\log(T)}$. Implementing these latter algorithms requires either a full enumeration of the exponentially-sized policy space or access to a non-standard optimization oracle. In comparison, the computational steps required by our algorithms are relatively standard, and our methods can be implemented efficiently in a range of practically interesting problem settings.

The rest of this chapter is organized as follows. In Section 3.2 we formally introduce the setting and corresponding assumptions. In Section 3.3 we present the algorithm and analysis of the algorithm for the semi-bandit setting and in Section 3.4 we do the same for the full-bandit setting. In Section 3.5 we provide lower bounds and finally, in Section 3.6 we empirically evaluate our algorithms.

3.2 Preliminaries

We are considering a nonstochastic bandit problem with combinatorial actions and contexts provided in each timestep. Given an action set $\mathcal{A} \subseteq \{0,1\}^K$, a context space $\mathcal{X} \subseteq \mathbb{R}^d$, and a distribution $\mathcal{D}$ over $\mathcal{X}$, our learning protocol can be described as follows. In each round t:

1. The non-oblivious adversary picks a loss matrix $\boldsymbol{\Theta}_t \in \mathbb{R}^{d\times K}$
2. The environment draws an independent context vector $\boldsymbol{x}_t \sim \mathcal{D}$
3. The learner observes $\boldsymbol{x}_t$ and picks an action from the action set $\boldsymbol{a}_t \in \mathcal{A} \subseteq \{0,1\}^K$
4. The learner incurs the loss $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t$
5. The learner observes:
   full-bandit: $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t$
   semi-bandit: $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \odot \boldsymbol{a}_t$

where $\odot$ is the elementwise multiplication of two vectors. We assume without loss of generality, that $\mathcal{A}$ contains K linearly independent vectors.

Additional notation. Each element $\boldsymbol{a}$ of the action set $\mathcal{A}$ is called an action. Each one of the K dimensions that make up the action set is called a sub-action, and a single action $\boldsymbol{a}$ has at most m active sub-actions, $\|\boldsymbol{a}\|_1 \leq m$. We use $\lambda_{\min}(\cdot)$

[3]This follows from discretizing the space of loss matrices at a resolution of order $1/T$, and considering the class of greedy policies with respect to this cover. Details of how such an argument can be fully worked out are non-trivial, and a fully rigorous argument may likely lead to a worse regret bound.

to denote the smallest eigenvalue of a matrix or tensor, $\|\cdot\|_{\text{op}}$ to denote the operator norm of a matrix, which is given by the largest eigenvalue $\lambda_{\max}(\cdot)$ for square positive definite matrices, and $\boldsymbol{e}_i$ to denote the basis vector in direction i. The filtration over all past events at a timestep t is given by $\mathcal{F}_t$.

Core assumptions. Our results rely on the following assumptions that are rather standard in the combinatorial and contextual bandit literature.

- The distribution $\mathcal{D}$ from which the contexts are independently drawn is known and satisfies $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \Sigma \succ 0I$, with $\lambda_{\min}^{\boldsymbol{\Sigma}} \coloneqq \lambda_{\min}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top])$ as the smallest eigenvalue of the covariance matrix;
- there exists a $\sigma > 0$ such that $\|\boldsymbol{x}\|_2 \leq \sigma$ holds $\mathcal{D}$-almost surely;
- $\max_t \|\boldsymbol{\Theta}_t\|_F \leq G$ for some $G > 0$, where $\|\cdot\|_F$ is the Frobenius norm;
- $\max_t \|(\boldsymbol{\Theta}_t)_{\cdot,i}\|_2 \leq R$ for all $i \in d$, where $(\cdot)_{\cdot,i}$ is the $i-$th column of a matrix;
- $\max_t \|\boldsymbol{x}^\top \boldsymbol{\Theta}_t\|_\infty \leq 1$ holds $\mathcal{D}$-almost surely.

The regret in our setting is defined by the best context-to-action mapping $\pi : \mathcal{X} \to \mathcal{A}$ in hindsight

$$\mathcal{R}_T = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \left(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \pi(\boldsymbol{x}_t)\right)\right] ,$$

where Π is the set of all context to action mappings.

3.3 Semi-Bandits

In this section, we focus on the semi-bandit feedback, though some of the tools and methods will also be applicable in the full-bandit setting. In the semi-bandit setting, we are able to observe $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \odot \boldsymbol{a}_t$ after playing an action, where $\odot$ is the Hadamard or elementwise product. The algorithm we present in this section, Algorithm 3, is based on FTRL. Next we introduce our loss estimators, first we will introduce an unbiased estimator $\hat{\boldsymbol{\Theta}}_t \in \mathbb{R}^{d \times K}$, defined as

$$(\hat{\boldsymbol{\Theta}}_t)_{\cdot,k} = \boldsymbol{\Sigma}_{t,k}^{-1} \boldsymbol{x}_t (\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k (\boldsymbol{a}_t)_k , \tag{3.3}$$

where $\boldsymbol{\Sigma}_{t,k} = \mathbb{E}_{\boldsymbol{a}_t, \boldsymbol{x}}\left[(\boldsymbol{a}_t)_k \boldsymbol{x}\boldsymbol{x}^\top \mid \mathcal{F}_t\right]$. We show that $\hat{\boldsymbol{\Theta}}_t$ is an unbiased estimator of $\boldsymbol{\Theta}_t$ as

$$\mathop{\mathbb{E}}_{\boldsymbol{a}_t, \boldsymbol{x}_t}\left[(\hat{\boldsymbol{\Theta}}_t)_{\cdot,k} \mid \mathcal{F}\right] = \boldsymbol{\Sigma}_{t,k}^{-1} \mathop{\mathbb{E}}_{\boldsymbol{a}_t, \boldsymbol{x}_t}\left[\boldsymbol{x}_t \boldsymbol{x}_t^\top (\boldsymbol{a}_t)_k \mid \mathcal{F}\right] (\boldsymbol{\Theta}_t)_{\cdot,k} = (\boldsymbol{\Theta}_t)_{\cdot,k}$$

where the last equality follows from the definition of $\boldsymbol{\Sigma}_{t,k}$. However, this estimator has to find $\boldsymbol{\Sigma}_{t,k} = \mathbb{E}_{\boldsymbol{a}_t, \boldsymbol{x}}\left[(\boldsymbol{a}_t)_k \boldsymbol{x}\boldsymbol{x}^\top \mid \mathcal{F}_t\right]$ explicitly. Depending on the context distribution, closed form computation of that expectation might be impossible or at least

Require: learning rate $\eta > 0$, exploration rate $\gamma \in (0, 1)$
Require: exploration set $\mathcal{E} \subseteq \mathcal{A}$
for t in $[T]$ **do**
  Observe $\boldsymbol{x}_t$
  Compute

$$\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t)) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathrm{Conv}(\mathcal{A})} \sum_{\tau=1}^{t-1} \boldsymbol{x}_t^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{v} + R(\boldsymbol{v}) \ , \tag{3.1}$$

  Find probability distribution $p_t(\cdot|\boldsymbol{x}_t)$ such that $\mathbb{E}_{\boldsymbol{a} \sim p_t(\cdot|\boldsymbol{x}_t)}[\boldsymbol{a}] = \boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t))$
  Set

$$\pi_t(\boldsymbol{a}|X) = (1-\gamma) p_t(\boldsymbol{a}|X) + \gamma \frac{\mathbb{1}[\boldsymbol{a} \in \mathcal{E}]}{|\mathcal{E}|} \tag{3.2}$$

  Draw and play $\boldsymbol{a}_t \sim \pi_t(\cdot|\boldsymbol{x}_t)$
  Observe loss $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \odot \boldsymbol{a}_t$ and compute $\tilde{\boldsymbol{\Theta}}_t$ using (3.4).
end for

Algorithm 3: CO_2-FTRL

computationally infeasible. Thus, we will instead approximate that expectation by drawing samples using a process called Matrix Geometric Resampling (MGR) (Neu and Olkhovskaya, 2020), leading to a slightly biased estimator

$$(\tilde{\boldsymbol{\Theta}}_t)_{\cdot,k} = \hat{\boldsymbol{\Sigma}}_{t,k}^{+} \boldsymbol{x}_t (\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k (\boldsymbol{a}_t)_k \ , \tag{3.4}$$

where $\hat{\boldsymbol{\Sigma}}_{t,k}^{+}$ is a biased estimate of $\boldsymbol{\Sigma}_{t,k}$. The full details of this estimator, the approximation, and the MGR are explored in Chapter 3.3.2. Given the loss estimator of past rounds, Algorithm 3 first computes

$$\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x})) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathrm{Conv}(\mathcal{A})} \sum_{\tau=1}^{t-1} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{v} + R(\boldsymbol{v}) \ , \tag{3.5}$$

where

$$\tilde{\boldsymbol{L}}_t(\boldsymbol{x}) = \sum_{\tau=1}^{t-1} \tilde{\boldsymbol{\Theta}}_\tau^\top \boldsymbol{x} \qquad \text{and} \qquad R(\boldsymbol{v}) = \frac{1}{\eta} \sum_{k=1}^{K} \Big((\boldsymbol{v})_k \log(\boldsymbol{v})_k - (\boldsymbol{v})_k \Big) \ . \tag{3.6}$$

The vector $\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t) \in \mathbb{R}^K$ estimates the cumulative losses of all sub-actions given the current context and past observations. Indeed, this is the only place where we actually use the context in the entire algorithm. Thus, Algorithm 3 can be interpreted as the non-contextual combinatorial algorithms of Koolen, Warmuth, and Kivinen (2010) and Audibert, Bubeck, and Lugosi (2014) run with a more sophisticated, context dependent, estimator.

As is the case with many FTRL algorithms, we aim to play actions that are close to the output of the optimization problem $\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t))$ in expectation. For that we

need to find a distribution over actions $\boldsymbol{p}_t \in \Delta_{\mathcal{A}}$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t))$. A more complete discussion on when that is efficiently possible can be found in Section 4.4 and the references therein.

Finally, we need to ensure that we play each sub-action sufficiently often. For that we construct an exploration set $\mathcal{E} \subseteq \mathcal{A}$ such that there is at least one $\boldsymbol{a} \in \mathcal{E}$ satisfying $(\boldsymbol{a})_k = 1$ for each $k \in [K]$. Then mixing in the exploration set to the distribution over actions found as derived from $\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}_t))$, sampling an action and computing a new loss estimator are the last steps of the algorithm. The full algorithm is spelled out in Algorithm 3.

The main result of this section is the regret bound of CO_2-FTRL in Theorem 2.

Theorem 2. *With* $M = \frac{\log(\gamma m)}{\beta\lambda^{\boldsymbol{\Sigma}}_{\min}}$, $\gamma = \min\left\{\frac{1}{2}, \sqrt{\frac{(1+\log(K/m))|\mathcal{E}|\log(T)}{T\beta\lambda^{\boldsymbol{\Sigma}}_{\min}}}\right\}$, *and* $\eta = \min\left\{\frac{\log(2)}{(M+1)}, \sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{2KdT}}\right\}$, *$CO_2$-FTRL (Algorithm 3) satisfies*

$$\mathcal{R}_T \in O\left(\sqrt{mKT\left(1+\log\frac{K}{m}\right)\max\left\{d, \frac{m\sigma^2\log(T)}{\lambda^{\boldsymbol{\Sigma}}_{\min}}\right\}}\right) .$$

This theorem is implied by Theorem 17 in Appendix 3.B. The rest of this section is dedicated to proving and discussing this result. First, we interrogate the MGR and how biased the loss estimator is in Section 3.3.2, then we introduce a regret decomposition inspired by Neu and Olkhovskaya (2020), that uses a so called *Ghost Sample* to temporary decouple the regret from the randomness of the contexts in Section 3.3.1. Both the MGR and the regret decomposition will also be useful in the full-bandit setting.

3.3.1 Regret Decomposition and Ghost Sample

The analysis pertaining to the contextual bandit part of the problem is inspired by Neu and Olkhovskaya (2020). We can exploit the fact that the context distribution is stationary and that all contexts are drawn independently. That is $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \sim \mathcal{D}$. The *Ghost Sample* is then simply yet another sample $\boldsymbol{x}_0 \sim \mathcal{D}$ that follows the same distribution and is independent of all other $\boldsymbol{x}_t$. This additional sample is only for the purposes of analysis and allows us to decompose the regret for each context. Using $\pi^*_T(\boldsymbol{x}) = \min_{\boldsymbol{a}\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^T \left(\boldsymbol{x}^\top \boldsymbol{\Theta}_t \boldsymbol{a}\right)\right]$ as the best context to action mapping in hindsight, we can use the tower rule, the fact that all $\boldsymbol{x}_t$ follow

the same distribution and the tower rule again to conclude that

$$\begin{aligned}
\mathcal{R}_T &= \max_{\pi\in\Pi} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_T\\ \boldsymbol{a}_1,\dots,\boldsymbol{a}_T}} \left[\sum_{t=1}^{T} \left(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \pi(\boldsymbol{x}_t)\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{F}} \left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_t\sim\mathcal{D}\\ \boldsymbol{a}_t\sim\boldsymbol{p}_t(\boldsymbol{x}_t)}} \left[\left(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \pi^*(\boldsymbol{x}_t)\right) \,\middle|\, \mathcal{F}_t\right]\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{F}} \left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_0\sim\mathcal{D}\\ \boldsymbol{a}_t\sim\boldsymbol{p}_t(\boldsymbol{x}_0)}} \left[\left(\boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \pi^*(\boldsymbol{x}_0)\right) \,\middle|\, \mathcal{F}_t\right]\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_0\sim\mathcal{D}} \left[\underbrace{\mathop{\mathbb{E}}_{\mathcal{F}} \left[\sum_{t=1}^{T} \left(\boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \pi^*(\boldsymbol{x}_0)\right) \,\middle|\, \boldsymbol{x}_0\right]}_{\mathcal{R}_T(\boldsymbol{x}_0)}\right] . \qquad (3.7)
\end{aligned}$$

If we fix any context $\boldsymbol{x} \in \mathcal{X}$, then $\mathcal{R}_T(\boldsymbol{x})$ is the regret we expect to incur if we would only observe $\boldsymbol{x}$ as our context and the main idea of the proofs is bounding $\mathcal{R}_T(\boldsymbol{x}_0)$ taking $\boldsymbol{x}_0$ as fixed and non-random and then taking an expectation over $\boldsymbol{x}_0$ at the very end. The *Ghost Sample* will be used in both the full-bandit and semi-bandit settings and we formalize it in a single lemma.

Lemma 3 (Neu and Olkhovskaya (2020, Equation (6)))**.** *Let $\tilde{\boldsymbol{\Theta}}_t$ be some estimator of $\boldsymbol{\Theta}_t$ that only depends on $\mathcal{F}_t$ and with bias $\boldsymbol{B}_t$ such that $\boldsymbol{\Theta}_t = \mathbb{E}[\tilde{\boldsymbol{\Theta}}_t] - \boldsymbol{B}_t$, then for any $\boldsymbol{x}_0 \sim \mathcal{D}$*

$$\mathcal{R}_T \le \mathop{\mathbb{E}}_{\boldsymbol{x}_0\sim\mathcal{D}} \left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] + 2 \mathop{\mathbb{E}}_{\boldsymbol{x}_0\sim\mathcal{D},\mathcal{F}} \left[\sum_{t=1}^{T} \max_{\boldsymbol{a}\in\mathcal{A}} \left\|\mathbb{E}\left[\boldsymbol{x}_0^\top \boldsymbol{B}_t \boldsymbol{a} \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right]\right\|_1\right] ,$$

where $\|\cdot\|_1$ is the L_1 norm equal to the absolute value in $\mathbb{R}$ and

$$\tilde{\mathcal{R}}_T(\boldsymbol{x}_0) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \left(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}_t - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \pi^*(\boldsymbol{x}_0)\right) \,\middle|\, \boldsymbol{x}_0\right] .$$

Proof. We start with Equation (3.7)

$$\mathcal{R}_T = \mathop{\mathbb{E}}_{\boldsymbol{x}_0\sim\mathcal{D}} \left[\mathcal{R}_T(\boldsymbol{x}_0)\right] .$$

We focus a single $\mathcal{R}_T(\boldsymbol{x})$. Then using the tower rule, $\boldsymbol{\Theta}_t = \tilde{\boldsymbol{\Theta}}_t - \boldsymbol{B}_t$ and then upper bounding $\boldsymbol{x}^\top \boldsymbol{B}_t \boldsymbol{a} \le |\boldsymbol{x}^\top \boldsymbol{B}_t \boldsymbol{a}| \le \max_{\boldsymbol{a}'\in\mathcal{A}} |\boldsymbol{x}^\top \boldsymbol{B}_t \boldsymbol{a}'|$, which holds for all $\boldsymbol{a} \in \mathcal{A}$, we

are able to replace $\boldsymbol{\Theta}$ by $\tilde{\boldsymbol{\Theta}}$ in exchange for an additive bias term for any $\boldsymbol{x} \in \mathcal{X}$

$$\begin{aligned}
\mathcal{R}_T(\boldsymbol{x}) &= \mathop{\mathbb{E}}_{\mathcal{F}}\left[\sum_{t=1}^{T}\left(\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{a}_t - \boldsymbol{x}^\top \boldsymbol{\Theta} \pi^*(\boldsymbol{x})\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{F}}\left[\sum_{t=1}^{T}\mathbb{E}\left[\left(\boldsymbol{x}^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t - \boldsymbol{x}^\top \boldsymbol{\Theta}_t \pi^*(\boldsymbol{x})\right) \,\middle|\, \mathcal{F}_t\right]\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{F}}\left[\sum_{t=1}^{T}\mathbb{E}\left[\left(\boldsymbol{x}^\top \left(\tilde{\boldsymbol{\Theta}}_t - \boldsymbol{B}_t\right) \boldsymbol{a}_t - \boldsymbol{x}^\top \left(\tilde{\boldsymbol{\Theta}}_t - \boldsymbol{B}_t\right) \pi^*(\boldsymbol{x})\right) \,\middle|\, \mathcal{F}_t\right]\right] \\
&\leq \tilde{\mathcal{R}}_T(\boldsymbol{x}) + 2 \mathop{\mathbb{E}}_{\mathcal{F}}\left[\sum_{t=1}^{T} \max_{\boldsymbol{a} \in \mathcal{A}} \left\| \mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{B}_t \boldsymbol{a} \,|\, \mathcal{F}_t\right]\right\|_1\right] .
\end{aligned}$$

We conclude the proof by putting both equations together

$$\mathcal{R}_T = \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}}\left[\mathcal{R}_T(\boldsymbol{x}_0)\right] \leq \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}}\left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] + 2 \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}, \mathcal{F}}\left[\sum_{t=1}^{T} \max_{\boldsymbol{a} \in \mathcal{A}} \left\| \mathbb{E}\left[\boldsymbol{x}_0^\top \boldsymbol{B}_t \boldsymbol{a} \,|\, \boldsymbol{x}_0, \mathcal{F}_t\right]\right\|_1\right] .$$

□

3.3.2 Matrix Geometric Resampling

All of the estimators introduced in the previous section require to compute an expectation over the context space $\mathcal{D}$ which can make a closed form computation impossible. A common idea to tackle this problem is to sample the expectation instead of computing it outright and while we can easily obtain an unbiased estimator of the expectation, we require the inverse of the expectation for which an unbiased estimator may not be available. Neu and Olkhovskaya (2020) address this issue by introducing Matrix Geometric Resampling (MGR) to obtain a biased estimate of the inverse of the expectation through sampling from the context distribution $\mathcal{D}$. A generalized version of the MGR from Neu and Olkhovskaya (2020) can be found in Algorithm 4. The principal input to the MGR is a sampling scheme, which is any way to produce independent and unbiased samples $\hat{\boldsymbol{P}}_k$ of an invertible matrix $\boldsymbol{P}$, such that $\mathbb{E}[\hat{\boldsymbol{P}}_k] = \boldsymbol{P}$. One example of a sampling scheme is Sampling Scheme 5, which is used by CO_2-FTRL. The MGR then computes a biased estimate $\boldsymbol{P}^{-1}$ only based on the samples $\hat{\boldsymbol{P}}_k$. The reader may be familiar with a geometric series from which the MGR derives it's name, that is

$$\sum_{i=0}^{M-1} r^i = \frac{1 - r^M}{1 - r},$$

for $M \in \mathbb{N}$ and $r \in (-1, 1)$. A generalization of the geometric series to linear operators is the von Neumann series which we slightly reformulate to obtain

$$\sum_{i=0}^{M-1} (\boldsymbol{I} - \boldsymbol{P})^i = \boldsymbol{P}^{-1} - (\boldsymbol{I} - \boldsymbol{P})^M \boldsymbol{P}^{-1}, \tag{3.8}$$

Require: Sampling Scheme S, $\beta > 0, M > 0$
1: **for** $k = 1, \ldots, M-1$: **do**
2: Draw $\hat{\boldsymbol{P}}_k$ according to S
3: Compute $\boldsymbol{C}_k = \prod_{j=1}^{k}(I - \beta\hat{\boldsymbol{P}}_j)$
4: **end for**
5: **Output** $\hat{\boldsymbol{P}}^+ = \beta\boldsymbol{I} + \beta\sum_{k=1}^{M-1}\boldsymbol{C}_k$

Algorithm 4: MGR

where $\boldsymbol{I}$ is the identity matrix, $M \in \mathbb{N}$ and $\boldsymbol{P}$ is some invertible matrix. This equation allows us to express the inverse of $\boldsymbol{P}$ as a sum, which can be sampled using only $\hat{\boldsymbol{P}}_k$ and an additional bias term that will vanish for $M \to \infty$ if $\|\boldsymbol{I} - \boldsymbol{P}\|_{\text{op}} \in (-1, 1)$. The condition on the bias is fulfilled if $\boldsymbol{P}$ is semi-definite and $\lambda_{\max}(\boldsymbol{P})$ is less than 2 (in fact we will even require that $\lambda_{\max}(\boldsymbol{P}) \leq 1$ for the MGR proof) and the multiplicative parameter β ensures that this condition on the largest eigenvalue is satisfied.

The core results of the MGR are captured in Lemma 4 below, the proof of which can be found in Appendix 3.A.

Lemma 4. *Let $\hat{\boldsymbol{P}}^+$ be defined by the MGR procedure (Algorithm 4) run for M iterations where each $\hat{\boldsymbol{P}}_k \in \mathbb{R}^{b\times b}$ drawn in Step 2 of Algorithm 4 is positive semi-definite, and such that $\mathbb{E}[\hat{\boldsymbol{P}}_k] = \boldsymbol{P}$, where $\boldsymbol{P}$ is also symmetric and positive semi-definite. Choose β such that $\beta \leq \frac{1}{\lambda_{\max}(\hat{\boldsymbol{P}}_k)}$ with probability 1 for all k, then the following three results hold*

1. $\operatorname{tr}\Big(\mathbb{E}_{\text{MGR}}[\boldsymbol{P}\hat{\boldsymbol{P}}^{+\top}\boldsymbol{P}\hat{\boldsymbol{P}}^+]\Big) < 2b$

2. $\mathbb{E}_{\text{MGR}}[\hat{\boldsymbol{P}}^+]\boldsymbol{P} = \boldsymbol{I} - (\boldsymbol{I} - \beta\boldsymbol{P})^M$ *and* $\boldsymbol{P}\,\mathbb{E}_{\text{MGR}}[\hat{\boldsymbol{P}}^+] = \boldsymbol{I} - (\boldsymbol{I} - \beta\boldsymbol{P})^M$

3. $\|\hat{\boldsymbol{P}}^+\|_{\text{op}} \leq (M+1)\beta$.

Computational efficiency of MGR. If i.i.d. samples from the context distribution $\mathcal{D}$ and from the distibution $\pi_t(\cdot|\boldsymbol{x})$ over $\mathcal{A}$ are both available through sampling oracles, then MGR can be run in time of order $MKd + Kd^2$ (Neu and Olkhovskaya, 2020), where we assume both oracles can return a random draw from their corresponding distributions in unit time. Conditions on $\mathcal{A}$ enabling an efficient implementation of the sampling oracle for $\pi_t(\cdot|\boldsymbol{x})$ are discussed in Sections 3.3 and 3.4.

Require: Context distribution $\mathcal{D}$, current policy π_t, sub-action k
1: Draw $\boldsymbol{x} \sim \mathcal{D}$
2: Draw $\boldsymbol{a} \sim \pi_t(\cdot|\boldsymbol{x})$
3: Output $(\boldsymbol{a})_k \boldsymbol{x}\boldsymbol{x}^\top$

Sampling-Scheme 5: CO_2-FTRL Sampling Scheme

3.3.3 Sampling Scheme and Loss Estimators

We now focus on the loss estimator in the semi-bandit setting. Recall from the definition in Equation (3.3) that the unbiased estimator is defined column wise

$$(\hat{\boldsymbol{\Theta}}_t)_{\cdot,k} = \boldsymbol{\Sigma}_{t,k}^{-1} \boldsymbol{x}_t (\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k (\boldsymbol{a}_t)_k \ ,$$

where $\boldsymbol{\Sigma}_{t,k} = \mathbb{E}_{\boldsymbol{a}_t, \boldsymbol{x}} \left[(\boldsymbol{a}_t)_k \boldsymbol{x}\boldsymbol{x}^\top \mid \mathcal{F}_t \right]$. $\boldsymbol{\Sigma}_{t,k}^{-1}$ is the potentially problematic element that we will replace by the MGR. For that we define S as spelled out in Sampling-Scheme 5, which draws unbiased samples from $(\boldsymbol{a}_t)_k \boldsymbol{x}\boldsymbol{x}^\top$ given the history $\mathcal{F}_t$. Plugging that into the MGR yields

$$\hat{\boldsymbol{\Sigma}}_{t,k}^{+} = \mathrm{MGR}(S(\mathcal{D}, \pi_t, k), \beta, M),$$

where β and M are hyperparameters of the algorithm. We define the corresponding loss estimator as

$$(\tilde{\boldsymbol{\Theta}}_t)_{\cdot,k} = \hat{\boldsymbol{\Sigma}}_{t,k}^{+} \boldsymbol{x}_t (\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k (\boldsymbol{a}_t)_k \ . \tag{3.9}$$

The bias of $\tilde{\boldsymbol{\Theta}}_t$ depends on β, the number of iterations we are running the MGR for M and the size of the smallest eigenvalue of $\boldsymbol{\Sigma}_{t,k}$, the exact result is given in Lemma 5, which is proven in Appendix 3.A. Controlling $\lambda_{\min}(\boldsymbol{\Sigma}_{t,k})$ is the only reason why we need to mix in an exploratory policy in step 5 of CO_2-FTRL (Algorithm 3).

Lemma 5. *Let* $\beta \leq \min_{t \in [T], k \in [K]} \frac{1}{\lambda_{\max}(\boldsymbol{\Sigma}_{t,k})}$ *and* $\tilde{\boldsymbol{\Theta}}_t$ *as defined in Equation* (3.9). *For all* $\boldsymbol{a} \in \mathcal{A}$ *and all* $\boldsymbol{x} \in \mathcal{X}$ *Algorithm 3 guarantees*

$$\mathbb{E}\left[\boldsymbol{x}^\top (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t) \boldsymbol{a} \,\middle|\, \mathcal{F} \right] \leq m\sigma R \exp\left(-\frac{M\beta\gamma}{|\mathcal{E}|} \lambda_{\min}^{\boldsymbol{\Sigma}} \right) \ ,$$

for all $t \in [T]$, *where* $\mathcal{E}$ *is the exploration set.*

3.3.4 Main Result

To prove Theorem 17, which implies Theorem 2, we need two more lemmas, the first is standard in the bandit literature. It upper bounds the regret by the size of the regularization and the variance of the estimator. The proof is essentially applying a result of Orabona (2019) for each context independently and the main body of the proof is checking the required assumptions, the full details of the proof can be found in Appendix 3.B.

Lemma 6. *Let $\beta = \frac{1}{\sigma^2}$ and let $\eta \leq \frac{\log(2)}{M+1}$. Let R as in Equation (3.6), then we have that for any $\boldsymbol{x} \in \mathcal{X}$, with $\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}))$ defined by (3.5) and any $\boldsymbol{u} \in \mathcal{A}$, it holds that*

$$\sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t (\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x})) - \boldsymbol{u}) \leq \frac{m\left(1 + \log\left(\frac{K}{m}\right)\right)}{\eta} + \eta \sum_{t=1}^{T} \sum_{k=1}^{K} (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t)_k^2 (\boldsymbol{w}_t(\boldsymbol{x}))_k \ .$$

Next, we need to control the $\sum_{k=1}^{K} (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t)_k^2 (\boldsymbol{w}_t(\boldsymbol{x}))_k$ terms in Lemma 6. First we show that $\mathbb{E}\left[(\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t)_k^2 (\boldsymbol{w}_t(\boldsymbol{x}))_k \,\middle|\, \mathcal{F}\right]$ is upper bounded by $2\,\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_{t,k} \hat{\boldsymbol{\Sigma}}_{t,k}^{+,\top} \boldsymbol{\Sigma}_{t,k} \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\right) \,\middle|\, \mathcal{F}\right]$, after which we can use Lemma 4 to control the bias. The result can be found in Lemma 7 and the complete proof in Appendix 3.B.

Lemma 7. *For any $\gamma \in (0, 1)$ and for all $t \in [T]$ we have that*

$$\mathbb{E}\left[\sum_{k=1}^{K} \left(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t\right)_k^2 \left(\boldsymbol{w}_t(\boldsymbol{x}_0)\right)_k \,\middle|\, \mathcal{F}_t\right] \leq \frac{2Kd}{1-\gamma} \ .$$

By combining Lemmas 5, 6, 7 we arrive at the final regret bound in Theorem 2, whose proof is implied by Theorem 17 in Appendix 3.B.

Theorem 2 (RESTATED). *With $M = \frac{\log(\gamma m)}{\beta \lambda_{\min}^{\boldsymbol{\Sigma}}}$, $\gamma = \min\left\{\frac{1}{2}, \sqrt{\frac{(1+\log(K/m))|\mathcal{E}|\log(T)}{T\beta\lambda_{\min}^{\boldsymbol{\Sigma}}}}\right\}$, and $\eta = \min\left\{\frac{\log(2)}{(M+1)}, \sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{2KdT}}\right\}$, CO_2-FTRL (Algorithm 3) satisfies*

$$\mathcal{R}_T \in O\left(\sqrt{mKT\left(1 + \log\frac{K}{m}\right)\max\left\{d, \frac{m\sigma^2\log(T)}{\lambda_{\min}^{\boldsymbol{\Sigma}}}\right\}}\right) \ .$$

As we will also discuss in Section 3.5, this result would be tight, were it not for the factor $\frac{m\sigma^2\log(T)}{\lambda_{\min}^{\boldsymbol{\Sigma}}}$, that appears in the max.

3.4 Full-Bandits

In this section we describe our results in the full-bandit setting, where we only have access to the total loss $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t$ incurred at each timestep t rather than the loss components $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \odot \boldsymbol{a}_t$ that we had access to in the semi-bandit setting. We start by describing how to construct an unbiased estimator for $\boldsymbol{\Theta}_t$. In order to do so we need to introduce several definitions related to tensors.

Tensor definitions. We will thoroughly introduce tensors in Appendix 3.C but we will give a quick start here to understand the ideas behind our algorithm. Let

$\boldsymbol{\Phi} \in \mathbb{R}^{d\times d\times K\times K}$ be a tensor and let $\boldsymbol{A} \in \mathbb{R}^{d\times K}$ be a matrix. Tensor $\boldsymbol{\Phi}$ acting on $\boldsymbol{A}$ is denoted by $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{B}$, where $\boldsymbol{B} \in \mathbb{R}^{d\times K}$ is a matrix with elements

$$\boldsymbol{B}_{i,k} = \sum_{a}^{d}\sum_{b}^{K} \boldsymbol{\Phi}_{i,a,b,k}\boldsymbol{A}_{a,b}$$

The definition of a tensor acting on a tensor is given in Definition 18 in Appendix 3.C. Tensors are associative in the sense that $\boldsymbol{\Phi}(\boldsymbol{\Psi}(\boldsymbol{A})) = \boldsymbol{\Phi}(\boldsymbol{\Psi})(\boldsymbol{A})$ where $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d\times K\times K}$, see Lemma 21.

The tensor product between matrices $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{d\times K}$ is defined as $(\boldsymbol{A}\otimes\boldsymbol{B}) = \boldsymbol{\Upsilon} \in \mathbb{R}^{d\times d\times K\times K}$, which has elements $\boldsymbol{\Upsilon}_{i,i',k,k'} = \boldsymbol{A}_{i,k}\boldsymbol{B}_{i',k'}$. We define an identity tensor $\boldsymbol{\mathcal{I}}$, which satisfies $\boldsymbol{\mathcal{I}}(\boldsymbol{A}) = \boldsymbol{A}$ for all $\boldsymbol{A}$. If the inverse exists, then the inverse of tensor $\boldsymbol{\Phi}$ is denoted by $\boldsymbol{\Phi}^{-1}$, which satisfies $\boldsymbol{\Phi}^{-1}(\boldsymbol{\Phi}) = \boldsymbol{\mathcal{I}}$. The central equality, which we use in our novel estimator, can be found in Lemma 8 whose proof is in Appendix 3.C.

Lemma 8. *$\boldsymbol{ABC}^\top = (\boldsymbol{A}\otimes\boldsymbol{C})(\boldsymbol{B})$, where $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ are matrices of appropriate size.*

With the above definitions and equalities, we are ready to construct our unbiased estimator defined as

$$\hat{\boldsymbol{\Theta}}_t = \boldsymbol{\Psi}_t^{-1}\Big(\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t\boldsymbol{a}_t\boldsymbol{a}_t^\top\Big) \tag{3.10}$$

where $\boldsymbol{\Psi}_t = \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{a}_t}\Big[\Big(\boldsymbol{x}_t\boldsymbol{x}_t^\top\otimes\boldsymbol{a}_t\boldsymbol{a}_t^\top\Big)\;\Big|\;\mathcal{F}_t\Big]$.

In Lemma 50 in Appendix 3.C we show that the inverse of tensor $\boldsymbol{\Psi}_t$ indeed exists. To see that $\hat{\boldsymbol{\Theta}}_t$ is unbiased, we use Lemma 8 to see that, conditioned on $\mathcal{F}_t$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{a}_t}\Big[\hat{\boldsymbol{\Theta}}_t\;\Big|\;\mathcal{F}_t\Big] = \mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{a}_t}\Big[\boldsymbol{\Psi}_t^{-1}\Big(\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t\boldsymbol{a}_t\boldsymbol{a}_t^\top\Big)\;\Big|\;\mathcal{F}_t\Big]\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{a}_t}\Big[\boldsymbol{\Psi}_t^{-1}\Big((\boldsymbol{x}_t\boldsymbol{x}_t^\top\otimes\boldsymbol{a}_t\boldsymbol{a}_t^\top)(\boldsymbol{\Theta}_t)\Big)\;\Big|\;\mathcal{F}_t\Big]\\
&= \boldsymbol{\Psi}_t^{-1}\Big(\boldsymbol{\Psi}_t(\boldsymbol{\Theta}_t)\Big)\;.
\end{aligned}$$

Now, using the associative property of tensors and the definition of inverses of tensors, we can see that

$$\boldsymbol{\Psi}_t^{-1}\Big(\boldsymbol{\Psi}_t(\boldsymbol{\Theta}_t)\Big) = \boldsymbol{\Psi}_t^{-1}\Big(\boldsymbol{\Psi}_t\Big)(\boldsymbol{\Theta}_t) = \boldsymbol{\mathcal{I}}(\boldsymbol{\Theta}_t) = \boldsymbol{\Theta}_t.$$

As in the semi-bandit setting, we make use of the MGR algorithm outlined in Section 3.3.2 to construct an unbiased estimator without having to evaluate an expectation over the context explicitly. In particular, we aim to apply the MGR to $\boldsymbol{\Psi}_t^{-1}$. Unfortunately, the MGR algorithm as stated in Section 3.2 only works

Require: $\eta > 0, \gamma \in (0,1), M > 0, \beta > 0$

1: Find probability distribution μ over $\mathcal{A}$ as defined by the Kiefer-Wolfowitz theorem (Theorem 49)
2: **for** t in $[T]$ **do**
3: Observe $\boldsymbol{x}_t$ and for all $\boldsymbol{a} \in \mathcal{A}$ set

$$\tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a}) = \exp\left(-\eta \sum_{\tau=1}^{t-1} \boldsymbol{x}_t^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{a}\right) \tag{3.12}$$

4: Draw $\boldsymbol{a}_t$ from

$$\pi_t(\boldsymbol{a}|\boldsymbol{x}_t) = \frac{(1-\gamma)\tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a})}{\sum_{\boldsymbol{a}' \in \mathcal{A}} \tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a}')} + \gamma \mu_{\boldsymbol{a}} \tag{3.13}$$

5: Observe loss $\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}$ and compute $\tilde{\boldsymbol{\Theta}}_t$ using (3.11) and Sampling Scheme 7.
6: **end for**

Algorithm 6: Exp3-Tensor

for matrices. To resolve this issue, we temporarily flatten the tensor to a matrix (Definition 31). It turns out that the MGR algorithm applied to the flattened samples of the expectation inside $\boldsymbol{\Psi}_t$ returns a matrix which we can then unflatten (Definition 32) into an inverse of $\boldsymbol{\Psi}_t$, with small bias.

The estimator that makes use of the MGR algorithm to estimate $\boldsymbol{\Psi}_t^{-1}$ is defined by

$$\tilde{\boldsymbol{\Theta}}_t = \hat{\boldsymbol{\Psi}}_t^+ \left(\boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t \boldsymbol{a}_t^\top\right) \tag{3.11}$$

where $\hat{\boldsymbol{\Psi}}_t^+ = \mathrm{MGR}(S(\mathcal{D}, \pi_t), \beta, M)^U$ is the unflattened output of the MGR when executed on S, the Sampling Scheme 7 as defined in Appendix 3.D. The number M of iterations the MGR is run for and β are hyper-parameters of the algorithm set according to Theorem 52.

As in the semi-bandit setting we need to control the variance of the estimator, which is done by ensuring that the eigenvalues of the tensor $\boldsymbol{\Psi}_t^{-1}$ are not too large, see Definition 41 for the definition of eigenvalues of tensors. We ensure this by employing an exploration distribution $\boldsymbol{\mu}$ over $\mathcal{A}$ based on the Kiefer-Wolfowitz theorem (see also Theorem 49 in the appendix). And the result is shown by arguing that the smallest eigenvalue of the flattened tensor $\boldsymbol{\Psi}_t^F$ is properly bounded, see Lemma 12 below.

The exploration scheme is mixed with a version of Exp3 (Auer, Cesa-Bianchi, Freund, and Schapire, 2002). In particular, we simply run Exp3 on all the actions in $\mathcal{A}$, an approach also used by Cesa-Bianchi and Lugosi (2012a). The full algorithm

is specified in Algorithm 6 and is aptly named Exp3-Tensor. Its regret bound can be found in Theorem 9, whose proof is implied by Theorem 52 in Appendix 3.D.

Theorem 9. *Algorithm 6 with appropriate tuning satisfies Furthermore, let*

$$\gamma = \min\left\{1, \sqrt{K\log(T)\frac{\log(|\mathcal{A}|)}{T\beta\lambda_{\min}^{\boldsymbol{\Sigma}}}}\right\} \qquad \eta = \min\left\{\frac{1}{m(M+1)}, \sqrt{\frac{\log(|\mathcal{A}|)}{Tm^2Kd}}\right\}$$
$$\beta = \frac{1}{\sigma^2 m} \qquad M = \max\left\{\frac{K\log(T)}{\beta m\lambda_{\min}^{\boldsymbol{\Sigma}}}, \sqrt{\frac{TK\log(T)}{\log(|\mathcal{A}|)\beta\lambda_{\min}^{\boldsymbol{\Sigma}}}}\right\},$$

$$\mathcal{R} \in O\left(m^{3/2}\sqrt{KT\log(K)\max\left\{d, \frac{m\sigma^2\log(T)}{\lambda_{\min}^{\boldsymbol{\Sigma}}}\right\}}\right).$$

Computational efficiency. The crucial steps in Exp3-Tensor are the computation of $\boldsymbol{\mu}$ via the Kiefer-Wolfowitz theorem and the computation of (3.12) and (3.13). Computing $\boldsymbol{\mu}$ exactly is not efficient in general. However, there are efficient algorithms that compute approximations to $\boldsymbol{\mu}$ (Lattimore and Szepesvári, 2020, Section 21.2, Note 3). Using this approximation to $\boldsymbol{\mu}$ does not deteriorate the order of regret. The running time for executing the remaining steps is essentially equivalent to the time it takes to run the corresponding steps in CombBand plus the runtime of the MGR procedure. Cesa-Bianchi and Lugosi (2012a) show various concrete examples of action sets $\mathcal{A}$ on which ComBand can be run efficiently. Since the MGR procedure can be run efficiently, this implies that Exp3-Tensor is efficient on a pair $(\mathcal{D}, \mathcal{A})$ whenever a sampling oracle for $\mathcal{D}$ is available and CombBand can be run efficiently on $\mathcal{A}$.

3.4.1 Proof Sketch

In the remainder of this section we provide a sketch of the proof of Theorem 52.

As a first step, observe that we need to control the bias of our estimator: the $\mathbb{E}\left[\boldsymbol{x}_0^\top \boldsymbol{A}_t \boldsymbol{a} \mid \mathcal{F}\right]$ term of Lemma 3. We do precisely this in the following Lemma.

Lemma 10. *Suppose that* $\beta \leq \frac{1}{\lambda_{\max}(\mathbf{\Psi}_t^F)}$. *Then for* $\tilde{\mathbf{\Theta}}_t$ *defined in Equation* (3.11) *and any* $\boldsymbol{a} \in \mathcal{A}$ *and any* $\boldsymbol{x}$ *in the support of* $\mathcal{D}$

$$\boldsymbol{x}^\top(\mathbf{\Theta}_t - \tilde{\mathbf{\Theta}}_t)\boldsymbol{a} \leq \sigma G\sqrt{m}\exp\left(-M\beta\frac{\gamma K\lambda_{\min}^{\mathbf{\Sigma}}}{m}\right) .$$

The proof of Lemma 10 can be found in Appendix 3.D. The proof is very similar to the proof of Lemma 5, with the main difference being the fact that we need to carefully track the effect of the flattening and unflattening operations.

Another part of the proof is bounding the regret of Exp3. The following result can be derived from the standard Exp3 analysis (Auer, Cesa-Bianchi, Freund, and Schapire, 2002) and is provided in Lemma 11.

Lemma 11. *Fix any* $\boldsymbol{x} \in \mathcal{X}$ *and suppose that* $\tilde{\mathbf{\Theta}}_t$ *and* $\eta > 0$ *are such that* $\max_t \eta|\boldsymbol{x}^\top\tilde{\mathbf{\Theta}}_t\boldsymbol{a}| < 1$ *for all* $\boldsymbol{a} \in \mathcal{A}$. *Then the regret of Algorithm 6 in context* $\boldsymbol{x}$ *satisfies*

$$\tilde{\mathcal{R}}_T(\boldsymbol{x}) \leq \frac{\log(|\mathcal{A}|)}{\eta} + \gamma\, U_T(\boldsymbol{x}) + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi_t(\cdot|\boldsymbol{x})}\left[(\boldsymbol{x}^\top\tilde{\mathbf{\Theta}}_t\boldsymbol{a})^2|\mathcal{F}\right]\right]$$

where $U_T(\boldsymbol{x}) = \sum_{t=1}^{T}\sum_{\boldsymbol{a}\in\mathcal{A}}\boldsymbol{\mu}_{\boldsymbol{a}}\boldsymbol{x}^\top\tilde{\mathbf{\Theta}}_t\Big(\boldsymbol{a} - \pi_T^*(\boldsymbol{x})\Big)$ *and* $\boldsymbol{\mu}$ *is the distribution on* $\mathcal{A}$ *defined by the Kiefer-Wolfowitz theorem.*

To ensure that we can apply Lemma 11, we only need to show that our learning rate is chosen correctly and that our estimator $\tilde{\mathbf{\Theta}}_t$ behaves nicely enough, which essentially boils down to controlling the smallest eigenvalue of the flattened tensor $\mathbf{\Psi}_t^F$. This is shown in Lemma 12.

Lemma 12. *For all* $t \geq 1$,

$$\lambda_{\min}(\mathbf{\Psi}_t^F) \geq \frac{\gamma K\lambda_{\min}(\mathbf{\Sigma})}{m}$$

Moreover, for $\eta \leq \frac{1}{m(M+1)}$, *any* $\boldsymbol{a} \in \mathcal{A}$, *and any* $\boldsymbol{x}$ *in the support of* $\mathcal{D}$ *it also holds that* $\eta\left|x^\top\tilde{\mathbf{\Theta}}_t\boldsymbol{a}\right| < 1$.

While the regret of Exp3 is relatively straightforward to control with the standard importance weighted estimator, here we face a complicated variance term $\mathbb{E}_{\boldsymbol{a}\sim\pi_t(\cdot,\boldsymbol{x})}\left[(\boldsymbol{x}^\top\tilde{\mathbf{\Theta}}_t\boldsymbol{a})^2 \mid \mathcal{F}\right]$ due to our choice of estimator $\tilde{\mathbf{\Theta}}_t$. Not only is $\tilde{\mathbf{\Theta}}_t$ biased, but we also need to significantly manipulate the $(\boldsymbol{x}^\top\tilde{\mathbf{\Theta}}_t\boldsymbol{a})^2$ term in order to recover an expression resembling the term we would have if we were to use the standard importance weighted estimator. We do so in Appendix 3.D, which leads to the following result.

Lemma 13. *Fix a $t \in [T]$ and let $\boldsymbol{a}_0 \sim \pi_t(\cdot|\boldsymbol{x}_0)$. Then*

$$\mathbb{E}\left[(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}_0)^2 \,\middle|\, \mathcal{F}\right] \leq 2m^2 Kd$$

To finish the proof of Theorem 9, we only need to assemble the pieces we have collected so far. Applying Lemma 13 to the right-hand side of Lemma 11 gives.

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_0}[\hat{\mathcal{R}}_T(\boldsymbol{x}_0)] \leq \frac{\log(|\mathcal{A}|)}{\eta} + \eta T m^2 Kd + \gamma U_T(\boldsymbol{x}_0)$$

The remaining steps are applying this result to Lemma 3, applying Lemma 10, and tuning η, γ, M and β accordingly. All details can be found in Theorem 52 in Appendix 3.D.

3.5 Lower Bounds

In this section we provide lower bounds for both the full- and semi-bandit settings. All details related to the results in this section can be found in Appendix 3.E. Our lower bounds hold for a large class of algorithms which we call orthogonal algorithms. Informally, if two contexts $\boldsymbol{x}_\tau$ and $\boldsymbol{x}_t$ are orthogonal, then orthogonal algorithms do not use information from round $\tau < t$ to compute a prediction for round t. Essentially all algorithms using the estimators in Sections 3.3 and 3.4 are orthogonal algorithms. The lower bound for the semi-bandit setting is provided below.

Theorem 14. *In the semi-bandit setting, any orthogonal algorithm must suffer $\Omega(\sqrt{dmKT})$ regret.*

Theorem 14 is implied by Theorem 54, whose proof follows from a reduction to online learning with stochastic feedback graphs (Esposito, Fusco, Hoeven, and Cesa-Bianchi, 2022).

For the full-bandit setting the result can be found below. The result is implied by Theorem 55, whose proof follows from carefully adapting the lower bound of Cohen, Hazan, and Koren (2017) to work in our setting.

Theorem 15. *In the full-bandit setting, any orthogonal algorithm must suffer $\tilde{\Omega}(m^{3/2}\sqrt{dKT})$ regret.*

Ignoring factors logarithmic in K, our upper bounds in both settings have an extra $\max\left\{d, \frac{m\sigma^2 \log(T)}{\lambda^{\boldsymbol{\Sigma}}_{\min}}\right\}$ factor. Although the term d is captured by our lower bounds, the $\frac{m\sigma^2 \log(T)}{\lambda^{\boldsymbol{\Sigma}}_{\min}}$ term is not. In particular, in the construction of our lower bounds $\sigma = 1$ and $\lambda^{\boldsymbol{\Sigma}}_{\min} = 1/d$. Thus, with the same set of losses as in the lower bound, our algorithms suffer $\tilde{O}(m^2\sqrt{dKT})$ regret in the full-bandit setting and $\tilde{O}(m\sqrt{dKT})$ regret in the semi-bandit setting. This implies that for $m = 1$ our algorithms as well as the algorithm of Neu and Olkhovskaya (2020) have a tight regret bound. However, a gap of $\sqrt{m}$ exists when m is a parameter of the problem.

3.6 Experiments

The full code for the experiments can be found here[4].

To the best of our knowledge, our work is the first one in this setting, and thus there are no natural strong baselines to compare against in experiments. Our baselines are RealLinExp3 (Neu and Olkhovskaya, 2020) and two versions of CombBand (Cesa-Bianchi and Lugosi, 2012a). RealLinExp3 uses contextual information but ignores any structure in the action set, implying that each action is treated independently from the others. We compare to two versions of ComBand: Running one ComBand instance per context (ComBand OPC) and the vanilla ComBand that ignores contexts entirely. Both of these are able to exploit the combinatorial nature of actions. Although our algorithms can handle arbitrary context distributions (no expectation needs to be computed thanks to the MGR procedure), to accommodate ComBand OPC, we run our experiments on a finitely supported distribution $\mathcal{D}$.

Let $\mathcal{B}_{K,m} = \{\boldsymbol{x} : \boldsymbol{x} \in \{0,1\}^K, \|\boldsymbol{x}\|_1 = m\}$ be the set containing all m-sized subsets of a base set of K elements. We use $\mathcal{B}_{d,m}$ with $(d,m) \in \Big\{(3{,}1), (5{,}2), (12{,}3)\Big\}$ to define context spaces $\mathcal{X}$, and $\mathcal{B}_{K,m}$ with $(K,m) \in \Big\{(3{,}1), (5{,}2), (8{,}3)\Big\}$ to define action sets $\mathcal{A}$. Our experiments are run over a length of 10^5 timesteps and averaged over 10 repetitions, with the relatively modest sample size induced by computational constraints. The losses $\boldsymbol{\Theta}_t$ are generated as follows: for each $i \in [d]$ we choose m subactions that are "good actions". This means that $(\boldsymbol{\Theta}_t)_{i,j}$ is drawn from a Bernoulli(0.4) distribution if j is a good action in i and from a Bernoulli(0.5) distribution otherwise. The contexts are drawn uniformly at random.

In our experiments, we did not use the MGR procedure for any of the algorithms. While MGR enjoys a $O(MKd + Kd^2)$ scaling in computational cost (Neu and Olkhovskaya, 2020), due to the nature of the context distribution using the unbiased estimator turns out to be significantly faster, as the M factor in the scaling of the runtime of MGR is of order $1/\gamma$ and there exists a highly optimized implementation of matrix inversion in NumPy (Harris, Millman, Walt, Gommers, Virtanen, Cournapeau, Wieser, Taylor, Berg, Smith, Kern, Picus, Hoyer, Kerkwijk, Brett, Haldane, Río, Wiebe, Peterson, Gérard-Marchant, Sheppard, Reddy, Weckesser, Abbasi, Gohlke, and Oliphant, 2020). We stress that this direct computation is only possible due to the simple setting and if the expectation cannot be computed then the MGR is necessary.

[4] https://github.com/LukasZierahn/Combinatorial-Contextual-Bandits

3.6.1 Full-Bandit Setting

The results for the full-bandit setting with theoretical tuning can be found in Figure 3.F.1. As expected the comparative performance of RealLinExp3 deteriorates with a more complicated combinatorial element and improves with a more difficult contextual element. Curiously, non-contextual ComBand sometimes outperforms ComBand OPC. It is possible that some actions are by random chance better across multiple contexts and thus ignoring contexts can lead to better results in the short term as the algorithm is able to exploit those better actions. This phenomenon appears most prevalent when the number of contexts is large, i.e., in the (5,2)- and $(12, 3)$-context cases. We use uniform random exploration for Exp3-Tensor to ease implementation.

While the exploration rate for Exp3-Tensor is very large (equal to $\gamma = 51.66\%$ in the case where $d = 12$ and $K = 8$), the algorithm is only marginally falling behind the other algorithms (see Figure 3.F.1) with exploration rate of 1.32% for RealLinExp3, 5.38% for Comband, and 18.64% for the Comband One-Per-Context.

The results for the full-bandit setting with $1/\sqrt{T}$ tuning can be found in Figure 3.F.2. The results are comparable to the ones obtained with theoretical tuning. This means that the results for Exp3-Tensor are on par with or slightly worse than the results of other the other algorithms. Interestingly, RealLinExp3 seems to be performing even better in the settings without a combinatorial aspect but continues to struggle with a strong combinatorial aspect.

Unfortunately, with both theoretical tuning and $1/\sqrt{T}$ tuning Exp3-Tensor was at best on par with the other algorithms. We conjecture that the artificial scenarios with a finite number of contexts, which we designed to accommodate our baselines, are not sufficiently expressive for Exp3-Tensor to exploit and that Exp3-Tensor is especially useful in settings with continuous context distributions.

3.6.2 Semi-Bandit Setting

To efficiently implement Line 4 of Algorithm 3 we use Warmuth and Kuzmin (2008, Algorithm 4).

One of the algorithms we compare with in the semi-bandit setting is a variant of the Online Stochastic Mirror Descent (OSMD) algorithm presented in Audibert, Bubeck, and Lugosi (2014, Figure 3). Specifically, our variant uses the same estimator as that algorithm. However, we use FTRL instead of OSMD. We use exploration parameter $\gamma = \sqrt{\frac{K}{Tm}}$ and learning rate $\eta = \sqrt{\frac{m \log(\frac{K}{m})}{TK}}$. From here on, we refer to this algorithm as Non-Contextual (NC) FTRL. The list of algorithms in the semi-bandit setting is thus CO_2-FTRL, RealLinExp3, NC FTRL, and NC FTRL OPC.

The results for the theoretical learning rates can be found in Figure 3.F.3. RealLinExp3 overlaps with the full-bandit case and the results relative to the other algorithms are similar: RealLinExp3's relative performance decreases as the problem becomes more combinatorial. NC FTRL OPC does well in general and especially so if the number of contexts is small. However, NC FTRL OPC is outperformed by NC FTRL in the (12,3)-context case. CO_2-FTRL is on par with the other algorithms in most experiments, although in some experiments it is outperformed and sometimes it outperforms other algorithms. Similar to the high γ in the full-bandit case, CO_2-FTRL uses $\gamma = 24.11\%$ in the most complicated case which might contribute to regret.

The results for $1/\sqrt{T}$ tuning can be found in Figure 3.F.4. When using the $1/\sqrt{T}$ tuning CO_2-FTRL is on par with the best competitors in the simpler cases and outperforming in the more complicated problem instances. This also suggests that CO_2-FTRL is less sensitive to miss-specified tunings and perhaps performs better with more aggressive tuning.

3.6.3 Conclusion

We conjecture that the relative simple artificial scenarios with a finite number of contexts, which we designed to accommodate our baselines, may not be complicated enough to demonstrate why the rather sophisticated approach of EXP3-Tensor is necessary. More experiments, using more general context spaces and improved tunings, are necessary to gain a better understanding of the algorithms' behavior, especially with continuous contexts.

Unlike EXP3-Tensor, CO_2-FTRL does outperforms the baseline algorithms. We believe the semi-bandit setting being a simpler setting might mean the presented scenarios are complicated enough to distinguish CO_2-FTRL from the baselines.

3.7 Discussion

As already discussed in Section 3.5, our bounds are tight with respect to all parameters except m. More precisely, there is an extra factor $\sqrt{m}$ in the upper bounds for both semi and full bandit settings. We leave open the question whether this extra term is necessary or not.

One direction for future work is to empirically evaluate our algorithm as well as the baselines we specified in Section 3.6 in more general experimental settings. The experiments in Section 3.6 were specified to accommodate the baselines and it would be interesting to understand how all algorithms fare under different circumstances. One such circumstance could be replacing the discrete distribution over context in our experiments with a continuous distribution, even though it is not clear how to accommodate all baselines for such a context distribution.

Appendix

3.A Matrix Geometric Resampling - Proofs

Lemma 4 (RESTATED). *Let $\hat{\boldsymbol{P}}^+$ be defined by the MGR procedure (Algorithm 4) run for M iterations where each $\hat{\boldsymbol{P}}_k \in \mathbb{R}^{b\times b}$ drawn in Step 2 of Algorithm 4 is positive semi-definite, and such that $\mathbb{E}[\hat{\boldsymbol{P}}_k] = \boldsymbol{P}$, where $\boldsymbol{P}$ is also symmetric and positive semi-definite. Choose β such that $\beta \leq \frac{1}{\lambda_{\max}(\hat{\boldsymbol{P}}_k)}$ with probability 1 for all k, then the following three results hold*

1. $\operatorname{tr}\Big(\mathbb{E}_{\mathrm{MGR}}[\boldsymbol{P}\hat{\boldsymbol{P}}^{+\top}\boldsymbol{P}\hat{\boldsymbol{P}}^{+}]\Big) < 2b$

2. $\mathbb{E}_{\mathrm{MGR}}[\hat{\boldsymbol{P}}^{+}]\boldsymbol{P} = \boldsymbol{I} - (\boldsymbol{I} - \beta\boldsymbol{P})^M$ *and* $\boldsymbol{P}\,\mathbb{E}_{\mathrm{MGR}}[\hat{\boldsymbol{P}}^{+}] = \boldsymbol{I} - (\boldsymbol{I} - \beta\boldsymbol{P})^M$

3. $\|\hat{\boldsymbol{P}}^{+}\|_{\mathrm{op}} \leq (M+1)\beta$.

Proof. The second and third statement of the lemma are proven in Section 4.2 of Neu and Olkhovskaya (2020). While a similar statement to the first statement of the lemma can be found in Neu and Olkhovskaya (2020), there is a transpose missing from their statement compared to ours. In particular, since $\hat{\boldsymbol{P}}^{+}$ might not be symmetric it may be that $\hat{\boldsymbol{P}}^{+\top} \neq \hat{\boldsymbol{P}}^{+}$. Therefore, we need to prove the first statement of the lemma.

A central part of our consideration of that will be $\hat{\boldsymbol{P}}^{+\top}$, which we explore now. For that we will employ the definitions of $\hat{\boldsymbol{P}}^{+\top}$ and $\boldsymbol{C}_k$ as given by the MGR procedure in Algorithm 4 and the fact that $(\boldsymbol{I} - \beta\hat{\boldsymbol{P}}_k)$ is symmetric as $\hat{\boldsymbol{P}}_k$ is

symmetric by assumption.

$$\begin{aligned}
\hat{\boldsymbol{P}}^{+\top} &= \left(\beta \sum_{k=0}^{M} \boldsymbol{C}_k\right)^{\top} \\
&= \beta \sum_{k=0}^{M} \boldsymbol{C}_k^{\top} \\
&= \beta \sum_{k=0}^{M} \left(\prod_{j=1}^{k} (\boldsymbol{I} - \beta \hat{\boldsymbol{P}}_k)\right)^{\top} \\
&= \beta \sum_{k=0}^{M} \prod_{j=1}^{k} (\boldsymbol{I} - \beta \hat{\boldsymbol{P}}_{k-j+1})^{\top} \\
&= \beta \sum_{k=0}^{M} \prod_{j=1}^{k} \underbrace{(\boldsymbol{I} - \beta \hat{\boldsymbol{P}}_{k-j+1})}_{\boldsymbol{D}_{k-j+1}},
\end{aligned}$$

where we also introduced the notation $\boldsymbol{D}_j = (\boldsymbol{I} - \beta \hat{\boldsymbol{P}}_j)$.

Now we are equipped to focus on $\hat{\boldsymbol{P}}^{+\top} \boldsymbol{P} \hat{\boldsymbol{P}}^{+}$, which we do by using the above equation for $\hat{\boldsymbol{P}}^{+\top}$. We multiply out to obtain

$$\begin{aligned}
\hat{\boldsymbol{P}}^{+\top} \boldsymbol{P} \hat{\boldsymbol{P}}^{+} &= \left(\beta \sum_{k=0}^{M} \prod_{j=1}^{k} \boldsymbol{D}_{k-j+1}\right) \boldsymbol{P} \left(\beta \sum_{k=0}^{M} \prod_{j=1}^{k} \boldsymbol{D}_j\right) \\
&= \beta^2 \sum_{k=0}^{M} \sum_{k'=0}^{M} \left(\prod_{j=1}^{k} \boldsymbol{D}_{k-j+1}\right) \boldsymbol{P} \left(\prod_{j=1}^{k'} \boldsymbol{D}_j\right).
\end{aligned}$$

We can write out

$$\left(\prod_{j=1}^{k} \boldsymbol{D}_{k-j+1}\right) \boldsymbol{P} \left(\prod_{j=1}^{k'} \boldsymbol{D}_j\right) = \boldsymbol{D}_k \boldsymbol{D}_{k-1} \ldots \boldsymbol{D}_2 \boldsymbol{D}_1 \boldsymbol{P} \boldsymbol{D}_1 \boldsymbol{D}_2 \ldots \boldsymbol{D}_{k'-1} \boldsymbol{D}_{k'} \,, \quad (3.14)$$

which will help illustrates the effects of taking the expectation over $\hat{\boldsymbol{P}}_j$ for all $j \in [M]$. We start by looking at the individual terms of the sum. If $k \leq k'$ then there are $k'-k$ terms $\boldsymbol{D}_j$ that appear in $\prod_{j=1}^{k'} \boldsymbol{D}_j$ but not in $\prod_{j=1}^{k} \boldsymbol{D}_{k-j+1}$. Similarly, if $k' \leq k$ then there are $k - k'$ terms $\boldsymbol{D}_j$ that appear in $\prod_{j=1}^{k} \boldsymbol{D}_{k-j+1}$ but not in $\prod_{j=1}^{k'} \boldsymbol{D}_j$. Define $k_{\min} = \min(k, k')$ and $k_{\max} = \max(k, k')$, then by linearity of the expectation and the tower rule we have that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\hat{\boldsymbol{P}}^{+\top}\boldsymbol{P}\hat{\boldsymbol{P}}^{+}\right] \\
&= \mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}\left(\prod_{j=1}^{k}\boldsymbol{D}_{k-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k'}\boldsymbol{D}_{j}\right)\right] \\
&= \beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\left(\prod_{j=1}^{k}\boldsymbol{D}_{k-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k'}\boldsymbol{D}_{j}\right)\right] \\
&= \beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_1,\ldots,\hat{\boldsymbol{P}}_{k_{\min}}}\left[\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_{k_{\min}},\ldots,\hat{\boldsymbol{P}}_{k_{\max}}}\left[\left(\prod_{j=1}^{k}\boldsymbol{D}_{k-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k'}\boldsymbol{D}_{j}\right)\middle|\hat{\boldsymbol{P}}_1,\ldots,\hat{\boldsymbol{P}}_{k_{\min}}\right]\right] \\
&= \beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k]_+}\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_1,\ldots,\hat{\boldsymbol{P}}_{k_{\min}}}\left[\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{k_{\min}-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{j}\right)\right](\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k']_+}\,,
\end{aligned}
\tag{3.15}
$$

where $[x]_+ = \max\{x, 0\}$. Next, fix some $k, k' \in [M]$, now we will just inspect the expectation in the previous equation. Here we recall Equation 3.14 and then move from the inside out, pick some $j \in [k]$, then for any matrix $\boldsymbol{H}$ that is commutative with $\boldsymbol{P}$, i.e., $\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{H}$ we can see that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_j}[\boldsymbol{D}_j\boldsymbol{H}\boldsymbol{D}_j] &= \mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_j}[(\boldsymbol{I}-\beta\hat{\boldsymbol{P}}_j)\boldsymbol{H}(\boldsymbol{I}-\beta\hat{\boldsymbol{P}}_j)] \\
&= \mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_j}[\boldsymbol{H}-\beta\hat{\boldsymbol{P}}_j\boldsymbol{H}-\beta\boldsymbol{H}\hat{\boldsymbol{P}}_j+\beta^2\hat{\boldsymbol{P}}_j\boldsymbol{H}\hat{\boldsymbol{P}}_j] \\
&\preceq \mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_j}[\boldsymbol{H}-\beta\hat{\boldsymbol{P}}_j\boldsymbol{H}-\beta\boldsymbol{H}\hat{\boldsymbol{P}}_j+\beta\boldsymbol{H}\hat{\boldsymbol{P}}_j] \\
&= \boldsymbol{H}-\beta\boldsymbol{P}\boldsymbol{H} \\
&= (\boldsymbol{I}-\beta\boldsymbol{P})\boldsymbol{H},
\end{aligned}
$$

where we used $\hat{\boldsymbol{P}}_j \preceq \lambda_{\max}(\hat{\boldsymbol{P}}_j)\boldsymbol{I} \preceq \beta^{-1}\boldsymbol{I}$ for the inequality. If $\boldsymbol{H}$ and $\boldsymbol{P}$ commute then so do $\boldsymbol{H}-\beta\boldsymbol{P}\boldsymbol{H}$ and $\boldsymbol{P}$. Thus we can now use the above idea recursively $k_{\min}$ times in total, starting with $\boldsymbol{H} = \boldsymbol{P}$.

$$
\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_1,\ldots,\hat{\boldsymbol{P}}_k}\left[\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{k_{\min}-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{j}\right)\right] \preceq \boldsymbol{P}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\min}}
$$

Plugging this into Equation (3.15) we find

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\hat{\boldsymbol{P}}^{+\top}\boldsymbol{P}\hat{\boldsymbol{P}}^{+}\right] \\
&= \mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\beta^2\left(\prod_{j=1}^{k}\boldsymbol{D}_{k-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k'}\boldsymbol{D}_{j}\right)\right] \\
&= \beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}\mathop{\mathbb{E}}_{\hat{\boldsymbol{P}}_1,\ldots,\hat{\boldsymbol{P}}_{k_{\min}}}\left[(\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k]_+}\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{k_{\min}-j+1}\right)\boldsymbol{P}\left(\prod_{j=1}^{k_{\min}}\boldsymbol{D}_{j}\right)(\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k']_+}\right] \\
&\preceq \beta^2\sum_{k=0}^{M}\sum_{k'=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k]_+}\boldsymbol{P}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\min}}(\boldsymbol{I}-\beta\boldsymbol{P})^{[k_{\max}-k']_+} \\
&= \beta^2\boldsymbol{P}\sum_{k=0}^{M}\sum_{k'=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}}. \qquad (3.16)
\end{aligned}
$$

Let $\boldsymbol{B}_{k',k} = (\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}}$. We can order the double sum as $\sum_{k=0}^{M}\sum_{k'=0}^{M}\boldsymbol{B}_{k,k'} = 2\sum_{k=0}^{M}\sum_{k'=k}^{M}\boldsymbol{B}_{k,k'} - \sum_{k=0}^{M}\boldsymbol{B}_{k,k}$ since $\boldsymbol{B}_{k,k'} = \boldsymbol{B}_{k',k}$. Thus, by using that $(\boldsymbol{I}-\beta\boldsymbol{P}) \preceq \boldsymbol{I}$, which follows from $\beta \leq \frac{1}{\lambda_{\max}(\boldsymbol{P})}$, we can see that

$$
\begin{aligned}
\sum_{k=0}^{M}\sum_{k'=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}} &= 2\sum_{k=0}^{M}\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}} - \sum_{k=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k} \\
&\preceq 2\sum_{k=0}^{M}\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}} \\
&= 2\sum_{k=0}^{M}\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k'} \\
&= 2\sum_{k=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k}\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k'-k}
\end{aligned}
$$

where in the third equality we use that $k_{\max} = \max(k,k') = k'$. Re-indexing $\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k'-k}$ and applying the Neummann series like Equation (3.8) together with the fact that both $\boldsymbol{I}-\beta\boldsymbol{P}$ and $\boldsymbol{P}$ are positive semi-definite and then the same Equation (3.8) again, we can follow that

$$
\begin{aligned}
2\sum_{k=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k}\sum_{k'=k}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k'-k} &= 2\sum_{k=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k}(\beta^{-1}\boldsymbol{P}^{-1} - (\boldsymbol{I}-\beta\boldsymbol{P})^{M-k}\beta^{-1}\boldsymbol{P}^{-1}) \\
&\preceq 2\sum_{k=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k}\beta^{-1}\boldsymbol{P}^{-1} \\
&= 2(\beta^{-1}\boldsymbol{P}^{-1} - (\boldsymbol{I}-\beta\boldsymbol{P})^{M}\beta^{-1}\boldsymbol{P}^{-1})\beta^{-1}\boldsymbol{P}^{-1} \\
&\preceq 2\beta^{-2}\boldsymbol{P}^{-2}. \qquad (3.17)
\end{aligned}
$$

Using Equations (3.17) and (3.16) we may conclude the proof as

$$\begin{aligned}\operatorname{tr}\Big(\mathop{\mathbb{E}}_{\mathrm{MGR}}[\boldsymbol{P}^\top \hat{\boldsymbol{P}}^{+\top}\boldsymbol{P}\hat{\boldsymbol{P}}^{+}]\Big) &\leq \operatorname{tr}\left(\boldsymbol{P}^\top\beta^2\boldsymbol{P}\sum_{k=0}^{M}\sum_{k'=0}^{M}(\boldsymbol{I}-\beta\boldsymbol{P})^{k_{\max}}\right)\\ &\leq 2\operatorname{tr}\left(\boldsymbol{P}^\top\beta^2\boldsymbol{P}(\beta^{-2}\boldsymbol{P}^{-2})\right)\\ &= 2\operatorname{tr}\left(\boldsymbol{P}\beta^2\boldsymbol{P}\beta^{-2}\boldsymbol{P}^{-2}\right)\\ &= 2b.\end{aligned}$$

□

Lemma 5 (RESTATED). *Let* $\beta \leq \min_{t\in[T],k\in[K]}\frac{1}{\lambda_{\max}(\boldsymbol{\Sigma}_{t,k})}$ *and* $\tilde{\boldsymbol{\Theta}}_t$ *as defined in Equation* (3.9)*. For all* $\boldsymbol{a}\in\mathcal{A}$ *and all* $\boldsymbol{x}\in\mathcal{X}$ *Algorithm 3 guarantees*

$$\mathbb{E}\left[\boldsymbol{x}^\top(\boldsymbol{\Theta}_t-\tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a}\,\middle|\,\mathcal{F}\right] \leq m\sigma R\exp\left(-\frac{M\beta\gamma}{|\mathcal{E}|}\lambda^{\boldsymbol{\Sigma}}_{\min}\right)\ ,$$

for all $t\in[T]$*, where* $\mathcal{E}$ *is the exploration set.*

Proof. We focus on $\tilde{\boldsymbol{\Theta}}$ first and start by plugging in the definition of $\tilde{\boldsymbol{\Theta}}$ (Equation (3.4)), splitting the expectation due to the independence of the MGR process and then applying Fact 2 of Lemma 4

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}\,\middle|\,\mathcal{F}_t\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\boldsymbol{x}^\top(\tilde{\boldsymbol{\Theta}}_t)_{\cdot,k}(\boldsymbol{a})_k\,\middle|\,\mathcal{F}_t\right]\\ &= \mathbb{E}\left[\sum_{k=1}^{K}\boldsymbol{x}^\top\hat{\boldsymbol{\Sigma}}^{+}_{t,k}\boldsymbol{x}_t(\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k(\boldsymbol{a})_k\,\middle|\,\mathcal{F}_t\right]\\ &= \sum_{k=1}^{K}\boldsymbol{x}^\top\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\hat{\boldsymbol{\Sigma}}^{+}_{t,k}\right]\mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{a}_t}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top(\boldsymbol{a}_t)_k\,\middle|\,\mathcal{F}_t\right](\boldsymbol{\Theta}_t)_{\cdot,k}(\boldsymbol{a})_k\\ &= \sum_{k=1}^{K}\boldsymbol{x}^\top(\boldsymbol{I}-(\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k})^M)(\boldsymbol{\Theta}_t)_{\cdot,k}(\boldsymbol{a})_k\ .\end{aligned}$$

We plug this into the initial quantity of interest

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}^\top(\boldsymbol{\Theta}_t-\tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a}\,\middle|\,\mathcal{F}_t\right] &= \sum_{k=1}^{K}\boldsymbol{x}^\top(\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k})^M(\boldsymbol{\Theta}_t)_{\cdot,k}(\boldsymbol{a})_k\\ &= \boldsymbol{x}^\top(\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k})^M\boldsymbol{\Theta}_t\boldsymbol{a}\\ &\leq \|\boldsymbol{x}\|_2\left\|(\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k})^M\right\|_{\mathrm{op}}\|\boldsymbol{\Theta}_t\boldsymbol{a}\|_2\ ,\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm for matrices. Since $\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k}$ is positive definite, by the choice of β, the operator norm is equal to the largest eigenvalue

$$\left\|(\boldsymbol{I}-\beta\boldsymbol{\Sigma}_{t,k})^M\right\|_{\mathrm{op}} = (1-\beta\lambda_{\min}(\boldsymbol{\Sigma}_{t,k}))^M \leq \exp(-M\beta\lambda_{\min}(\boldsymbol{\Sigma}_{t,k}))\ ,$$

where we also used $1 + x \leq \exp(x)$, which holds for all x. We also find that

$$\|\boldsymbol{\Theta}_t \boldsymbol{a}\|_2 = \sqrt{\sum_{i=1}^{d} \left(\sum_{k=1}^{K} (\boldsymbol{\Theta}_t)_{i,k} (\boldsymbol{a})_k \right)^2} \leq m \sqrt{\sum_{i=1}^{d} (\boldsymbol{\Theta}_t)_{i,k^*}^2} \leq mR \ ,$$

where we used $k^* := \arg\max_{k \in [K]} \|(\boldsymbol{\Theta}_t)_{\cdot,k}\|_2$ and the definition of R. Together with the assumptions that $\|\boldsymbol{x}\|_2 \leq \sigma$, we thus obtain

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{x}^\top (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t) \boldsymbol{a} \;\middle|\; \mathcal{F}_t \right] &\leq m\sigma R \exp(-M\beta \lambda_{\min}(\boldsymbol{\Sigma}_{t,k})) \\ &\leq m\sigma R \exp\left(-\frac{M\beta\gamma}{|\mathcal{E}|} \lambda_{\min}^{\boldsymbol{\Sigma}} \right) \ , \end{aligned}$$

where we used in the last inequality that by construction of $\mathcal{E}$, π_t guarantees that $\mathbb{P}_t\Big((\boldsymbol{a}_t)_k = 1\Big) \geq \gamma |\mathcal{E}|^{-1}$ and thus $\lambda_{\min}(\boldsymbol{\Sigma}_{t,k}) \geq \gamma \lambda_{\min}^{\boldsymbol{\Sigma}} |\mathcal{E}|^{-1}$, for all t and k. □

3.B Semi-Bandits - Proofs

Lemma 6 (RESTATED)**.** *Let $\beta = \frac{1}{\sigma^2}$ and let $\eta \leq \frac{\log(2)}{M+1}$. Let R as in Equation (3.6), then we have that for any $\boldsymbol{x} \in \mathcal{X}$, with $\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}))$ defined by (3.5) and any $\boldsymbol{u} \in \mathcal{A}$, it holds that*

$$\sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t (\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x})) - \boldsymbol{u}) \leq \frac{m\left(1 + \log\left(\frac{K}{m}\right)\right)}{\eta} + \eta \sum_{t=1}^{T} \sum_{k=1}^{K} (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t)_k^2 (\boldsymbol{w}_t(\boldsymbol{x}))_k \ .$$

Proof. Fix any $\boldsymbol{x} \in \mathcal{X}$. We define the unconstrained potential as

$$\tilde{\boldsymbol{w}}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x})) = \arg\min_{\boldsymbol{v} \in \mathbb{R}^K} \sum_{\tau=1}^{t-1} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{v} + \frac{1}{\eta} R(\boldsymbol{v}) \ , \tag{3.18}$$

We will use Orabona (2019, Lemma 7.16) to show the result. We restate a less general version in our notation here.

Lemma 16 (Orabona (2019, Lemma 7.16))**.** *Let $\mathcal{A}$ to be non-empty and closed and $\arg\min_{\boldsymbol{a} \in \mathcal{A}} F_t(\boldsymbol{a})$ exists and is non-empty. Assume $R(\cdot)$ is twice differentiable with a positive definite Hessian in the interior of its domain. Then, for all $t \in [T]$ there exists a $\boldsymbol{z}_t$ on the line segment between $\boldsymbol{w}(\tilde{\boldsymbol{L}}_t(\boldsymbol{x}))$ and $\tilde{\boldsymbol{w}}(\tilde{\boldsymbol{L}}_{t+1}(\boldsymbol{x}))$ such that the following holds for any $\boldsymbol{u} \in \mathcal{A}$*

$$\sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t (\boldsymbol{w}_t(\tilde{\boldsymbol{L}}_t(\boldsymbol{x})) - \boldsymbol{u}) \leq R(\boldsymbol{u}) - R(\boldsymbol{w}_1(\tilde{\boldsymbol{L}}_1(\boldsymbol{x}))) + \frac{1}{2} \sum_{t=1}^{T} \|\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t\|^2_{\left(\frac{\partial^2}{(\partial \boldsymbol{z}_t)^2} R(\boldsymbol{z}_t)\right)^{-1}},$$

where $\|\boldsymbol{v}\|^2_{\left(\frac{\partial^2}{(\partial \boldsymbol{z}_t)^2} R(\boldsymbol{z}_t)\right)^{-1}} = \boldsymbol{v}^\top \left(\frac{\partial^2}{(\partial \boldsymbol{z}_t)^2} R(\boldsymbol{z}_t)\right)^{-1} \boldsymbol{v}$.

We recall that $R(\boldsymbol{v}) = \frac{1}{\eta}\sum_{k=1}^{K}\Big((\boldsymbol{v})_k \log(\boldsymbol{v})_k - (\boldsymbol{v})_k\Big)$ and since the Hessian of R at $\boldsymbol{v}$ is a diagonal matrix with diagonal elements $1/\boldsymbol{v}(1), \dots, 1/\boldsymbol{v}(K)$ the Hessian is positive definite for all $\boldsymbol{v} \in \operatorname{int}\Big(\operatorname{Conv}(\mathcal{A})\Big)$ and thus the requirements of the lemma are met. We proceed to bound $-R(\boldsymbol{v})$ and for any $\boldsymbol{v} \in \operatorname{Conv}(\mathcal{A})$ we have that

$$\begin{aligned}
-R(\boldsymbol{v}) &= -\frac{1}{\eta}\sum_{k=1}^{K}\Big((\boldsymbol{v})_k \log(\boldsymbol{v})_k - (\boldsymbol{v})_k\Big) \\
&= \frac{1}{\eta}\|\boldsymbol{v}\|_1 \sum_{k=1}^{K}\left(\frac{(\boldsymbol{v})_k}{\|\boldsymbol{v}\|_1}\log\frac{1}{(\boldsymbol{v})_k}\right) + \frac{\|\boldsymbol{v}\|_1}{\eta} \\
&\le \frac{1}{\eta}\|\boldsymbol{v}\|_1 \log\left(\sum_{k=1}^{K}\frac{(\boldsymbol{v})_k}{\|\boldsymbol{v}\|_1}\frac{1}{(\boldsymbol{v})_k}\right) + \frac{\|\boldsymbol{v}\|_1}{\eta} \\
&\le \frac{m\Big(1+\log(\frac{K}{m})\Big)}{\eta}, \qquad (3.19)
\end{aligned}$$

where in the second inequality we used Jensen's inequality and in the last inequality we used that $\boldsymbol{x}\log(K/\boldsymbol{x}) + \boldsymbol{x}$ is increasing in $\boldsymbol{x}$ for $\boldsymbol{x} \in [1, K]$ and the assumption that $\|\boldsymbol{v}\|_1 \le m$.

Now, to control the $\|\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\|^2_{\left(\frac{\partial^2}{(\partial z_t)^2}R(z_t)\right)^{-1}}$ term we need to control z_t. Recall that by assumption $|(\boldsymbol{x}^\top\boldsymbol{\Theta}_t)_k| \le 1$ for any $\boldsymbol{x} \in \mathcal{X}$. Therefore we have that

$$\begin{aligned}
\max_{k\in[K]}\left|\eta\boldsymbol{x}^\top(\tilde{\boldsymbol{\Theta}}_t)_k\right| &= \max_{k\in[K]}\left|\eta\boldsymbol{x}^\top(\tilde{\boldsymbol{\Theta}}_t)_k\right| \\
&= \max_{k\in[K]}\left|\eta\boldsymbol{x}^\top\hat{\boldsymbol{\Sigma}}^{+}_{t,k}\boldsymbol{x}_t(\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k\right| \\
&\le \max_{k\in[K]}\left|\eta\boldsymbol{x}^\top\hat{\boldsymbol{\Sigma}}^{+}_{t,k}\boldsymbol{x}_t\right| \\
&\le \eta\sigma^2\max_{k\in[K]}\|\hat{\boldsymbol{\Sigma}}^{+}_{t,k}\|_{\mathrm{op}} \\
&\le \eta\sigma^2\beta(M+1) \\
&\le \log(2),
\end{aligned}$$

where the second inequality is Hölder's inequality, the third inequality is due to Lemma 4, and the last equality holds since by assumption $\eta \le \frac{\log(2)}{(M+1)}$ and $\beta = \frac{1}{\sigma^2}$. Since $(\tilde{\boldsymbol{w}}_{t+1}(\boldsymbol{x}))_k = (\boldsymbol{w}_t(\boldsymbol{x}))_k \exp(-\eta(\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t)_k)$ (Orabona, 2019, Remark 11.1) this means that $(\tilde{\boldsymbol{w}}_{t+1}(\boldsymbol{x}))_k \in [0.5\boldsymbol{w}_t(\boldsymbol{x}), 2\boldsymbol{w}_t(\boldsymbol{x})]$. Now, since $\boldsymbol{z}_t$ is on the line segment between $\tilde{\boldsymbol{w}}_{t+1}(\boldsymbol{x})$ and $\boldsymbol{w}_t(\boldsymbol{x})$, it follows that $(\boldsymbol{z}_t)_k \le 2(\boldsymbol{w}_t(\boldsymbol{x}))_k$ for all $k \in [K]$. This in turn implies that

$$\|\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\|^2_{\left(\frac{\partial^2}{(\partial z_t)^2}R(z_t)\right)^{-1}} \le 2\|\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\|^2_{\left(\frac{\partial^2}{(\partial \boldsymbol{w}_t(\boldsymbol{x}))^2}R(\boldsymbol{w}_t(\boldsymbol{x}))\right)^{-1}},$$

which, when combined with Lemma 16 and Equation (3.19) completes the proof. □

Lemma 7 (RESTATED). *For any $\gamma \in (0,1)$ and for all $t \in [T]$ we have that*

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t\right)_k^2 \left(\boldsymbol{w}_t(\boldsymbol{x}_0)\right)_k \,\middle|\, \mathcal{F}_t\right] \leq \frac{2Kd}{1-\gamma}\,.$$

Proof. By using the assumption that $\boldsymbol{x}_t^\top(\boldsymbol{\Theta}_t)_{\cdot,k} \leq 1$ we can see that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\mathbb{E}\left[\sum_{k=1}^{K}(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t)_k^2(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\mathbb{E}\left[\sum_{k=1}^{K}\left(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{x}_t(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k\right)^2(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \,\middle|\, \mathcal{F}_t\right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\mathbb{E}\left[\sum_{k=1}^{K}(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{x}_t)^2(\boldsymbol{a}_t)_k(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \,\middle|\, \mathcal{F}_t\right].
\end{aligned}$$

We continue by writing out the square,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\mathbb{E}\left[\sum_{k=1}^{K}(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{x}_t)^2(\boldsymbol{a}_t)_k(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\mathbb{E}\left[\sum_{k=1}^{K}\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{x}_t\boldsymbol{x}_t^\top(\boldsymbol{a}_t)_k\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{x}_0(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\sum_{k=1}^{K}\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{x}_0(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \mathcal{F}_t\right],
\end{aligned}$$

where in the last equality we used the definition of $\boldsymbol{\Sigma}_{t,k}$. Now, using the definition of π_t we can see that

$$\begin{aligned}
&(1-\gamma)\mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\sum_{k=1}^{K}\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{x}_0(\boldsymbol{w}_t(\boldsymbol{x}_0))_k \,\middle|\, \mathcal{F}_t\right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{x}_0, \boldsymbol{a}\sim\pi_t(\cdot|\boldsymbol{x}_0)}\left[\sum_{k=1}^{K}\boldsymbol{x}_0^\top \hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{x}_0(\boldsymbol{a})_k \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_0, \boldsymbol{a}\sim\pi_t(\cdot|\boldsymbol{x}_0)}\left[\sum_{k=1}^{K}\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{x}_0\boldsymbol{x}_0^\top(\boldsymbol{a})_k) \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\sum_{k=1}^{K}\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{t,k}^{+}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{\Sigma}_{t,k}) \,\middle|\, \mathcal{F}_t\right] \\
&= \mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\sum_{k=1}^{K}\operatorname{tr}(\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+\top}\boldsymbol{\Sigma}_{t,k}\hat{\boldsymbol{\Sigma}}_{t,k}^{+}) \,\middle|\, \mathcal{F}_t\right] \leq 2Kd,
\end{aligned}$$

where the last inequality follows from Lemma 4. Dividing by $(1-\gamma)$ completes the proof. □

Theorem 17. *Let* $\beta = \frac{1}{\sigma^2}$, *let* $\eta \leq \frac{\log(2)}{(M+1)}$, *and let* $\gamma \in (0, \frac{1}{2})$. *Algorithm 3 guarantees*

$$\mathcal{R} \leq \frac{m\left(1 + \log\left(\frac{K}{m}\right)\right)}{\eta} + 2\eta TKd + 2\gamma Tm + 2Tm\sigma R e^{-M\beta\gamma\frac{\lambda^{\boldsymbol{\Sigma}}_{\min}}{|\mathcal{E}|}}$$

Furthermore, if

$$M = \frac{\log(\gamma m)}{\beta\lambda^{\boldsymbol{\Sigma}}_{\min}}, \qquad \gamma = \min\left\{\frac{1}{2}, \sqrt{\frac{(1+\log(K/m))|\mathcal{E}|\log(T)}{T\beta\lambda^{\boldsymbol{\Sigma}}_{\min}}}\right\},$$

$$\text{and } \eta = \min\left\{\frac{\log(2)}{(M+1)}, \sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{2KdT}}\right\}$$

then Algorithm 3 guarantees

$$\begin{aligned}\mathcal{R} \leq \sqrt{2dmKT(1+\log(K/m))} + 6m\sqrt{\sigma^2(1+\log(K/m))\frac{T|\mathcal{E}|\log(T)}{\lambda^{\boldsymbol{\Sigma}}_{\min}}} \\ + 2m\sigma R + \frac{m(1+\log(K/m))}{\log(2)} + \frac{2|\mathcal{E}|\log(T)m\sigma^2(1+\log(K/m))}{\lambda^{\boldsymbol{\Sigma}}_{\min}}.\end{aligned}$$

Proof. Recall that

$$\tilde{\mathcal{R}}_T(\boldsymbol{x}) = \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{w}_t(\boldsymbol{x}) - \boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\pi^*_T(\boldsymbol{x})\right)\right]$$

and $\pi^*_T(\boldsymbol{x}) = \min_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{x}^\top\boldsymbol{\Theta}_t\boldsymbol{a}\right)\right]$. We by separating the usual regret from the exploration and applying Lemma 3:

$$\mathcal{R} \leq (1-\gamma)\mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] + 2\mathop{\mathbb{E}}_{\boldsymbol{x}_0,\mathcal{F}}\left[\sum_{t=1}^{T}\max_{\boldsymbol{a}\in\mathcal{A}}\left|\mathop{\mathbb{E}}_{\tilde{\boldsymbol{\Theta}}_t}\left[\boldsymbol{x}_0^\top(\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a} \;\middle|\; \mathcal{F}_t\right]\right|\right] + 2\gamma Tm$$

where we used that $\mathbb{E}_{\boldsymbol{a}_t}[\boldsymbol{a}_t|\mathcal{F}_t, \boldsymbol{x}_0] = (1-\gamma)\boldsymbol{w}_t(\boldsymbol{x}_0) + \frac{\gamma}{|\mathcal{E}|}\sum_{A\in\mathcal{E}}\boldsymbol{a}$ and that $\frac{\gamma}{|\mathcal{E}|}\sum_{A\in\mathcal{E}}\boldsymbol{x}_0^\top\boldsymbol{\Theta}_t\boldsymbol{a} \leq \gamma m$.

Note that since $\boldsymbol{\Sigma}_{t,k} \preceq \boldsymbol{\Sigma} \preceq \sigma^2\boldsymbol{I}$ setting $\beta = \frac{1}{\sigma^2} \leq \frac{1}{\lambda_{\max}(\boldsymbol{\Sigma}_{t,k})}$ is a valid choice to use in Lemma 5. By Lemma 5 we have that

$$2\mathop{\mathbb{E}}_{\boldsymbol{x}_0,\mathcal{F}}\left[\sum_{t=1}^{T}\max_{\boldsymbol{a}\in\mathcal{A}}\left|\mathop{\mathbb{E}}_{\tilde{\boldsymbol{\Theta}}_t}\left[\boldsymbol{x}_0^\top(\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a} \;\middle|\; \mathcal{F}_t\right]\right|\right] \leq 2Tm\sigma R\exp\left(-\frac{M\beta\gamma}{|\mathcal{E}|}\lambda^{\boldsymbol{\Sigma}}_{\min}\right).$$

Now, by Lemmas 6 and 7 we have that

$$(1-\gamma)\mathop{\mathbb{E}}_{\boldsymbol{x}_0}\left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] \leq \frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{\eta} + 2\eta TKd$$

Combining the above we find

$$\mathcal{R}_T \leq \frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{\eta} + 2\eta TKd + 2\gamma Tm + 2Tm\sigma R \exp\left(-\frac{M\beta\gamma}{|\mathcal{E}|}\lambda_{\min}^{\mathbf{\Sigma}}\right).$$

Now, setting $M = \frac{|\mathcal{E}|\log(T)}{\gamma\beta\lambda_{\min}^{\mathbf{\Sigma}}}$ we find

$$\mathcal{R}_T \leq \frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{\eta} + 2\eta TKd + 2\gamma Tm + 2m\sigma R.$$

Set $\gamma = \min\left\{1, \sqrt{\frac{(1+\log(K/m))|\mathcal{E}|\log(T)}{T\beta\lambda_{\min}^{\mathbf{\Sigma}}}}\right\}$ to find that $M = \max\left\{\frac{|\mathcal{E}|\log(T)}{\beta\lambda_{\min}^{\mathbf{\Sigma}}}, \sqrt{T\frac{|\mathcal{E}|\log(T)}{(1+\log(K/m))\beta\lambda_{\min}^{\mathbf{\Sigma}}}}\right\}$ and

$$\mathcal{R}_T \leq \frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{\eta} + 2\eta TKd + 2m\sqrt{(1+\log(K/m))\frac{T|\mathcal{E}|\log(T)}{\beta\lambda_{\min}^{\mathbf{\Sigma}}}} + 2m\sigma R.$$

Finally, setting $\eta = \min\left\{\frac{\log(2)}{(M+1)}, \sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{2KdT}}\right\}$ and replacing M by its value we find

$$\begin{aligned}
\mathcal{R}_T &\leq \sqrt{2dmKT(1+\log(K/m))} + 2m\sqrt{(1+\log(K/m))\frac{T|\mathcal{E}|\log(T)}{\beta\lambda_{\min}^{\mathbf{\Sigma}}}} \\
&\quad + 2m\sigma R + \frac{(M+1)m(1+\log(K/m))}{\log(2)} \\
&\leq \sqrt{2dmKT(1+\log(K/m))} + 6m\sqrt{(1+\log(K/m))\frac{T|\mathcal{E}|\log(T)}{\beta\lambda_{\min}^{\mathbf{\Sigma}}}} \\
&\quad + 2m\sigma R + \frac{m(1+\log(K/m))}{\log(2)} + \frac{2|\mathcal{E}|\log(T)m(1+\log(K/m))}{\beta\lambda_{\min}^{\mathbf{\Sigma}}},
\end{aligned}$$

after which replacing $\beta = \frac{1}{\sigma^2}$ completes the proof. □

3.C Full-Bandits - Tensors

In this section of the appendix we will rigorously introduce tensors. While tensors enjoy a wide employment in physics and other fields, to the best of our knowledge this is their first usage in bandit literature which justifies some background on tensors. Nevertheless, we only define some narrow concepts which may or may not align with how tensors have been used before.

Let $\mathbb{R}^d$ be a vectorspace made up of (column) vectors $\boldsymbol{x} \in \mathbb{R}^d$ equipped with the standard basis. Covectors are now (row) vectors, which are elements of the

dual space that are linear functions that map vectors to the reals, i.e., covectors are elements of the form $f : \mathbb{R}^d \to \mathbb{R}$. Thus we can combine a vector and a covector to a single real number or we could combine two covectors and get a function that takes two vectors as arguments and then returns a real number.

Tensors are now made up of vectors and covectors and we will write the degree of a vector as degree(a, b), where a is the number of vector and b the number of covector elements. A vector is of degree$(1, 0)$, a covector is degree$(0, 1)$, a matrix is of degree$(1, 1)$. We will primarily be interested in tensors of degree(2,2). While vectors and matrices are also tensors, from here on out we will usually only refer to degree(2,2) tensors as tensors, which we denote by $\boldsymbol{\Phi}, \boldsymbol{\Psi} \in \mathbb{R}^{m\times m\times n\times n}$. It is not required for general tensors that the first two and last two dimension agree but that will be the case for all tensors we will consider. Furthermore, we will also follow the notation introduced by Einstein, 1916 and index vector elements by a lower index like this $\boldsymbol{x}_i$ and covectors with an upper index like this $\boldsymbol{x}^i$. Since a matrix is made up of a vector and covector part, we will index it as follows $\boldsymbol{A}_i{}^j$. The Einstein notation also seeks to omit a lot of the sums and clutter associated with writing tensors usually. Instead of writing all sums explicitly, when an indexing variable appears once in a vector and once in a covector component then we are implicitly summing over the variable, if it only appears once then it is an index, let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$:

$$\boldsymbol{A}_i{}^j \boldsymbol{x}_j = \sum_j^m \boldsymbol{A}_i{}^j \boldsymbol{x}_j = (\boldsymbol{B}\boldsymbol{x})_i$$
$$\boldsymbol{A}_i{}^k \boldsymbol{B}_k{}^j = \sum_k^m \boldsymbol{A}_i{}^k \boldsymbol{B}_k{}^j = (\boldsymbol{A}\boldsymbol{B})_{i,j}.$$

Finally, we will extend the notation by curly brackets that collect all free parameters and compile them to a single object, which gives greater clarity on which parameters act as free indices and which are being summed over:

$$\boldsymbol{A}\boldsymbol{x} = \sum_j \boldsymbol{A}_i{}^j \boldsymbol{x}_j = \{\boldsymbol{A}_i{}^j \boldsymbol{x}_j\}_i$$
$$\boldsymbol{A}\boldsymbol{B} = \sum_k \boldsymbol{A}_i{}^k \boldsymbol{B}_k{}^j = \{\boldsymbol{A}_i{}^k \boldsymbol{B}_k{}^j\}_i{}^j$$
$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_i \sum_j \boldsymbol{x}^{\top\, i} \boldsymbol{A}_i{}^j \boldsymbol{x}_j = \{\boldsymbol{x}^{\top\, i} \boldsymbol{A}_i{}^j \boldsymbol{x}_j\}_{\mathbb{R}},$$

where $\{\cdot\}_{\mathbb{R}}$ is a real number, i.e., all indices are being summed over. To reiterate, $\{\boldsymbol{\Psi}_a{}^b{}_c{}^d\}_a{}^b{}_c{}^d$ is a tensor, $\boldsymbol{\Psi}_a{}^b{}_c{}^d \in \mathbb{R}$ is an element of $\boldsymbol{\Psi}$ at the position a, b, c, d, and $\{\boldsymbol{\Psi}_a{}^b{}_c{}^d\}_{\mathbb{R}} \in \mathbb{R}$ is obtained by summing over all indices of the tensor.

Definition 18. *Let $\boldsymbol{\Psi}, \boldsymbol{\Phi} \in \mathbb{R}^{m\times m\times n\times n}$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, then we define the*

following basic operations

$$\begin{aligned}
\boldsymbol{\Psi}(\boldsymbol{A}) &= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_a{}^d \\
\boldsymbol{\Psi} \cdot \boldsymbol{A} &= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_d{}^e\}_a{}^b{}_c{}^e \\
\boldsymbol{\Psi}(\boldsymbol{\Phi}) &= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Phi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d \\
\boldsymbol{A}(\boldsymbol{B}) &= \{\boldsymbol{A}_a{}^b \boldsymbol{B}_b{}^a\}_{\mathbb{R}} \\
\boldsymbol{A} \cdot \boldsymbol{\Psi} &= \{\boldsymbol{A}_e{}^a \boldsymbol{\Psi}_a{}^b{}_c{}^d\}_e{}^b{}_c{}^d.
\end{aligned}$$

Here it is important to pay close attention to the dimensions of each element. $\boldsymbol{\Psi}(\boldsymbol{A})$ is degree(1,1) tensor, i.e., a matrix while $\boldsymbol{\Psi} \cdot \boldsymbol{A}$ is a degree(2,2) tensor. All of the above operations are linear.

We denote by $\delta_{a,b}$ the Kronecker delta defined as $\delta_{a,b} = \mathbb{1}[a = b]$ with the indicator function $\mathbb{1}$ which is 1 if the condition is true and 0 otherwise. The neutral element in our tensor space is given in the following definition.

Definition 19. *We will call the neutral element of the tensor space $\boldsymbol{\mathcal{I}}$ and define it as follows $\boldsymbol{\mathcal{I}} = \{\delta_{a,b}\delta_{c,d}\}_a{}^b{}_c{}^d = \{\mathbb{1}[a = b \wedge c = d]\}_a{}^b{}_c{}^d$.*

Thus, $\boldsymbol{\mathcal{I}}_a{}^b{}_c{}^d$ is one if $a = b$ and $c = d$ and zero otherwise, in a sense the elements of any tensor where $\boldsymbol{\Psi}_a{}^a{}_b{}^b$ for some a, b can be seen as the diagonal of the tensor and we will define the trace in terms of these elements later. $\boldsymbol{\mathcal{I}}$ is then the tensor with ones on the diagonal and zeros everywhere else.

Observe that $\boldsymbol{\mathcal{I}}$ in fact is the identity for all operations defined above:

$$\begin{aligned}
\boldsymbol{\mathcal{I}}(\boldsymbol{A}) &= \{\boldsymbol{\mathcal{I}}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_a{}^d = \boldsymbol{A} \\
\boldsymbol{\mathcal{I}}(\boldsymbol{\Psi}) &= \{\boldsymbol{\mathcal{I}}_a{}^b{}_c{}^d \boldsymbol{\Psi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d = \boldsymbol{\Psi}.
\end{aligned}$$

We will define the inverse of a tensor in terms of this neutral element $\boldsymbol{\mathcal{I}}$ as follows.

Definition 20. *Let $\boldsymbol{\Phi}$ be a tensor, we will call $\boldsymbol{\Psi}$ an inverse of $\boldsymbol{\Phi}$ if*

$$\boldsymbol{\Phi}(\boldsymbol{\Psi}) = \boldsymbol{\mathcal{I}} \quad \Longleftrightarrow \quad \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Phi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d = \{\delta_{a,e}\delta_{f,d}\}_a{}^e{}_f{}^d.$$

If such a $\boldsymbol{\Psi}$ exists, it we denoted it by $\boldsymbol{\Phi}^{-1}$ and we call $\boldsymbol{\Phi}$ invertible.

Lemma 21. *Let $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$ be tensors of equal dimension and let $\boldsymbol{A}$ be a matrix of appropriate dimension. Tensors are associative:*

$$\boldsymbol{\Phi}(\boldsymbol{\Psi}(\boldsymbol{A})) = (\boldsymbol{\Phi}(\boldsymbol{\Psi}))(\boldsymbol{A}).$$

Proof. The proof follows from repeatedly applying the definition of $\boldsymbol{\Psi}(\boldsymbol{A})$ and $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ in 18.

$$\begin{aligned}\boldsymbol{\Phi}(\boldsymbol{\Psi}(\boldsymbol{A})) &= \boldsymbol{\Phi}(\{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_a{}^d) \\ &= \{\boldsymbol{\Phi}_e{}^a{}_d{}^f \boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_e{}^f \\ &= \{(\boldsymbol{\Phi}(\boldsymbol{\Psi}))_e{}^b{}_c{}^f \boldsymbol{A}_b{}^c\}_e{}^f \\ &= (\boldsymbol{\Phi}(\boldsymbol{\Psi}))(\boldsymbol{A}).\end{aligned}$$

□

We now introduce our definition of the Frobenius inner product.

Definition 22. *Let $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$ be tensors. The Frobenius inner product between $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ is defined as*

$$\langle \boldsymbol{\Psi}, \boldsymbol{\Phi} \rangle_F = \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Phi}_b{}^a{}_d{}^c\}_{\mathbb{R}}.$$

For matrices $\boldsymbol{A}, \boldsymbol{B}$ the Frobenius inner product is defined as

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \{\boldsymbol{A}_a{}^b \boldsymbol{B}_b{}^a\}_{\mathbb{R}} = \boldsymbol{A}(\boldsymbol{B}) = \boldsymbol{B}(\boldsymbol{A}).$$

The following Lemma characterises how the inner product acts on inverses.

Lemma 23. *Let $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m\times n\times n}$, be invertible, then*

$$\langle \boldsymbol{\Psi}, \boldsymbol{\Psi}^{-1} \rangle_F = mn.$$

Proof. For this proof, it is clearer to write the sums explicitly, first we apply the definition of $\langle \cdot, \cdot \rangle_F$ (Definition 22), then reordering the sums, then applying the definition of $\boldsymbol{\Psi}(\boldsymbol{\Phi})$ (Definition 18), while the last steps only consist of applying the definition of $\boldsymbol{\Psi}^{-1}$ (Definition 20) and $\boldsymbol{\mathcal{I}}$ (Definition 19).

$$\begin{aligned}\langle \boldsymbol{\Psi}, \boldsymbol{\Psi}^{-1} \rangle_F &= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Psi}^{-1}{}_b{}^a{}_d{}^c\}_{\mathbb{R}} \\ &= \sum_a^m \sum_b^m \sum_c^n \sum_d^n \boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Psi}^{-1}{}_b{}^a{}_d{}^c \\ &= \sum_a^m \sum_d^n \sum_c^n \sum_b^m \boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Psi}^{-1}{}_b{}^a{}_d{}^c \\ &= \sum_a^m \sum_d^n \boldsymbol{\Psi}(\boldsymbol{\Psi}^{-1})_a{}^a{}_d{}^d \\ &= \sum_a^m \sum_d^n \boldsymbol{\mathcal{I}}_a{}^a{}_d{}^d \\ &= mn.\end{aligned}$$

□

We will also need to define the transpose of the tensor.

Definition 24. *Let $\boldsymbol{\Psi}$ be a tensor, then the transpose of $\boldsymbol{\Psi}$ is defined as*

$$\boldsymbol{\Psi}^{\top}{}_{a}{}^{b}{}_{c}{}^{d} = \boldsymbol{\Psi}_{b}{}^{a}{}_{d}{}^{c}.$$

A tensor is called symmetric iff it is invariant under transpose, i.e., $\boldsymbol{\Psi}^{\top} = \boldsymbol{\Psi}$.

Lemma 25. *Let $\boldsymbol{\Phi}$ be an invertible tensor, then transposing and inverting can be interchanged, i.e.,*

$$(\boldsymbol{\Psi}^{-1})^{\top} = (\boldsymbol{\Psi}^{\top})^{-1}.$$

Proof. We will show the claim be showing that $(\boldsymbol{\Psi}^{-1})^{\top}$ is the inverse of $\boldsymbol{\Psi}^{\top}$ by first applying the definition of $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ (Definition 18), applying the definition of the transpose (Definition 24) twice, reordering and applying the transpose to the entire expression and finally applying $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ again:

$$\begin{aligned}
(\boldsymbol{\Psi}^{-1})^{\top}(\boldsymbol{\Psi}^{\top}) &= \{(\boldsymbol{\Psi}^{-1})^{\top}{}_{a}{}^{b}{}_{c}{}^{d}\boldsymbol{\Psi}^{\top}{}_{b}{}^{e}{}_{f}{}^{c}\}_{a}{}^{e}{}_{f}{}^{d} \\
&= \{\boldsymbol{\Psi}^{-1}{}_{b}{}^{a}{}_{d}{}^{c}\boldsymbol{\Psi}_{e}{}^{b}{}_{c}{}^{f}\}_{a}{}^{e}{}_{f}{}^{d} \\
&= \{\boldsymbol{\Psi}_{e}{}^{b}{}_{c}{}^{f}\boldsymbol{\Psi}^{-1}{}_{b}{}^{a}{}_{d}{}^{c}\}_{a}{}^{e}{}_{f}{}^{d} \\
&= \left(\{\boldsymbol{\Psi}_{e}{}^{b}{}_{c}{}^{f}\boldsymbol{\Psi}^{-1}{}_{b}{}^{a}{}_{d}{}^{c}\}_{e}{}^{a}{}_{d}{}^{f}\right)^{\top} \\
&= (\boldsymbol{\Psi}(\boldsymbol{\Psi}^{-1}))^{\top}.
\end{aligned}$$

□

We write the tensor product as $\otimes$. It will take two arbitrary tensors as input and output another tensor, we will only concretely define the following for two vectors $\boldsymbol{x}, \boldsymbol{y}$ and two matrices $\boldsymbol{A}, \boldsymbol{B}$:

Definition 26. *Let $\boldsymbol{x}, \boldsymbol{y}$ be vectors and $\boldsymbol{A}, \boldsymbol{B}$ be matrices, then we define the tensor product $\otimes$ as follows*

$$\begin{aligned}
(\boldsymbol{x} \otimes y) &= \boldsymbol{x}_{a}\boldsymbol{y}^{\top\, b} = \boldsymbol{x}\boldsymbol{y}^{\top} \\
(\boldsymbol{A} \otimes \boldsymbol{B}) &= \boldsymbol{A}_{a}{}^{b}\boldsymbol{B}^{\top}{}_{c}{}^{d}.
\end{aligned}$$

It is important to notice that $\boldsymbol{x} \otimes \boldsymbol{y}$ is a matrix of degree$(1,1)$ and $(\boldsymbol{A} \otimes \boldsymbol{B})$ is a tensor of degree$(2,2)$. The following Lemma details the relationship between the tensor product and tensor matrix operations.

Lemma 8 (RESTATED)**.** $\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}^{\top} = (\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{B})$*, where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are matrices of appropriate size.*

Proof. We start from the right-hand side and apply the definition of the tensor product (Definition 26), followed by the definition of $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ (Definition 18) and finish by rearranging.

$$\begin{aligned}(\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{B}) &= (\{\boldsymbol{A}_a{}^b \boldsymbol{C}^\top{}_c{}^d\}_a{}^b{}_c{}^d)(\boldsymbol{B}) \\ &= \{\boldsymbol{A}_a{}^b \boldsymbol{C}^\top{}_c{}^d \boldsymbol{B}_b{}^c\}_a{}^d \\ &= \{\boldsymbol{A}_a{}^b \boldsymbol{B}_b{}^c \boldsymbol{C}^\top{}_c{}^d\}_a{}^d \\ &= \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}^\top.\end{aligned}$$

□

Lemma 27. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be symmetric matrices such that $\boldsymbol{A} = \boldsymbol{A}^\top$ and $\boldsymbol{B} = \boldsymbol{B}^\top$, now $(\boldsymbol{A} \otimes \boldsymbol{B})$ is a symmetric tensor, i.e.,*

$$(\boldsymbol{A} \otimes \boldsymbol{B}) = (\boldsymbol{A} \otimes \boldsymbol{B})^\top$$

Proof. Fix some a, b, c, d. First we use the definition of the tensor product (Definition 26), and then we use the fact that both $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric. Then we use the definition of transposing for matrices and finally recognize the tensor product again

$$\begin{aligned}(\boldsymbol{A} \otimes \boldsymbol{B})_a{}^b{}_c{}^d &= \boldsymbol{A}_a{}^b \boldsymbol{B}_c{}^d \\ &= \boldsymbol{A}^\top{}_a{}^b \boldsymbol{B}^\top{}_c{}^d \\ &= \boldsymbol{A}_b{}^a \boldsymbol{B}_d{}^c \\ &= (\boldsymbol{A} \otimes \boldsymbol{B})^\top{}_a{}^b{}_c{}^d.\end{aligned}$$

□

Next we introduce summing over tensors.

Definition 28. *Let $\boldsymbol{\Psi}, \boldsymbol{\Phi}$ be tensors, then $\boldsymbol{\Psi} + \boldsymbol{\Phi} = \{\boldsymbol{\Psi}_a{}^b{}_c{}^d + \boldsymbol{\Phi}_a{}^b{}_c{}^d\}_a{}^b{}_c{}^d$.*

Lemma 29. *Let $\boldsymbol{B}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_N$ be matrices, then*

$$\sum_{n=1}^{N} (\boldsymbol{B} \otimes \boldsymbol{A}_n) = (\boldsymbol{B} \otimes \sum_{n=1}^{N} \boldsymbol{A}_n).$$

Proof. We start by applying the definition of the tensor outer product (Definition 26), applying the definition of summing tensors (Definition 28) N times, using

the associative property of $\boldsymbol{B}_a{}^b$ for any given a, b and finish by applying the definition of the tensor product (Definition 26) again.

$$\begin{aligned}
\sum_{n=1}^{N}(\boldsymbol{B}\otimes\boldsymbol{A}_n) &= \sum_{n=1}^{N}\{\boldsymbol{B}_a{}^b\boldsymbol{A}_n^\top{}_c{}^d\}_a{}^b{}_c{}^d \\
&= \{\sum_{n=1}^{N}\boldsymbol{B}_a{}^b\boldsymbol{A}_n^\top{}_c{}^d\}_a{}^b{}_c{}^d \\
&= \{\boldsymbol{B}_a{}^b\left(\sum_{n=1}^{N}\boldsymbol{A}_n^\top\right)_c{}^d\}_a{}^b{}_c{}^d \\
&= (\boldsymbol{B}\otimes\sum_{n=1}^{N}\boldsymbol{A}_n)
\end{aligned}$$

□

Lemma 30. *Let $\boldsymbol{\Psi}$ be a tensor and let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{w}$ be vectors, then*

$$\boldsymbol{x}^\top\boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{w}^\top)\boldsymbol{v} = \langle\boldsymbol{\Psi}, (\boldsymbol{y}\boldsymbol{x}^\top\otimes\boldsymbol{v}\boldsymbol{w}^\top)\rangle_F$$

and

$$\langle\boldsymbol{w}\boldsymbol{x}^\top, \boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{v}^\top)\rangle_F = \langle\boldsymbol{\Psi}, (\boldsymbol{y}\boldsymbol{x}^\top\otimes\boldsymbol{v}\boldsymbol{w}^\top)\rangle_F$$

Proof. For the first statement we write $\boldsymbol{x}^\top\boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{w}^\top)\boldsymbol{v}$ in tensor notation and then apply the Frobenius inner product (Definition 22) alongside the definition of the tensor product (Definition 26)

$$\begin{aligned}
\boldsymbol{x}^\top\boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{w}^\top)\boldsymbol{v} &= \{\boldsymbol{x}^\top{}^a\boldsymbol{\Psi}_a{}^b{}_c{}^d\boldsymbol{y}_b\boldsymbol{w}^\top{}^c\boldsymbol{v}_d\}_\mathbb{R} \\
&= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d\boldsymbol{x}^\top{}^a\boldsymbol{y}_b\boldsymbol{w}^\top{}^c\boldsymbol{v}_d\}_\mathbb{R} \\
&= \langle\boldsymbol{\Psi}, \{\boldsymbol{x}^\top{}^a\boldsymbol{y}_b\boldsymbol{w}^\top{}^c\boldsymbol{v}_d\}_b{}^a{}_d{}^c\rangle_F \\
&= \langle\boldsymbol{\Psi}, \{(\boldsymbol{y}\boldsymbol{x}^\top)_b{}^a(\boldsymbol{v}\boldsymbol{w}^\top)_d{}^c\}_b{}^a{}_d{}^c\rangle_F \\
&= \langle\boldsymbol{\Psi}, (\boldsymbol{y}\boldsymbol{x}^\top\otimes\boldsymbol{v}\boldsymbol{w}^\top)\rangle_F.
\end{aligned}$$

For the second statement we apply the common definition of the Frobenius inner product for matrices, followed by applying the first half of the lemma

$$\begin{aligned}
\langle\boldsymbol{w}\boldsymbol{x}^\top, \boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{v}^\top)\rangle_F &= \{\boldsymbol{w}_b\boldsymbol{x}^\top{}^a\boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{v}^\top)_a{}^b\}_\mathbb{R} \\
&= \boldsymbol{x}^\top\boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{v}^\top)\boldsymbol{w} \\
&= \langle\boldsymbol{\Psi}, (\boldsymbol{y}\boldsymbol{x}^\top\otimes\boldsymbol{v}\boldsymbol{w}^\top)\rangle_F.
\end{aligned}$$

□

Next we relate previous definitions and operations to standard linear algebra. For that we show a certain equivalency between matrices and tensors as well as matrices and vectors by introducing the two following definitions. The first of which is the flattening operation that can act on a tensor or matrix.

Definition 31. *Let $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m\times n\times n}$ be a tensor and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix, then*

$$\begin{aligned}\boldsymbol{\Psi}^F{}_a{}^b &= \boldsymbol{\Psi}_{(a \mod m)+1}{}^{(b \mod m)+1}{}_{\lfloor b/m\rfloor+1}{}^{\lfloor a/m\rfloor+1}\\ \boldsymbol{A}^F{}_a &= \boldsymbol{A}_{(a \mod m)+1}{}^{\lfloor a/m\rfloor+1}\end{aligned}$$

Furthermore, $\boldsymbol{\Psi}^F \in \mathbb{R}^{mn\times mn}$ and $\boldsymbol{A} \in \mathbb{R}^{mn}$.

Note that in the definition above there is a +1 is the indexes. These are necessary since our indices start counting from 1 and not from 0.

Similarly, we define how to unflatten a matrix or a vector.

Definition 32. *Let $\boldsymbol{A} \in \mathbb{R}^{mn\times mn}$ be a matrix and $\boldsymbol{x} \in \mathbb{R}^{mn}$ be a vector, then we will define*

$$\begin{aligned}\boldsymbol{A}^U{}_a{}^b{}_c{}^d &= \boldsymbol{A}_{(a-1)+(d-1)m}{}^{(b-1)+(c-1)m}\\ \boldsymbol{x}^U{}_a{}^b &= \boldsymbol{x}_{(a-1)+(b-1)m}\end{aligned}$$

Furthermore $\boldsymbol{A}^U \in \mathbb{R}^{m\times m\times n\times n}$ and $\boldsymbol{x}^U \in \mathbb{R}^{m\times n}$.

In light of this equivalency between tensors and matrices, one might wonder if we do really require this extensive notion of tensors in the first place. This incredulity might even be reinforced upon recognizing that the flattening of the tensor product of these two matrices $(\boldsymbol{A}\otimes\boldsymbol{B})^F$ yields the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Next, we show that unflattening a flattened tensor recovers the original tensor.

Lemma 33. *Let $\boldsymbol{\Psi}$ be a tensor, then unflattening is the inverse of flattening*

$$(\boldsymbol{\Psi}^F)^U = \boldsymbol{\Psi}.$$

Proof. Fix some a, b, c, d, then we first apply the definition of unflattening (Definition 32) followed by the definition of flattening (Definition 31)

$$\begin{aligned}(\boldsymbol{\Psi}^F)^U{}_a{}^b{}_c{}^d &= \boldsymbol{\Psi}^F{}_{(a-1)+(d-1)m}{}^{(b-1)+(c-1)m}\\ &= \boldsymbol{\Psi}_{((a-1)+(d-1)m \mod m)+1}{}^{((b-1)+(c-1)m \mod m)+1}{}_{(\lfloor(b-1)+(c-1)m/m\rfloor)+1}{}^{(\lfloor(a-1)+(d-1)m/m\rfloor)+1}\\ &= \boldsymbol{\Psi}_a{}^b{}_c{}^d.\end{aligned}$$

□

Lemma 34. *Let $\boldsymbol{\Psi}$ be a tensor, then $\boldsymbol{\Psi}^F$ is symmetric if and only if $\boldsymbol{\Psi}$ is symmetric.*

Proof. Let $\mathbf{\Psi}$ be symmetric, we will now show that $\mathbf{\Psi}^F$ is symmetric. For that we use the definition of flattening (Definition 31) alongside the symmetry of $\mathbf{\Psi}$.

$$\begin{aligned}\mathbf{\Psi}^F{}_a{}^b &= \mathbf{\Psi}_{(a \mod m)+1}{}^{(b \mod m)+1}{}_{\lfloor b/m \rfloor+1}{}^{\lfloor a/m \rfloor+1} \\ &= \mathbf{\Psi}_{(b \mod m)+1}{}^{(a \mod m)+1}{}_{\lfloor a/m \rfloor+1}{}^{\lfloor b/m \rfloor+1} = \mathbf{\Psi}^F{}_b{}^a\end{aligned}$$

For the other direction let $\mathbf{\Psi}^F$ be symmetric, now we show that $\mathbf{\Psi}$ is too. First we use the definition of unflatteing (Definition 32) alongside the symmetry of $\mathbf{\Psi}^F$ to show

$$\mathbf{\Psi}_a{}^b{}_c{}^d = \mathbf{\Psi}^F{}_{(a-1)+(d-1)m}{}^{(b-1)+(c-1)m} = \mathbf{\Psi}^F{}_{(b-1)+(c-1)m}{}^{(a-1)+(d-1)m} = \mathbf{\Psi}_b{}^a{}_d{}^c.$$

□

Lemma 35. *Let $\mathbf{\Psi}, \mathbf{\Phi} \in \mathbb{R}^{m\times m\times n\times n}$ be tensors and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ a matrix, then $\mathbf{\Psi}$ acting on $\boldsymbol{A}$ or $\mathbf{\Phi}$ is equivalent in the higher or lower dimensional space*

$$\begin{aligned}\mathbf{\Psi}(\mathbf{\Phi}) &= (\mathbf{\Psi}^F \mathbf{\Phi}^F)^U \\ \mathbf{\Psi}(\boldsymbol{A}) &= (\mathbf{\Psi}^F \boldsymbol{A}^F)^U\end{aligned}$$

where $\mathbf{\Psi}^F\mathbf{\Phi}^F$ is employing the usual matrix multiplication.

Proof. We start with the first claim by using the definition $\mathbf{\Psi}(\boldsymbol{A})$ and $\mathbf{\Phi}(\mathbf{\Psi})$ (Definition 18), writing the sums explicitly, reindexing, applying the definition of flattening (Definition 31) and recognizing a matrix product before finally using the definition of unflattening (Definition 32)

$$\begin{aligned}\mathbf{\Psi}(\mathbf{\Phi}) &= \{\mathbf{\Psi}_a{}^b{}_c{}^d \mathbf{\Phi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d \\ &= \{\sum_{b=1}^{m}\sum_{c=1}^{n} \mathbf{\Psi}_a{}^b{}_c{}^d \mathbf{\Phi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d \\ &= \{\sum_{i=1}^{mn} \mathbf{\Psi}_a{}^{(i \mod m)+1}{}_{\lfloor i/m \rfloor+1}{}^d \mathbf{\Phi}_{(i \mod m)+1}{}^e{}_f{}^{\lfloor i/m \rfloor+1}\}_a{}^e{}_f{}^d \\ &= \{\sum_{i=1}^{mn} \mathbf{\Psi}^F{}_{(a-1)+(d-1)m}{}^i \mathbf{\Phi}^F{}_i{}^{(e-1)+(f-1)m}\}_a{}^e{}_f{}^d \\ &= \{(\mathbf{\Psi}^F\mathbf{\Phi}^F)_{(a-1)+(d-1)m}{}^{(e-1)+(f-1)m}\}_a{}^e{}_f{}^d \\ &= (\mathbf{\Psi}^F\mathbf{\Phi}^F)^U.\end{aligned}$$

The second part of the proof follows the exact same steps except recognizing a

matrix vector product instead of a matrix product in the second to last step

$$\begin{aligned}
\boldsymbol{\Psi}(\boldsymbol{A}) &= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_a{}^d \\
&= \{\sum_{b=1}^{m}\sum_{c=1}^{n} \boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{A}_b{}^c\}_a{}^d \\
&= \{\sum_{i=1}^{mn} \boldsymbol{\Psi}_a{}^{(i \mod m)+1}{}_{\lfloor i/m \rfloor+1}{}^d \boldsymbol{A}_{(i \mod m)+1}{}^{\lfloor i/m \rfloor+1}\}_a{}^d \\
&= \{\sum_{i=1}^{mn} \boldsymbol{\Psi}^F{}_{(a-1)+(d-1)m)}{}^i \boldsymbol{A}^F{}_i\}_a{}^d \\
&= \{(\boldsymbol{\Psi}^F \boldsymbol{A}^F)_{(a-1)+(d-1)m)}\}_a{}^d \\
&= (\boldsymbol{\Psi}^F \boldsymbol{A}^F)^U.
\end{aligned}$$

□

Lemma 36. *Let $\boldsymbol{\Phi}$ be an invertible tensor, then flattening and inverting can be interchanged, i.e.*

$$(\boldsymbol{\Psi}^{-1})^F = (\boldsymbol{\Psi}^F)^{-1}$$

Proof. We will show that $(\boldsymbol{\Psi}^{-1})^F$ is the inverse of $\boldsymbol{\Psi}^F$ by first using the fact that flattening and unflattening cancel another (Lemma 33), followed by applying $\boldsymbol{\Psi}(\boldsymbol{\Phi}) = (\boldsymbol{\Psi}^F \boldsymbol{\Phi}^F)^U$ (Lemma 35) and the definition of $\boldsymbol{\Psi}^{-1}$ (Definition 20)

$$\begin{aligned}
\boldsymbol{\Psi}^F(\boldsymbol{\Psi}^{-1})^F &= ((\boldsymbol{\Psi}^F(\boldsymbol{\Psi}^{-1})^F)^U)^F \\
&= (\boldsymbol{\Psi}(\boldsymbol{\Psi}^{-1}))^F \\
&= (\boldsymbol{\mathcal{I}})^F \\
&= \boldsymbol{I} \\
\Rightarrow (\boldsymbol{\Psi}^{-1})^F &= (\boldsymbol{\Psi}^F)^{-1}
\end{aligned}$$

□

Lemma 37. *Let $\boldsymbol{\Psi}$ be a tensor, then transposing and flattening can be interchanged, i. e.*

$$\boldsymbol{\Psi}^{\top F} = \boldsymbol{\Psi}^{F\top}$$

Proof. Fix some a, b. First we apply the definition of flattening (Definition 31) and then the definition of transposing (Definition 24). We then proceed using the definition of flattening followed by the definition of transposing one more time

$$\begin{aligned}
\boldsymbol{\Psi}^{\top F}{}_a{}^b &= \boldsymbol{\Psi}^\top{}_{(a \mod m)+1}{}^{(b \mod m)+1}{}_{\lfloor b/m \rfloor+1}{}^{\lfloor a/m \rfloor+1} \\
&= \boldsymbol{\Psi}{}_{(b \mod m)+1}{}^{(a \mod m)+1}{}_{\lfloor a/m \rfloor+1}{}^{\lfloor b/m \rfloor+1} \\
&= \boldsymbol{\Psi}{}_{(b \mod m)+1}{}^{(a \mod m)+1}{}_{\lfloor a/m \rfloor+1}{}^{\lfloor b/m \rfloor+1} \\
&= \boldsymbol{\Psi}^F{}_b{}^a \\
&= \boldsymbol{\Psi}^{F\top}{}_a{}^b .
\end{aligned}$$

□

Next we will show two quick facts about how flattening and the tensor product interact

Lemma 38. *Let $\boldsymbol{\Psi}$ be a tensor and let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{w}$ be vectors, then*

$$(\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top) = \left((\boldsymbol{x}\boldsymbol{w}^\top)^F \otimes (\boldsymbol{v}\boldsymbol{y}^\top)^F\right)^U .$$

Proof. We start by applying the definition of the tensor product (Definition 26), followed by the fact that flattening and unflattening cancel another (Lemma 33), then the definition of flattening (Definition 31), some rearranging before applying the same flattening definition for matrices twice, lastly we apply the tensor product one more time

$$\begin{aligned}
(\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top) &= \{\boldsymbol{x}_{\ a}\boldsymbol{y}^{\top\ b}\boldsymbol{v}_{\ c}\boldsymbol{w}^{\top\ d}\}_a^{\ b}{}_c^{\ d} \\
&= ((\{x_{\ a}\boldsymbol{y}^{\top\ b}\boldsymbol{v}_{\ c}\boldsymbol{w}^{\top\ d}\}_a^{\ b}{}_c^{\ d})^F)^U \\
&= \left(\{\boldsymbol{x}_{(a \mod m)+1}\boldsymbol{y}^{\top\ (b \mod m)+1}\boldsymbol{v}_{\lfloor b/m\rfloor+1}\boldsymbol{w}^{\top\ \lfloor a/m\rfloor+1}\}_a^{\ b}\right)^U \\
&= \left(\{\boldsymbol{x}_{(a \mod m)+1}\boldsymbol{w}^{\top\ \lfloor a/m\rfloor+1}\boldsymbol{v}_{\lfloor b/m\rfloor+1}\boldsymbol{y}^{\top\ (b \mod m)+1}\}_a^{\ b}\right)^U \\
&= \left(\{(\boldsymbol{x}\boldsymbol{w}^\top)^F{}_a((\boldsymbol{v}\boldsymbol{y}^\top)^F)^{\top\ b}\}_a^{\ b}\right)^U \\
&= \left((\boldsymbol{x}\boldsymbol{w}^\top)^F \otimes (\boldsymbol{v}\boldsymbol{y}^\top)^F\right)^U .
\end{aligned}$$

□

Lemma 39. *Let $\boldsymbol{\Psi}$ be a tensor and let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{w}$ be vectors, then one can also perform the Frobenius inner product between $\boldsymbol{\Psi}$ and $(\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top)$ in a lower dimension*

$$\langle \boldsymbol{\Psi}, (\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top)\rangle_F = ((\boldsymbol{y}\boldsymbol{v}^\top)^F)^\top \boldsymbol{\Psi}^F (\boldsymbol{x}\boldsymbol{w}^\top)^F$$

Proof. First we apply the definition of the Frobenius inner product (Definition 22), then the definition of the tensor product, (Definition 26), some rearranging, then we change the indexing by using the definition of flattening (Definition 31) on $\boldsymbol{\Psi}$ and then twice again on the vectors, finally we recognize the product between two tensors and a matrix

$$\begin{aligned}
\langle \boldsymbol{\Psi}, (\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top)\rangle_F &= \{\boldsymbol{\Psi}_a^{\ b}{}_c^{\ d}(\boldsymbol{x}\boldsymbol{y}^\top \otimes \boldsymbol{w}\boldsymbol{v}^\top)_b^{\ a}{}_d^{\ c}\}_{\mathbb{R}} \\
&= \{\boldsymbol{\Psi}_a^{\ b}{}_c^{\ d}\boldsymbol{x}_{\ b}\boldsymbol{y}^{\top\ a}\boldsymbol{v}_{\ d}\boldsymbol{w}^{\top\ c}\}_{\mathbb{R}} \\
&= \{\boldsymbol{\Psi}_a^{\ b}{}_c^{\ d}\boldsymbol{y}^{\top\ a}\boldsymbol{x}_{\ b}\boldsymbol{w}^{\top\ c}\boldsymbol{v}_{\ d}\}_{\mathbb{R}} \\
&= \{\boldsymbol{\Psi}^F{}_a^{\ b}\boldsymbol{y}^{\top\ (a \mod m)+1}\boldsymbol{x}_{\ (b \mod m)+1}\boldsymbol{w}^{\top\ \lfloor b/m\rfloor+1}\boldsymbol{v}_{\ \lfloor a/m\rfloor+1}\}_{\mathbb{R}} \\
&= \{\boldsymbol{\Psi}^F{}_a^{\ b}((\boldsymbol{y}\boldsymbol{v}^\top)^F)^{\top\ a}(\boldsymbol{x}\boldsymbol{w}^\top)^F{}_b\}_{\mathbb{R}} \\
&= ((\boldsymbol{y}\boldsymbol{v}^\top)^F)^\top \boldsymbol{\Psi}^F (\boldsymbol{x}\boldsymbol{w}^\top)^F .
\end{aligned}$$

□

Lemma 40. *Let $\boldsymbol{A}$ be a matrix and $\boldsymbol{x}, \boldsymbol{y}$ be vectors, then*

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{x}\boldsymbol{y}^\top)^{F\top} \boldsymbol{A}^F$$

Proof. First we write $\boldsymbol{x}^\top \boldsymbol{A}$ in the tensor notation and regroup $\boldsymbol{y}$ and $\boldsymbol{x}^\top$ to a single matrix. Then we re-index by writing the sums explicitly. Now we can transpose $\boldsymbol{y}\boldsymbol{x}^\top$ and use the flatten operator (Definition 31). Finally we write in Einstein notation again and conclude by recognizing the quantity on the left-hand side.

$$\begin{aligned}
\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} &= \{\boldsymbol{x}^{\top\, a} \boldsymbol{A}_a{}^b \boldsymbol{y}_b\}_{\mathbb{R}} \\
&= \{(\boldsymbol{y}\boldsymbol{x}^\top)_b{}^a \boldsymbol{A}_a{}^b\}_{\mathbb{R}} \\
&= \sum_i (\boldsymbol{y}\boldsymbol{x}^\top)_{\lfloor i/m \rfloor + 1}{}^{(i \mod m)+1} \boldsymbol{A}_{(i \mod m)+1}{}^{\lfloor i/m \rfloor + 1} \\
&= \sum_i (\boldsymbol{x}\boldsymbol{y}^\top)_{(i \mod m)+1}{}^{\lfloor i/m \rfloor + 1} \boldsymbol{A}_{(i \mod m)+1}{}^{\lfloor i/m \rfloor + 1} \\
&= \sum_i (\boldsymbol{x}\boldsymbol{y}^\top)^F{}_i \boldsymbol{A}^F{}_i \\
&= \{(\boldsymbol{x}\boldsymbol{y}^\top)^{F\top\, i} \boldsymbol{A}^F{}_i\}_{\mathbb{R}} \\
&= (\boldsymbol{x}\boldsymbol{y}^\top)^{F\top} \boldsymbol{A}^F.
\end{aligned}$$

□

Next we will define eigenmatrices and eigenvalues for our definition of tensors.

Definition 41. *We will call a scalar $\lambda \in \mathbb{R}$ an eigenvalue of a tensor $\boldsymbol{\Psi}$ if there exists a matrix $\boldsymbol{A}$ such that*

$$\boldsymbol{\Psi}(\boldsymbol{A}) = \lambda \boldsymbol{A}.$$

Such a matrix $\boldsymbol{A}$ is then called an eigenmatrix of $\boldsymbol{\Psi}$.

In the next Lemma we will show that flattened tensors and tensors have the same eigenvalues, i.e. $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}^F$ have the same eigenvalues.

Lemma 42.
Let $\lambda'_1, \ldots, \lambda'_{k'}$ be the eigenvalues of $\boldsymbol{\Psi}$ with eigenmatrices $\boldsymbol{A}'_1, \ldots, \boldsymbol{A}'_{k'}$ and let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of $\boldsymbol{\Psi}^F$ with eigenvectors $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$, then $k = k'$ and there exists a permutation $\sigma \in S_k$ such that

$$\begin{aligned}
\lambda_1, \ldots, \lambda_k &= \lambda'_{\sigma(1)}, \ldots, \lambda'_{\sigma(k)} \\
\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k &= \boldsymbol{A}'^F_{\sigma(1)}, \ldots, \boldsymbol{A}'^F_{\sigma(k)}.
\end{aligned}$$

Proof. Let all variables be defined as in the lemma. We now show that all eigenmatrices of $\boldsymbol{\Psi}$ are eigenvectors of $\boldsymbol{\Psi}^F$, thus let $\boldsymbol{A}'$ be an eigenmatrix of $\boldsymbol{\Psi}$ with eigenvalue λ'

$$\boldsymbol{\Psi}^F(\boldsymbol{A}'^F) = (\boldsymbol{\Psi}(\boldsymbol{A}'))^F = (\lambda'\boldsymbol{A}')^F = \lambda'(\boldsymbol{A}')^F,$$

where we used the fact that an operation can be performed in lower or higher dimensional space (Lemma 35). Next we will use the same lemma again to conclude that all eigenvectors of $\boldsymbol{\Psi}^F$ are eigenmatrices of $\boldsymbol{\Psi}$ if unflattened and thus let $\boldsymbol{A}$ be an eigenvector of $\boldsymbol{\Psi}^F$ and λ the corresponding eigenvalue

$$\boldsymbol{\Psi}(\boldsymbol{A}) = (\boldsymbol{\Psi}^F(\boldsymbol{A}^F))^U = (\lambda\boldsymbol{A}^F)^U = \lambda\boldsymbol{A}.$$

Since all eigenvectors of $\boldsymbol{\Psi}^F$ correspond to an eigenmatrix of $\boldsymbol{\Psi}$ and all eigenmatrices of $\boldsymbol{\Psi}^F$ correspond to an eigenvector of $\boldsymbol{\Psi}^F$, it is clear that $k = k'$ and that there exists an $\sigma \in S_k$ such that $\lambda_1, \dots, \lambda_k = \lambda'_{\sigma(1)}, \dots, \lambda'_{\sigma(k)}$ and $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k = \boldsymbol{A}'^F_{\sigma(1)}, \dots, \boldsymbol{A}'^F_{\sigma(k)}$. □

Lemma 43. *Let $\boldsymbol{C} \in \mathbb{R}^{m\times m}, \boldsymbol{A} \in \mathbb{R}^{n\times n}$ be matrices with eigenvalues $\lambda_1, \dots, \lambda_m$ and $\lambda'_1, \dots, \lambda'_n$ respectively, then the eigenvalues of $(\boldsymbol{C} \otimes \boldsymbol{A})$ will be*

$$\lambda_i\lambda'_j, \forall i \in [m], j \in [n].$$

Proof. By Lemma 42 we can conclude that $(\boldsymbol{C}\otimes\boldsymbol{A})$ has mn (not necessarily unique) eigenvalues in total and we will show that $\lambda_i\lambda'_j$ for each of the mn combinations of $i \in [n]$ and $j \in [m]$ is an eigenvalue of $(\boldsymbol{C} \otimes \boldsymbol{A})$.
Fix some $i \in [n]$ and $j \in [m]$ with eigenvectors $\boldsymbol{x}_i$ and $\boldsymbol{x}'_j$, then use the fact that $\boldsymbol{ACB}^\top = (\boldsymbol{C} \otimes \boldsymbol{A})(\boldsymbol{B})$ (Lemma 8) to show

$$\begin{aligned}(\boldsymbol{C} \otimes \boldsymbol{A})((\boldsymbol{x}_i \otimes \boldsymbol{x}'_j)) &= \boldsymbol{C}(\boldsymbol{x}_i \otimes \boldsymbol{x}'_j)\boldsymbol{A}^\top \\ &= \boldsymbol{C}\boldsymbol{x}_i\boldsymbol{x}'^\top_j\boldsymbol{A}^\top \\ &= \boldsymbol{C}\boldsymbol{x}_i(\boldsymbol{A}\boldsymbol{x}'_j)^\top \\ &= \lambda_i\boldsymbol{x}_i(\lambda'_j\boldsymbol{x}'_j)^\top \\ &= \lambda_i\lambda'_j(\boldsymbol{x}_i \otimes \boldsymbol{x}'_j).\end{aligned}$$

□

Definition 44. *We will define a notion of the trace for tensors as follows*

$$\operatorname{tr}(\boldsymbol{\Psi}) = \{\boldsymbol{\Psi}_a{}^a{}_c{}^c\}_{\mathbb{R}}$$

Similarly, the trace for a matrix $\boldsymbol{A}$ can be written in tensor notation as

$$\operatorname{tr}(\boldsymbol{A}) = \{\boldsymbol{A}_a{}^a\}_{\mathbb{R}}$$

Lemma 45. *Let $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m\times n\times n}$ be a tensor, then the trace of a flattened tensor is equivalent to the trace of the tensor*

$$\mathrm{tr}(\boldsymbol{\Psi}) = \mathrm{tr}(\boldsymbol{\Psi}^F).$$

And if $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ are some matrices then the trace of the outer product is the product of the traces of the matrices.

$$\mathrm{tr}((\boldsymbol{A} \otimes \boldsymbol{B})) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$$

where the trace on the right-hand side of either equation is the classic trace for matrices.

Proof. The first result can be shown by writing out the definition of the trace (Definition 44) followed by reindexing and applying the definition of flattening (Definition 31) alongside the definition of the trace for matrices

$$\mathrm{tr}(\boldsymbol{\Psi}) = \sum_{a=1}^{m}\sum_{c=1}^{n} \boldsymbol{\Psi}_a{}^a{}_c{}^c = \sum_{a=1}^{mn} \boldsymbol{\Psi}_{(a \mod m)+1}{}^{(a \mod m)+1}{}_{\lfloor a/m\rfloor+1}{}^{\lfloor a/m\rfloor+1} = \mathrm{tr}(\boldsymbol{\Psi}^F).$$

The second result follows by applying the definition of the trace (Definition 44) alongside the definition of the tensor product (definition26), followed by applying the definition of the trace for matrices twice

$$\mathrm{tr}((\boldsymbol{A} \otimes \boldsymbol{B})) = \sum_{a=1}^{m}\sum_{c=1}^{n} \boldsymbol{A}_a{}^a \boldsymbol{b}_c{}^c = \mathrm{tr}(\boldsymbol{A}) \sum_{c=1}^{n} \boldsymbol{B}_c{}^c = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}).$$

□

Next we will see how the Frobenius inner product and the trace interact

Lemma 46. *Let $\boldsymbol{\Psi}, \boldsymbol{\Phi}$ be tensors, then Frobenius product can be written as a trace*

$$\langle \boldsymbol{\Psi}, \boldsymbol{\Phi} \rangle_F = \mathrm{tr}(\boldsymbol{\Psi}(\boldsymbol{\Phi})).$$

Proof. We start from the trace and apply the definition of one tensor being applied to another tensor (Definition 18), then the definition of the trace (Definition 44), followed by the definition of the Frobenius product (Definition 22)

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{\Psi}(\boldsymbol{\Phi})) &= \mathrm{tr}(\{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Phi}_b{}^e{}_f{}^c\}_a{}^e{}_f{}^d) \\
&= \{\boldsymbol{\Psi}_a{}^b{}_c{}^d \boldsymbol{\Phi}_b{}^a{}_d{}^c\}_{\mathbb{R}} \\
&= \langle \boldsymbol{\Psi}, \boldsymbol{\Phi} \rangle_F.
\end{aligned}$$

□

Lemma 47. *Let $\boldsymbol{A}, \boldsymbol{B}$ be matrices, then*

$$\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^T)$$

Proof. First we apply the definition of matrix multiplication (Definition 18), then the common definition of the trace for matrices. We then transpose both $\boldsymbol{A}$ and $\boldsymbol{B}$ and then we reassemble by executing the first two steps backwards.

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{B}) &= \operatorname{tr}(\{\boldsymbol{A}_a^{T\ b}\boldsymbol{B}_b{}^c\}_a{}^c) \\
&= \{\boldsymbol{A}_a^{T\ b}\boldsymbol{B}_b{}^a\}_{\mathbb{R}} \\
&= \{\boldsymbol{A}_b{}^a\boldsymbol{B}_a^{T\ b}\}_{\mathbb{R}} \\
&= \operatorname{tr}(\{\boldsymbol{A}_b{}^a\boldsymbol{B}_a^{T\ b}\}_a{}^c) \\
&= \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^T).
\end{aligned}$$

□

Lemma 48. *Let $\boldsymbol{\Psi}$ be a tensor and $\boldsymbol{A}, \boldsymbol{B}$ be matrices, then*

$$\operatorname{tr}(\boldsymbol{\Psi}(\boldsymbol{A}) \cdot \boldsymbol{B}) = \langle \boldsymbol{A}^\top, \boldsymbol{\Psi}^\top(\boldsymbol{B}^\top)\rangle_F$$

where $\cdot$ is used to express the common matrix multiplication between $\boldsymbol{\Psi}(\boldsymbol{A})$ and $\boldsymbol{B}$.

Proof. We start by using the definition of $\boldsymbol{\Psi}(\boldsymbol{A})$ (Definition 18), immediately followed by the definitions for matrix multiplication and the trace (Definition 44). Then we transpose, recognize the definition of $\boldsymbol{\Psi}(\boldsymbol{A})$ again and finally apply the definition of the Frobenius product (Definition 22)

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Psi}(\boldsymbol{A}) \cdot \boldsymbol{B}) &= \operatorname{tr}(\{\boldsymbol{\Psi}_a{}^{b}{}_c{}^d\boldsymbol{A}_b{}^c\}_a{}^d \cdot \boldsymbol{B}) \\
&= \operatorname{tr}(\{\boldsymbol{\Psi}_a{}^{b}{}_c{}^d\boldsymbol{A}_b{}^c\boldsymbol{B}_d{}^e\}_a{}^e) \\
&= \{\boldsymbol{\Psi}_a{}^{b}{}_c{}^d\boldsymbol{A}_b{}^c\boldsymbol{B}_d{}^a\}_{\mathbb{R}} \\
&= \{\boldsymbol{\Psi}^\top{}_b{}^a{}_d{}^c\boldsymbol{A}^\top{}_c{}^b\boldsymbol{B}^\top{}_a{}^d\}_{\mathbb{R}} \\
&= \{(\boldsymbol{\Psi}^\top(\boldsymbol{B}^\top))_b{}^c\boldsymbol{A}^\top{}_c{}^b\}_{\mathbb{R}} \\
&= \langle \boldsymbol{A}^\top, \boldsymbol{\Psi}^\top(\boldsymbol{B}^\top)\rangle_F.
\end{aligned}$$

□

3.D Full-Bandits - Proofs

Theorem 49 (Kiefer-Wolfowitz Theorem, Lattimore and Szepesvári (2020, Theorem 21.1))**.** *Let $\mathcal{G} \subset \mathbb{R}^K$ be a convex set with span$(\mathcal{G}) = \mathbb{R}^K$ and $\|\boldsymbol{g}\|_2 \leq \sqrt{m}$ for*

Require: Context distribution $\mathcal{D}$, current policy π_t
1: Draw $\boldsymbol{x} \sim \mathcal{D}$
2: Draw $\boldsymbol{a} \sim \pi_t(\cdot|\boldsymbol{x})$
3: **Output** $(\boldsymbol{x}\boldsymbol{x}^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)^F$

Sampling-Scheme 7: Tensor-Exp3 Sampling Scheme

all $\boldsymbol{g} \in \mathcal{G}$. Then there exists a probability distribution over the points of $\mathcal{G}$, $\mu_{\boldsymbol{g}} \in \mathbb{R}$ for all $\boldsymbol{g} \in \mathcal{G}$ with $\boldsymbol{\mu} \in \Delta_K$ such that

$$K = \max_{\boldsymbol{g} \in \mathcal{G}} \|\boldsymbol{g}\|^2_{\boldsymbol{V}^{-1}},$$

where $\boldsymbol{V} = \sum_{\boldsymbol{g} \in \mathcal{G}} \mu_{\boldsymbol{g}} \boldsymbol{g}\boldsymbol{g}^\top$. Furthermore, $\lambda_{\min}(\boldsymbol{V}) \geq \frac{K}{m}$.

Proof. A $\mu_{\boldsymbol{g}}$ such that $K = \max_{\boldsymbol{g} \in \mathcal{G}} \|\boldsymbol{g}\|^2_{\boldsymbol{V}^{-1}}$ with $\boldsymbol{V}$ as above exists by Lattimore and Szepesvári (2020, Theorem 21.1). Left to prove is $\lambda_{\min}(V) \geq \frac{K}{m}$. We apply $K = \max_{\boldsymbol{g} \in \mathcal{G}} \|\boldsymbol{g}\|^2_{\boldsymbol{V}^{-1}}$ and continue as follows

$$K = \max_{g \in \mathcal{G}} \boldsymbol{g}^\top \boldsymbol{V}^{-1} \boldsymbol{g} = m \max_{\boldsymbol{g} \in \mathcal{G}} \frac{\boldsymbol{g}^\top}{\sqrt{m}} \boldsymbol{V}^{-1} \frac{\boldsymbol{g}}{\sqrt{m}} \leq m \lambda_{\max}\left(\boldsymbol{V}^{-1}\right) = m \lambda_{\min}\left(\boldsymbol{V}\right),$$

where we used the fact that $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|_2^2 \|\boldsymbol{A}\|_{\text{op}}$ and the fact that $\|\mathcal{G}\|_{\text{op}} = \lambda_{\max}(\mathcal{G})$ as $\mathcal{G}$ is positive definite by its construction. Dividing by m provides the desired result. □

Lemma 50. $\boldsymbol{\Psi}_t = \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{a}_t}\left[\left(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top\right) \middle| \mathcal{F}\right]$ *is invertible.*

Proof. We know from Lemma 12 that $\lambda_{\min}(\boldsymbol{\Psi}_t^F) \geq \frac{\gamma K \lambda_{\min}(\boldsymbol{\Sigma})}{m} > 0$. It thus follows that $\boldsymbol{\Psi}_t^F$ is invertible and we conclude the proof by using the fact that $\boldsymbol{\Psi}_t$ is invertible if $\boldsymbol{\Psi}_t^F$ is (Lemma 36). □

We explicitly define a sampling scheme usable by the MGR for the full-bandit case in Sampling Scheme 7.

Lemma 51. *Samples generated by the sampling method detailed in Sampling Scheme 7 are unbiased samples of $\boldsymbol{\Psi}_t$.*

Proof. To show that $(\boldsymbol{x}\boldsymbol{x}^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)$ is indeed an unbiased sample of $\boldsymbol{\Psi}_t$ it is sufficient to take the expectation over $(\boldsymbol{x}\boldsymbol{x}^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)$ explicitly

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}, \boldsymbol{a} \sim \pi_t(\cdot|\boldsymbol{x})}\left[(\boldsymbol{x}\boldsymbol{x}^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)\right] = \boldsymbol{\Psi}_t.$$

□

Lemma 11 (RESTATED). *Fix any $\boldsymbol{x} \in \mathcal{X}$ and suppose that $\tilde{\boldsymbol{\Theta}}_t$ and $\eta > 0$ are such that $\max_t \eta |\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}| < 1$ for all $\boldsymbol{a} \in \mathcal{A}$. Then the regret of Algorithm 6 in context $\boldsymbol{x}$ satisfies*

$$\tilde{\mathcal{R}}_T(\boldsymbol{x}) \leq \frac{\log(|\mathcal{A}|)}{\eta} + \gamma\, U_T(\boldsymbol{x}) + \eta\, \mathbb{E}\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi_t(\cdot|\boldsymbol{x})} \left[(\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2 | \mathcal{F}\right]\right]$$

where $U_T(\boldsymbol{x}) = \sum_{t=1}^{T} \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\mu}_{\boldsymbol{a}} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\boldsymbol{a} - \pi_T^*(\boldsymbol{x})\Big)$ and $\boldsymbol{\mu}$ is the distribution on $\mathcal{A}$ defined by the Kiefer-Wolfowitz theorem.

Proof. The proof will follow the classical Exp3 analysis. Define

$$p_t(\boldsymbol{a}) = \frac{\tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a})}{\sum_{\boldsymbol{a}'\in\mathcal{A}} \tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a}')}$$

By recognizing that $p(\boldsymbol{a})$ is the exponential weights distribution we can apply Van der Hoeven, Van Erven, and Kotlowski (2018, Lemma 1) to find

$$\begin{aligned}
&\sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\Big(\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_t(\boldsymbol{a}|\boldsymbol{x}) \boldsymbol{a} \Big) - \pi^\star(\boldsymbol{x}) \Big) \\
&= (1-\gamma) \sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \boldsymbol{a} - \pi^\star(\boldsymbol{x}) \Big) + \gamma \underbrace{\sum_{t=1}^{T} \sum_{\boldsymbol{a}\in\mathcal{A}} \mu_{\boldsymbol{a}} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \left(\boldsymbol{a} - \pi_T^*(\boldsymbol{x})\right)}_{U_T(\boldsymbol{x})} \\
&\leq (1-\gamma) \left(\frac{\log(|\mathcal{A}|)}{\eta} + \sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \boldsymbol{a} \Big) \right) \\
&\quad + \frac{(1-\gamma)}{\eta} \log \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \exp\Big(-\eta \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} \Big) \Big) + \gamma U_T(\boldsymbol{x}) \,. \qquad (3.20)
\end{aligned}$$

Since by assumption $\eta |\boldsymbol{x} \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}| \leq 1$ we may apply $\exp(-z) \leq 1 - z + z^2$ for $|z| \leq 1$ to find

$$\begin{aligned}
&\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \boldsymbol{a} \Big) + \frac{1}{\eta} \log \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \exp\Big(-\eta \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} \Big) \Big) \\
&\leq \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \Big(\sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \boldsymbol{a} \Big) + \frac{1}{\eta} \log \Big(1 - \sum_{\boldsymbol{a}\in\mathcal{A}} p_t(\boldsymbol{a}) \eta \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} + \eta^2 \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a}) (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2 \Big) \\
&\leq \eta \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a}) (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2,
\end{aligned}$$

where the last inequality is because $\log(1+z) \leq z$ for $|z| \leq 1$. Using the above inequality in Equation (3.20) we find

$$\begin{aligned}
\sum_{t=1}^{T} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \left(\Big(\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_t(\boldsymbol{a}|\boldsymbol{x}) \boldsymbol{a} \Big) - \pi^\star(\boldsymbol{x}) \right) &\leq \frac{\log(|\boldsymbol{a}|)}{\eta} + \eta \sum_{t=1}^{T} \sum_{\boldsymbol{a}\in\mathcal{A}} (1-\gamma) p(\boldsymbol{a}) (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2 + \gamma U_T(\boldsymbol{x}) \\
&\leq \frac{\log(|\boldsymbol{a}|)}{\eta} + \eta \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi_t(\cdot|\boldsymbol{x})} (\boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2 + \gamma U_T(\boldsymbol{x}),
\end{aligned}$$

which completes the proof after taking expectations.

□

Lemma 12 (RESTATED). *For all $t \geq 1$,*

$$\lambda_{\min}(\boldsymbol{\Psi}_t^F) \geq \frac{\gamma K \lambda_{\min}(\boldsymbol{\Sigma})}{m}$$

Moreover, for $\eta \leq \frac{1}{m(M+1)}$, any $\boldsymbol{a} \in \mathcal{A}$, and any $\boldsymbol{x}$ in the support of $\mathcal{D}$ it also holds that $\eta \left| x^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} \right| < 1$.

Proof. Let $\boldsymbol{\Psi}_t$ be as defined in Equation (3.10), then

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{\Psi}_t^F) &= \lambda_{\min}(\boldsymbol{\Psi}_t) \\
&= \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t, \boldsymbol{a}_t}[(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top) | \mathcal{F}]\right) \\
&= \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}[\mathop{\mathbb{E}}_{\boldsymbol{a}_t}[(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top) | \boldsymbol{x}_t, \mathcal{F}]]\right)
\end{aligned}$$

where we first used the fact that $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Psi}_t^F$ agree on eigenvalues (Lemma 42) and plugged in the definition of $\boldsymbol{\Psi}_t$ (Equation (3.10)).

Next we will write the expectation over $\boldsymbol{a}_t$ explicitly, plug in the definition of π_t (Equation (3.13)) and use the fact that

$$\sum_{\boldsymbol{a} \in \mathcal{A}} (1-\gamma) \frac{w_t(\boldsymbol{x}_t, \boldsymbol{a})}{\sum_{\boldsymbol{a}' \in \mathcal{A}} w_t(\boldsymbol{x}_t, \boldsymbol{a}')} \geq 0$$

alongside the fact that $(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top)$ only has positive eigenvalues to continue like follows

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{\Psi}_t^F) &= \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}\left[\sum_{\boldsymbol{a} \in \mathcal{A}} \pi_t(\boldsymbol{a}|\boldsymbol{x}_t)(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)\right]\right) \\
&= \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}\left[\sum_{\boldsymbol{a} \in \mathcal{A}} \left((1-\gamma)\frac{w_t(\boldsymbol{x}_t, \boldsymbol{a})}{\sum_{\boldsymbol{a}' \in \mathcal{A}} w_t(\boldsymbol{x}_t, \boldsymbol{a}')} + \gamma \mu_{\boldsymbol{a}}\right)(\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)\right]\right) \\
&\geq \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}\left[\sum_{\boldsymbol{a} \in \mathcal{A}} \gamma \mu_{\boldsymbol{a}} (\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top)\right]\right)
\end{aligned}$$

Now the only thing that is left to do is to apply Lemma 29, use Lemma 43 to recognize that $\lambda_{\min}(\boldsymbol{A} \otimes \boldsymbol{B}) = \lambda_{\min}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{B})$ and use the Kiefer-Wolfowitz theorem

(Theorem 49).

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Psi}_t^F) &\geq \lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}\left[\sum_{\boldsymbol{a}\in\mathcal{A}}\gamma\mu_{\boldsymbol{a}}(\boldsymbol{x}_t\boldsymbol{x}_t^\top\otimes\boldsymbol{a}\boldsymbol{a}^\top)\right]\right)\\ &= \gamma\lambda_{\min}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_t}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]\otimes\sum_{\boldsymbol{a}\in\mathcal{A}}\mu_{\boldsymbol{a}}\boldsymbol{a}\boldsymbol{a}^\top\right)\\ &= \gamma\lambda_{\min}(\boldsymbol{\Sigma})\lambda_{\min}\left(\sum_{\boldsymbol{a}\in\mathcal{A}}\mu_{\boldsymbol{a}}\boldsymbol{a}\boldsymbol{a}^\top\right)\\ &= \frac{\gamma K\lambda_{\min}(\boldsymbol{\Sigma})}{m}\end{aligned}$$

We now prove the second claim of the lemma. Let $\boldsymbol{a}\in\mathcal{A}$ and $\boldsymbol{x}$ in the support of $\mathcal{D}$. We can start by writing the in the definition of $\tilde{\boldsymbol{\Theta}}_t$, then upper bounding $\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t\boldsymbol{a}_t$ by m and then using Lemma 30.

$$\begin{aligned}|\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}| &= \left|\boldsymbol{x}^\top\hat{\boldsymbol{\Psi}}_t^+(\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t\boldsymbol{a}_t\boldsymbol{a}_t^\top)\boldsymbol{a}\right|\\ &\leq m\left|\boldsymbol{x}^\top\hat{\boldsymbol{\Psi}}_t^+(\boldsymbol{x}_t\boldsymbol{a}_t^\top)\boldsymbol{a}\right|\\ &= m\left|((\boldsymbol{x}\boldsymbol{a}^\top)^F)^\top(\hat{\boldsymbol{\Psi}}_t^+)^F(\boldsymbol{x}_t\boldsymbol{a}_t^\top)^F\right|.\end{aligned}$$

Now, observe that for any $\boldsymbol{x}$ in the support of $\mathcal{D}$ and $\boldsymbol{a}\in\mathcal{A}$

$$\|(\boldsymbol{x}\boldsymbol{a}^\top)^F\|_2 = \sqrt{\sum_{i=1}^d\sum_{k=1}^K(\boldsymbol{x})_i^2(\boldsymbol{a})_k^2} = \sqrt{\left(\sum_{i=1}^d(\boldsymbol{x})_i^2\right)\left(\sum_{k=1}^K(\boldsymbol{a})_k^2\right)} \leq \sigma\sqrt{m}. \tag{3.21}$$

By using that $\|(\hat{\boldsymbol{\Psi}}_t^+)^F\|_{\text{op}} \leq (M+1)\beta$ by Lemma 4, Equation (3.21), and $\beta\leq\frac{1}{m\sigma^2}$ we can see that

$$\begin{aligned}|\boldsymbol{x}^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}| &\leq m\left|((\boldsymbol{x}\boldsymbol{a}^\top)^F)^\top(\hat{\boldsymbol{\Psi}}_t^+)^F(\boldsymbol{x}_t\boldsymbol{a}_t^\top)^F\right|\\ &\leq m\|(\boldsymbol{x}\boldsymbol{a}^\top)^F\|_2\|(\hat{\boldsymbol{\Psi}}_t^+)^F\|_{\text{op}}\|(\boldsymbol{x}_t\boldsymbol{a}_t^\top)^F\|_2\\ &\leq m^2\sigma^2\beta(M+1)\\ &\leq m(M+1).\end{aligned}$$

Using the fact that $\eta\leq\frac{2}{m(M+1)}$ allows us to conclude that $\eta|X^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}|\leq 1$ which finishes the proof □

Lemma 13 (RESTATED). *Fix a $t\in[T]$ and let $\boldsymbol{a}_0\sim\pi_t(\cdot|\boldsymbol{x}_0)$. Then*

$$\mathbb{E}\left[(\boldsymbol{x}_0^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}_0)^2\,\middle|\,\mathcal{F}\right]\leq 2m^2Kd$$

Proof. Most of this proof will be technical calculations, starting from the beginning by plugging in the definition of $\tilde{\boldsymbol{\Theta}}_t$ and upper bounding $(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t)^2$ by m^2

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}_0)^2 | \mathcal{F}\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2 | \mathcal{F}\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t)^2 (\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2 | \mathcal{F}\right] \\
&\leq m^2 \, \mathbb{E}\left[(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2 | \mathcal{F}\right]
\end{aligned}$$

Next we simplify $(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2$ in isolation. We do so by first expanding the square and then using the fact that $\boldsymbol{x}^\top \boldsymbol{y} = \operatorname{tr}(\boldsymbol{x}\boldsymbol{y}^\top)$. Then we use the fact that $\boldsymbol{C}\boldsymbol{B}\boldsymbol{A}^\top = (\boldsymbol{C} \otimes \boldsymbol{A})(\boldsymbol{B})$ by Lemma 8 to obtain

$$\begin{aligned}
(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2 &= \boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0 \boldsymbol{a}_0^\top (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top \boldsymbol{x}_0 \\
&= \operatorname{tr}\Big(\boldsymbol{x}_0 \boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0 \boldsymbol{a}_0^\top (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big) \\
&= \operatorname{tr}\Big((\boldsymbol{x}_0 \boldsymbol{x}_0^\top \otimes \boldsymbol{a}_0 \boldsymbol{a}_0^\top)(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big).
\end{aligned}$$

Here $\cdot$ denotes the classic matrix multiplication and is used to emphasis that the tensor $(\boldsymbol{x}_0 \boldsymbol{x}_0^\top \otimes \boldsymbol{a}_0 \boldsymbol{a}_0^\top)$ is acting on $\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)$.

We now use this result together with the fact that $\boldsymbol{x}_t, \boldsymbol{a}_t$ and $\boldsymbol{x}_0, \boldsymbol{a}_0$ are independent alongside the definition of $\boldsymbol{\Psi}_t$ as follows

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{x}_0^\top \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\operatorname{tr}\Big((\boldsymbol{x}_0 \boldsymbol{x}_0^\top \otimes \boldsymbol{a}_0 \boldsymbol{a}_0^\top)(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\operatorname{tr}\Big(\mathop{\mathbb{E}}_{\boldsymbol{x}_0, \boldsymbol{a}_0}\left[(\boldsymbol{x}_0 \boldsymbol{x}_0^\top \otimes \boldsymbol{a}_0 \boldsymbol{a}_0^\top)\right](\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\operatorname{tr}\Big(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big) \mid \mathcal{F}_t\right].
\end{aligned}$$

This is only possible as $\boldsymbol{x}_0 \sim \mathcal{D}$ and $\boldsymbol{a}_0 \sim \pi_t(\cdot \mid \boldsymbol{x}_0)$, as per assumption on $\boldsymbol{a}_0$.

We isolate the term inside the expectation again, $\operatorname{tr}\Big(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big)$, and use the fact that $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^\top) = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ (Lemma 47). We then use $\operatorname{tr}(\boldsymbol{\Psi}(\boldsymbol{A}) \cdot \boldsymbol{B}) = \langle \boldsymbol{A}^\top, \boldsymbol{\Psi}^\top(\boldsymbol{B}^\top)\rangle$ (Lemma 48). We finish by applying the fact that $\langle \boldsymbol{w}\boldsymbol{x}^\top, \boldsymbol{\Psi}(\boldsymbol{y}\boldsymbol{v}^\top)\rangle = \langle \boldsymbol{\Psi}, (\boldsymbol{y}\boldsymbol{x}^\top \otimes \boldsymbol{v}\boldsymbol{w}^\top)\rangle$ (Lemma 30) alongside the fact that tensors are associative (Lemma 21)

$$\begin{aligned}
\operatorname{tr}\Big(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)) \cdot (\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top))^\top\Big) &= \operatorname{tr}\Big((\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)))^\top \cdot \hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)\Big) \\
&= \langle (\boldsymbol{a}_t \boldsymbol{x}_t^\top), \hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+ (\boldsymbol{x}_t \boldsymbol{a}_t^\top)))\rangle \\
&= \langle \hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^+)), (\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top)\rangle.
\end{aligned}$$

By using the above equality we can thus see that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}\Big(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{x}_t\boldsymbol{a}_t^\top))\cdot(\hat{\boldsymbol{\Psi}}_t^{+}(\boldsymbol{x}_t\boldsymbol{a}_t^\top))^\top\Big)\mid\mathcal{F}_t\right]\\
&=\mathbb{E}\left[\langle\hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^{+})),(\boldsymbol{x}_t\boldsymbol{x}_t^\top\otimes\boldsymbol{a}_t\boldsymbol{a}_t^\top)\rangle\mid\mathcal{F}_t\right]\\
&=\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\left\langle\hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^{+})),\boldsymbol{\Psi}_t\right\rangle\,\middle|\,\mathcal{F}_t\right],
\end{aligned}$$

where in the last equality we used the linearity of the inner product. Observe that by Lemma 27 $\boldsymbol{\Psi}_t$ is symmetric and thus $\boldsymbol{\Psi}_t=\boldsymbol{\Psi}_t^\top$. Now, using that $\mathrm{tr}(\boldsymbol{\Psi}(\boldsymbol{\Phi}))=\langle\boldsymbol{\Psi},\boldsymbol{\Phi}\rangle$ for any tensors of appropriate dimension $\boldsymbol{\Phi},\boldsymbol{\Psi}$ by Lemma 46 and by Lemma 37 we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\left\langle\hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^{+})),\boldsymbol{\Psi}_t\right\rangle\,\middle|\,\mathcal{F}_t\right]\\
&=\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\mathrm{tr}(\boldsymbol{\Psi}_t^\top(\hat{\boldsymbol{\Psi}}_t^{+\top}(\boldsymbol{\Psi}_t(\hat{\boldsymbol{\Psi}}_t^{+}))))\,\middle|\,\mathcal{F}_t\right]\\
&=\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\mathrm{tr}(\boldsymbol{\Psi}_t^{\top F}\hat{\boldsymbol{\Psi}}_t^{+\top F}\boldsymbol{\Psi}_t^{F}\hat{\boldsymbol{\Psi}}_t^{+F})\,\middle|\,\mathcal{F}_t\right]\\
&=\mathop{\mathbb{E}}_{\mathrm{MGR}}\left[\mathrm{tr}(\boldsymbol{\Psi}_t^{F\top}\hat{\boldsymbol{\Psi}}_t^{+F\top}\boldsymbol{\Psi}_t^{F}\hat{\boldsymbol{\Psi}}_t^{+F})\,\middle|\,\mathcal{F}_t\right]\le 2Kd
\end{aligned}$$

where the last equality is due to Lemma 37, which states that we may switch flattening and transpose operations, and the inequality is due to Lemma 4.

By collecting the results we can bound

$$\mathbb{E}\left[(\boldsymbol{x}_0^\top\tilde{\boldsymbol{\Theta}}_t\boldsymbol{a}_0)^2\mid\mathcal{F}_t\right]\le m^2\,\mathbb{E}\left[(\boldsymbol{x}_0^\top\hat{\boldsymbol{\Psi}}_t^{+}(\boldsymbol{x}_t\boldsymbol{a}_t^\top)\boldsymbol{a}_0)^2\mid\mathcal{F}_t\right]\le 2m^2Kd,$$

which concludes the proof. □

Lemma 10 (RESTATED)**.** *Suppose that* $\beta\le\frac{1}{\lambda_{\max}(\boldsymbol{\Psi}_t^F)}$. *Then for* $\tilde{\boldsymbol{\Theta}}_t$ *defined in Equation* (3.11) *and any* $\boldsymbol{a}\in\mathcal{A}$ *and any* $\boldsymbol{x}$ *in the support of* $\mathcal{D}$

$$\boldsymbol{x}^\top(\boldsymbol{\Theta}_t-\tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a}\le\sigma G\sqrt{m}\exp\left(-M\beta\frac{\gamma K\lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right).$$

Proof. We will need to find the exact expectation of $\hat{\boldsymbol{\Psi}}_t^{+}$ and we will do that here by applying the definition of $\tilde{\boldsymbol{\Theta}}_t$ (Eqn. 3.11), followed by applying Lemma 8 and a tower rule and we finish by applying the MGR Lemma (Lemma 4)

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\Theta}}_t\mid\mathcal{F}_t]&=\mathbb{E}[\hat{\boldsymbol{\Psi}}_t^{+}\Big(\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Theta}_t\boldsymbol{a}_t\boldsymbol{a}_t^\top\Big)\mid\mathcal{F}_t]\\
&=\mathop{\mathbb{E}}_{\mathrm{MGR}_t}[\hat{\boldsymbol{\Psi}}_t^{+}(\boldsymbol{\Psi}_t(\boldsymbol{\Theta}_t))\mid\mathcal{F}_t]\\
&=(\mathop{\mathbb{E}}_{\mathrm{MGR}_t}[\hat{\boldsymbol{\Psi}}_t^{+F}\mid\mathcal{F}_t]\boldsymbol{\Psi}_t^F\boldsymbol{\Theta}_t^F)^U\\
&=(\boldsymbol{\Theta}_t^F-(\boldsymbol{I}-\beta\boldsymbol{\Psi}_t^F)^M\boldsymbol{\Theta}_t^F)^U.
\end{aligned}$$

Next we plug in the definition of $\hat{\boldsymbol{\Theta}}_t$, use the above equation, use Lemma 40 and finally use $\|(\boldsymbol{x}\boldsymbol{a}^\top)^F\|_2 \leq \sigma\sqrt{m}$ (Equation (3.21)) alongside $\|\boldsymbol{\Theta}_t\|_F \leq G$.

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{x}^\top(\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a} \mid \mathcal{F}_t]\|_1 &= \|\mathbb{E}[\boldsymbol{x}^\top(\boldsymbol{\Theta}_t - (\boldsymbol{\Theta}_t^F - (\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\boldsymbol{\Theta}_t^F)^U)\boldsymbol{a}]\|_1 \\
&= \|\mathbb{E}[\boldsymbol{x}^\top((\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\boldsymbol{\Theta}_t^F)^U\boldsymbol{a}]\|_1 \\
&= \|\mathbb{E}[(\boldsymbol{x}\boldsymbol{a}^\top)^{F\top}(\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\boldsymbol{\Theta}_t^F]\|_1 \\
&\leq \|\mathbb{E}\left[\|(\boldsymbol{x}\boldsymbol{a}^\top)^F\|_2\|(\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\|_{\text{op}}\|\boldsymbol{\Theta}_t^F\|_2\right]\|_1 \\
&\leq \sigma\sqrt{m}G\|(\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\|_{\text{op}}
\end{aligned}$$

Next we need to bound $\|(\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\|_{\text{op}}$ for which we will first use $\lambda_{\min}(\boldsymbol{\Psi}_t^F) \geq \frac{\gamma K\lambda_{\min}(\boldsymbol{\Sigma})}{m}$ (Lemma 12) alongside the fact that $\boldsymbol{\Psi}_t^F$ is positive semi-definite. Then we apply $1 - z \leq e^{-z}$ (which holds for all $z \in \mathbb{R}$).

$$\|(\boldsymbol{I} - \beta\boldsymbol{\Psi}_t^F)^M\|_{\text{op}} \leq \left(1 - \beta\frac{\gamma K\lambda_{\min}(\boldsymbol{\Sigma})}{m}\right)^M \leq \exp\left(-M\beta\frac{\gamma K\lambda_{\min}(\boldsymbol{\Sigma})}{m}\right)$$

Thus we can now follow that

$$\boldsymbol{x}^\top(\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t)\boldsymbol{a} \leq \sigma G\sqrt{m}\exp\left(-M\beta\frac{\gamma K\lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right) . \tag{3.22}$$

□

Theorem 52. *For any positive $\eta \leq \frac{2}{m(M+1)}$, $\beta \leq \frac{1}{\sigma^2 m}$ and any $\gamma \in (0,1)$ the expected regret of the algorithm satisfies*

$$R_T \leq \frac{\log(|\mathcal{A}|)}{\eta} + 2\eta Tm^2Kd + 2\gamma Tm + 4T\sigma G\sqrt{m}\exp\left(-M\beta\frac{\gamma K\lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right) ,$$

Furthermore, let

$$\gamma = \min\left\{1, \sqrt{K\log(T)\frac{\log(|\mathcal{A}|)}{T\beta\lambda_{\min}^{\boldsymbol{\Sigma}}}}\right\} \qquad \eta = \min\left\{\frac{1}{m(M+1)}, \sqrt{\frac{\log(|\mathcal{A}|)}{Tm^2Kd}}\right\}$$
$$\beta = \frac{1}{\sigma^2 m} \qquad M = \max\left\{\frac{K\log(T)}{\beta m\lambda_{\min}^{\boldsymbol{\Sigma}}}, \sqrt{\frac{TK\log(T)}{\log(|\mathcal{A}|)\beta\lambda_{\min}^{\boldsymbol{\Sigma}}}}\right\} ,$$

then

$$\mathcal{R} \leq 3\sqrt{\log(eK)\,Tm^3Kd} + 3\sqrt{m^4TK\frac{\sigma^2\log(eK)\log(T)}{\lambda_{\min}^{\boldsymbol{\Sigma}}}} + 4\sigma\sqrt{m}G + m^2\log(eK) + \frac{m^3 2\sigma^2 K\log(T)\log(eK)}{\lambda_{\min}^{\boldsymbol{\Sigma}}} .$$

Proof. Most of the work has been done in the previous lemmas already, now we only need to assemble them correctly, control the bias and check the conditions on the hyperparameters. Starting from Lemma 3 we have that

$$\mathcal{R}_T \leq \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}} \left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] + 2 \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}, \mathcal{F}} \left[\sum_{t=1}^{T} \max_{\boldsymbol{a} \in \mathcal{A}} \left\| \mathbb{E}\left[\boldsymbol{x}_0^\top \left(\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t\right) \boldsymbol{a} \,|\, \boldsymbol{x}_0, \mathcal{F}_t\right] \right\|_1 \right] .$$

We focus on the regret first. First, we know that $|\eta \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a}| < 1$ by Lemma 12, which means we can apply Lemma 11. Then also applying Lemma 13 yields

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}} \left[\tilde{\mathcal{R}}_T(\boldsymbol{x}_0)\right] &\leq \frac{\log(|\mathcal{A}|)}{\eta} + \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}} \left[\eta \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_t(\cdot, \boldsymbol{x}_0)} \left[(\boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a})^2 | \mathcal{F}_t\right] + \gamma U_T(\boldsymbol{x}_0)\right] \\
&\leq \frac{\log(|\mathcal{A}|)}{\eta} + 2\eta T m^2 K d + \mathop{\mathbb{E}}_{\boldsymbol{x}_0 \sim \mathcal{D}} \left[\gamma U_T(\boldsymbol{x}_0)\right] .
\end{aligned}$$

Next we need to bound $U_T(\boldsymbol{x}) = \sum_{t=1}^{T} \sum_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} \boldsymbol{x}^\top \tilde{\boldsymbol{\Theta}}_t \left(\boldsymbol{a} - \pi_T^*(\boldsymbol{x})\right)$ in expectation. First we multiply out and use the triangle inequality and finally upper bound $\boldsymbol{x}^\top \boldsymbol{\Theta}_t \boldsymbol{a} \leq m$ which holds for all $\boldsymbol{a}$ and apply Equation (3.22)

$$\begin{aligned}
\mathbb{E}[U_T(\boldsymbol{x}_0)] &= \mathop{\mathbb{E}}_{\mathcal{F}_T, \boldsymbol{x}_0} \left[\sum_{t=1}^{T} \sum_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} \boldsymbol{x}_0^\top \tilde{\boldsymbol{\Theta}}_t \left(\boldsymbol{a} - \pi_T^*(\boldsymbol{x}_0)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} \boldsymbol{x}_0^\top \left(\boldsymbol{\Theta}_t - (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t)\right) \left(\boldsymbol{a} - \pi_T^*(\boldsymbol{x}_0)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} \left(\left|\boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \boldsymbol{a}\right| + \left|\boldsymbol{x}_0^\top \boldsymbol{\Theta}_t \pi_T^*(\boldsymbol{x}_0)\right| + \left|\boldsymbol{x}_0^\top (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t) \boldsymbol{a}\right| + \left|\boldsymbol{x}_0^\top (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t) \pi_T^*(\boldsymbol{x}_0)\right|\right)\right] \\
&\leq \sum_{t=1}^{T} \sum_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} \left(2m + 2\sigma G \sqrt{m} \exp\left(-M\beta \frac{\gamma K \lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right)\right) \\
&= 2Tm + 2T\sigma G \sqrt{m} \exp\left(-M\beta \frac{\gamma K \lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right) .
\end{aligned}$$

By Lemma 10 we know the bias is small

$$\mathbb{E}\left[\boldsymbol{x}_0^\top (\boldsymbol{\Theta}_t - \tilde{\boldsymbol{\Theta}}_t) \boldsymbol{a} \mid \mathcal{F}\right] \leq \sigma G \sqrt{m} \exp\left(-M\beta \frac{\gamma K \lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right) .$$

Combining the last couple equations and using $\gamma \leq 1$ gives that

$$\mathcal{R}_T \leq \frac{\log(|\mathcal{A}|)}{\eta} + 2\eta T m^2 K d + 2\gamma T m + 4T\sigma G \sqrt{m} \exp\left(-M\beta \frac{\gamma K \lambda_{\min}^{\boldsymbol{\Sigma}}}{m}\right) ,$$

which proves the first result of the theorem. For the next result of the theorem, we first set $M = \frac{m \log(T)}{\beta \gamma K \lambda_{\min}^{\boldsymbol{\Sigma}}}$ to find

$$\mathcal{R} \leq \frac{\log(|\mathcal{A}|)}{\eta} + 2T\eta m^2 K d + 2\gamma T m + 4\sigma \sqrt{m} G$$

Next, set $\gamma = \min\left\{1, \sqrt{K\log(T)\frac{\log(|\mathcal{A}|)}{T\beta\lambda^{\mathbf{\Sigma}}_{\min}}}\right\}$ to find

$$\mathcal{R} \leq \frac{\log(|\mathcal{A}|)}{\eta} + 2T\eta m^2 Kd + 2m\sqrt{TK\frac{\log(|\mathcal{A}|)\log(T)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}} + 4\sigma\sqrt{m}G.$$

Finally, set $\eta = \min\left\{\frac{1}{m(M+1)}, \sqrt{\frac{\log(|\mathcal{A}|)}{Tm^2Kd}}\right\}$ and $M = \max\left\{\frac{K\log(T)}{\beta m\lambda^{\mathbf{\Sigma}}_{\min}}, \sqrt{\frac{TK\log(T)}{\log(|\mathcal{A}|)\beta\lambda^{\mathbf{\Sigma}}_{\min}}}\right\}$ to find

$$\begin{aligned}
\mathcal{R} &\leq 3\sqrt{\log(|\mathcal{A}|)Tm^2Kd} + mM\log(|\mathcal{A}|) + m\log(|\mathcal{A}|) + 2m\sqrt{TK\frac{\log(|\mathcal{A}|)\log(T)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}} + 4\sigma\sqrt{m}G \\
&\leq 3\sqrt{\log(|\mathcal{A}|)Tm^2Kd} + 3m\sqrt{TK\frac{\log(|\mathcal{A}|)\log(T)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}} + 4\sigma\sqrt{m}G + m\log(|\mathcal{A}|) \\
&\quad + \frac{mK\log(T)\log(|\mathcal{A}|)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}.
\end{aligned}$$

We examine $\log(|\mathcal{A}|)$ for the final result.

$$\begin{aligned}
\log(|\mathcal{A}|) &\leq \log\left(\sum_{j=1}^{m}\binom{K}{j}\right) \\
&\leq \log\left(\sum_{j=1}^{m}\left(\frac{eK}{j}\right)^j\right) \\
&\leq \log\left(m^m\left(\frac{eK}{m}\right)^m\right) \\
&= m\log\left(eK\right) \qquad (3.23)
\end{aligned}$$

where we used the well known fact that $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ (Knuth, 1997, §1.2.6 : Binomial Coefficients: Exercise 67) where e is Euler's number. This allows us to conclude

$$\begin{aligned}
\mathcal{R} &\leq 3\sqrt{\log\left(eK\right)Tm^3Kd} + 3\sqrt{m^3TK\frac{\log\left(eK\right)\log(T)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}} + 4\sigma\sqrt{m}G + m^2\log\left(eK\right) \\
&\quad + \frac{m^2K\log(T)\log\left(eK\right)}{\beta\lambda^{\mathbf{\Sigma}}_{\min}}.
\end{aligned}$$

which completes the proof of the second result of the theorem after using $\beta = \frac{1}{m\sigma^2}$. □

3.E Lower Bounds - Details

Before we prove the lower bound we first describe a peculiar property of our estimators. Suppose $\mathcal{X}$ consists of only basis vectors. Also suppose that in round

t the context was a basis vector in direction i. Then the feedback obtained at round t does not affect the algorithm's prediction at all subsequent rounds where the context is a basis vector in direction $i' \neq i$. More formally if $\boldsymbol{x}_t = e_i \neq e_j = \boldsymbol{x}_{t'}$ then for all $\boldsymbol{a}$

$$\boldsymbol{x}_{t'}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} = 0.$$

The proof of this statement can be found in Lemma 53.

This implies that Equation (3.1) in Algorithm 3 reduces to

$$\boldsymbol{w}_t(\boldsymbol{e}_i) = \operatorname*{arg\,min}_{\boldsymbol{a}\in \mathrm{Conv}(\mathcal{A})} \sum_{\tau=1}^{t-1} \boldsymbol{e}_i^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{a} + R(\boldsymbol{a}) = \operatorname*{arg\,min}_{\boldsymbol{a}\in \mathrm{Conv}(\mathcal{A})} \sum_{\tau<t\,:\,\boldsymbol{x}_\tau = \boldsymbol{e}_i} \boldsymbol{e}_i^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{a} + R(\boldsymbol{a})$$

Similarly Equation (3.12) of Tensor-Exp3 (Algorithm 6) reduces to

$$\tilde{\boldsymbol{w}}_t(\boldsymbol{x}_t, \boldsymbol{a}) = \exp\left(-\eta \sum_{\tau=1}^{t-1} \boldsymbol{x}_t^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{a}\right) = \exp\left(-\eta \sum_{\tau<t\,:\,X_\tau = \boldsymbol{e}_i} \boldsymbol{x}_t^\top \tilde{\boldsymbol{\Theta}}_\tau \boldsymbol{a}\right).$$

Hence, both algorithms ignore feedback from any previous round τ in which $\boldsymbol{x}_t \neq X_\tau$.

Lemma 53. *Let $\mathcal{X}$ consist of only basis vectors and pick some $t \in [T]$. Let $\boldsymbol{x}_{t'} \neq \boldsymbol{x}_t$ and let $\boldsymbol{a} \in \mathcal{A}$, then*

$$\boldsymbol{x}_{t'}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} = 0$$

holds for the biased and unbiased estimators of CO_2-FTRL (Equation (3.4) and Equation (3.3)) as well as the biased and unbiased estimators of Tensor-EXP3 (Equation (3.11) and Equation (3.10)).

Proof. First we introduce the concept of a n-sparse matrix which we define as a matrix such that $\boldsymbol{a}_{i,j} = 0$ if $i \not\equiv j \mod n$.

Now let $\boldsymbol{A}$ and $\boldsymbol{B}$ be n-sparse, then so is $\boldsymbol{B} + \boldsymbol{A}$ as well as $\boldsymbol{AB}$, which we can recognize by spelling out

$$\boldsymbol{AB} = \{\boldsymbol{B}_a{}^b \boldsymbol{A}_b{}^c\}_a{}^c.$$

Then $\boldsymbol{AB}$ at the index a, c can only be non-zero if there exist some b such that $a \equiv b \mod n$ and $c \equiv b \mod n$, which can only exist if $a \equiv c \mod n$.

Let $\hat{\boldsymbol{P}}^+$ be a sample of the MGR process, we can conclude that it is n-sparse if all samples $\hat{\boldsymbol{P}}_j$ are also n-sparse by writing out

$$\hat{\boldsymbol{P}}^+ = \beta \sum_{k=0}^{M} \prod_{j=1}^{k} (\boldsymbol{I} - \beta \hat{\boldsymbol{P}}_j).$$

Let $\hat{\boldsymbol{P}} \in \mathbb{R}^{d\times d}$ be a sample generated by the Sampling Scheme 5, used by CO_2-FTRL. Then $\hat{\boldsymbol{P}}$ is diagonal and thus d-sparse. It follows that $\boldsymbol{\Sigma}_{t,k}^{-1}$ is also diagonal for all k. Let $\boldsymbol{x}_{t'} = \boldsymbol{e}_i$ and $\boldsymbol{x}_t = \boldsymbol{e}_j$ and pick $k \in [K]$. We now consider the kth entry of the product $X_{t'}^\top \tilde{\boldsymbol{\Theta}}_t$, given by the biased estimator for CO_2-FTRL as defined in Equation (3.4). First we recognize that $\boldsymbol{x}_{t'}$ selects the ith row in the first equality. In the second equality we pull out the scalars $(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k$ and we finish in the last equality be recognizing that $\boldsymbol{x}_t$ selects the jth column.

$$\begin{aligned}(\boldsymbol{x}_{t'}^\top \tilde{\boldsymbol{\Theta}}_t)_k &= (\tilde{\boldsymbol{\Theta}}_t)_{i,k} \\ &= (\boldsymbol{\Sigma}_{t,k}^{-1}\boldsymbol{x}_t(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k)_i \\ &= (\boldsymbol{\Sigma}_{t,k}^{-1}\boldsymbol{x}_t)_i(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k \\ &= (\boldsymbol{\Sigma}_{t,k}^{-1})_{i,j}(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t)_k(\boldsymbol{a}_t)_k\end{aligned}$$

From here it is clear that $(\boldsymbol{\Sigma}_{t,k}^{-1})_{i,j}$ can only be non-zero if $i = j$ and if $j \neq i$, we conclude that

$$\boldsymbol{x}_{t'}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} = 0,$$

showing the first result.

For the biased estimator of Tensor-EXP3, as given in Equation (3.11), we investigate $(\boldsymbol{A} \otimes \boldsymbol{B})^F$ for some matrices $\boldsymbol{A}, \boldsymbol{B}$. By the definition of $\otimes$ (Definition 26) and flattening (Definition 31), we have that

$$\begin{aligned}(\boldsymbol{A} \otimes \boldsymbol{B})^F &= (\{\boldsymbol{A}_a{}^b \boldsymbol{B}^\top{}_c{}^d\}_a{}^b{}_c{}^d)^F \\ &= \{\boldsymbol{A}_{(a \mod m)+1}{}^{(b \mod m)+1} \boldsymbol{B}^\top{}_{\lfloor b/m \rfloor+1}{}^{\lfloor a/m \rfloor+1}\}_a{}^b.\end{aligned}$$

If $\boldsymbol{A}$ is now a diagonal matrix of dimension n, then it is clear that any entry of $(\boldsymbol{A}\otimes\boldsymbol{B})^F$ at the index a, b must be zero if $a \not\equiv b \mod n$, we conclude that $(\boldsymbol{A}\otimes\boldsymbol{B})^F$ is n-sparse. It follows that any $\hat{\boldsymbol{P}}$ drawn using Sampling Scheme 7, the sampling scheme associated with Tensor-Exp3, is d-sparse. As a conclusion $\hat{\boldsymbol{\Psi}}_t^{+\,F}$ is also d-sparse.

Unflattening (Definition 32) some n-sparse matrix $\boldsymbol{C} \in \mathbb{R}^{mn\times mn}$ can only be non-zero for some indices a, b, c, d if $a \equiv b \mod n$ as $\boldsymbol{C}^U{}_a{}^b{}_c{}^d = \boldsymbol{C}_{(a-1)+(d-1)n}{}^{(b-1)+(c-1)n}$.

We first apply the definition of $\tilde{\boldsymbol{\Theta}}_t$ (Equation (3.11)) and then apply Lemma 8

$$\begin{aligned}\boldsymbol{x}_{t'}^\top \tilde{\boldsymbol{\Theta}}_t \boldsymbol{a} &= \boldsymbol{x}_{t'} \hat{\boldsymbol{\Psi}}_t^+ \Big(\boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{a}_t \boldsymbol{a}_t^\top\Big)\boldsymbol{a} \\ &= \boldsymbol{x}_{t'}^\top \hat{\boldsymbol{\Psi}}_t^+ \Big((\boldsymbol{x}_t \boldsymbol{x}_t^\top \otimes \boldsymbol{a}_t \boldsymbol{a}_t^\top)\Big)(\boldsymbol{\Theta}_t)\boldsymbol{a} \\ &= \boldsymbol{x}_{t'}^\top \{\hat{\boldsymbol{\Psi}}_t^+{}_a{}^b{}_c{}^d \boldsymbol{x}_{t\,b} \boldsymbol{x}_t^\top{}^e \boldsymbol{a}_{t\,f} \boldsymbol{a}_t^\top{}^c\}_a{}^e{}_f{}^d (\boldsymbol{\Theta}_t)\boldsymbol{a} \\ &= \{\boldsymbol{x}_{t'}^\top{}^a \hat{\boldsymbol{\Psi}}_t^+{}_a{}^b{}_c{}^d \boldsymbol{x}_{t\,b} \boldsymbol{x}_t^\top{}^e \boldsymbol{a}_{t\,f} \boldsymbol{a}_t^\top{}^c\}_a{}^e{}_f{}^d (\boldsymbol{\Theta}_t)\boldsymbol{a} = 0,\end{aligned}$$

where in the third equality we used definitions of the tensor product (Definition 26) and of $\boldsymbol{\Psi}(\boldsymbol{\Phi})$ (Definition 18). The final equality is due to the fact that $\hat{\boldsymbol{\Psi}}_t^+{}_a{}^b{}_c{}^d = 0$ if $a \neq b$ that for any $\boldsymbol{x}_{t'} \neq \boldsymbol{x}_t$ and any A.

For the unbiased estimator of CO_2-FTRL, as given in Equation (3.3), it is enough to simply recognize that $\mathbb{E}_{\boldsymbol{a}_t,\boldsymbol{x}}\left[(\boldsymbol{a}_t)_k \boldsymbol{x}\boldsymbol{x}^\top \mid \mathcal{F}_t\right]$ is always diagonal if all $\boldsymbol{x}$ are basis vectors. The statement follows by the same arguments as above. For the unbiased estimator of Tensor-EXP3, as given in Equation (3.11), we recognize that applying the MGR with Sampling Scheme 7 and $M = \infty$ yields $\boldsymbol{\Psi}_t^{-1F}$ as shown in Section 3.2 of Neu and Olkhovskaya (2020). To find the i, j coordinate of $\boldsymbol{\Psi}_t^{-1F}$ we can write

$$(\boldsymbol{\Psi}_t^{-1F})_{i,j} = \left(\beta \sum_{k=0}^{\infty} \boldsymbol{C}_k\right)_{i,j} = \beta \sum_{k=0}^{\infty} (\boldsymbol{C}_k)_{i,j}\,,$$

where all $\boldsymbol{C}_k$, as defined by the MGR, are d-sparse as shown above. The rest of the argument follows as above by thus recognizing $\boldsymbol{\Psi}_t^{-1F}$ as d-sparse. □

Theorem 54. *In the semi-bandit setting, for all $T \geq 0.0064\, dK^3$ and for any possibly randomized orthogonal algorithm, there exists a sequence of losses $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_T$ such that $\mathcal{R} \geq 0.017\sqrt{dmKT}$.*

Proof. In the construction of the lower bound we consider sequences of losses that are independent of the actions of the learner and contexts. The contexts are basis vectors sampled uniformly at random. For any sequence of (randomized) context to action mapping π_t we have that

$$\begin{aligned}\mathcal{R} &= \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{\Theta}_t \pi_t(\boldsymbol{e}_i) - \min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{\Theta}_t \boldsymbol{a}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} \pi_t(\boldsymbol{e}_i) - \min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} \boldsymbol{a}\right].\end{aligned}$$

Note that introduced the ghost sample $\boldsymbol{x}_0 = \boldsymbol{e}_i$ as in Lemma 3.

Since we assume that π_t is a orthogonal algorithm, it does not use information from rounds in which $\boldsymbol{x}_\tau \neq \boldsymbol{e}_i$ for $\tau < t$ to compute π_t.

To prove the lower bound we will use Yao's minimax principle, which tells us that it is sufficient to provide a stochastic strategy for the adversary on which the expected regret of any deterministic algorithm is lower bounded. In the construction of the lower bound the action set $\mathcal{A}$ is the set of basis vectors. The losses are generated as follows. We sample Z from the uniform distribution over $[K]$. Conditioned on $Z = k$, the loss $(\boldsymbol{\Theta}_t)_{i,k'}$ is sampled from an independent Bernoulli distribution with mean $\frac{1}{2}$ if $k' \neq k$ and it is sampled from Bernoulli distribution with mean $\frac{1}{2} - \epsilon$ for some $\epsilon \in [0, \frac{1}{4}]$.

We follow the proof of Theorem 7 by Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022). Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022) construct a lower bound for online learning with stochastic feedback graphs, where the only edges in the feedback graphs are self loops which realise with probability $\frac{1}{d}$ (the lower bound

constructed by Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022) is more general, but for our results we only require this particular instance).

Random variables $T_1, \dots, T_K$ denote the number of times that the learner played an $\pi_t(\boldsymbol{e}_i) = \boldsymbol{a}_t^i$ such that $(\boldsymbol{a}_t^i)_k = 1$. For each $k \in [K]$ we introduce notations $\mathbb{P}_k$ and $\mathbb{E}_k$ to denote the probability and expectation with respect to the marginal distributions under which $Z = k$. From Equation (16) in (Esposito, Fusco, Hoeven, and Cesa-Bianchi, 2022) we have that for any deterministic algorithm

$$\mathcal{R} \geq \epsilon \left(T - \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_k[T_k] \right)$$

We also consider auxiliary distribution $\mathbb{P}_0$, which is equivalent to distribution $\mathbb{P}_k$ as specified above but with $\epsilon = 0$ for all k. We denote the corresponding expectation by $\mathbb{E}_0$. Denote by λ_t the feedback set in round t. Denote by $\lambda^t = (\lambda_1, \dots, \lambda_t)$ the tuple of feedback sets the learner has access to in round $t+1$.

Since the action $\pi_t(\boldsymbol{e}_i)$ is fully determined by λ^{t-1}, the central object of interest is the distribution over λ^{t-1} and in particular the information the learner gains from observing certain losses. Observe that if the action in round t given λ^{t-1} is not equal to $\boldsymbol{e}_k$ then the learner does not obtain any information. If the action of the learner given the history of losses is equal to $\boldsymbol{e}_k$ the learner only obtains information if $\boldsymbol{x}_t = \boldsymbol{e}_i$. To formalise this idea, we use Equation (17) of Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022):

$$\mathbb{E}_k[T_k] - \mathbb{E}_0[T_k] \leq \sqrt{\tfrac{1}{2} \sum_{t=1}^{T} \sum_{\lambda^{t-1}} \mathrm{KL}(\boldsymbol{P}_{0,t} \| \boldsymbol{P}_{k,t})},$$

where $\boldsymbol{P}_{k,t}(\lambda_t) = \mathbb{P}_k(\lambda_t | \lambda^{t-1})$ and KL is the KL-divergence. Since the distribution of λ_t given λ^{t-1} is the same under $\mathbb{P}_0$ and $\mathbb{P}_k$ when $\pi_t(\boldsymbol{e}_i) \neq \boldsymbol{e}_k$ the KL-divergence is 0. If $\pi_t(\boldsymbol{e}_i) \neq \boldsymbol{e}_k$ then with probability $\frac{1}{d}$ the learner observes the loss and with probability $1 - \frac{1}{d}$ the learner does not observe anything. Thus, by Equation (18) of Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022) we have that $\mathrm{KL}(\boldsymbol{P}_{0,t} \| \boldsymbol{P}_{k,t}) \leq 8 \log(4/3) \epsilon^2 \frac{1}{d}$. Thus, for all $T \geq 0.0064\, dK^3$, we can now simply follow the remainder of the proof of Theorem 7 of Esposito, Fusco, Hoeven, and Cesa-Bianchi (2022) and use the same parameters to arrive at

$$\left(\mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} \pi_t(\boldsymbol{e}_i) \right] - \mathbb{E}\left[\min_{A} \sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} A \right] \right) \geq 0.017 \sqrt{dKT}\ .$$

Therefore, we may conclude that any orthogonal algorithm satisfies

$$\mathcal{R} \geq 0.017 \sqrt{dKT}$$

which, after observing that $m = 1$, completes the proof. □

Theorem 55. *Suppose $T \geq dmK$ and that $K \geq 2m$. In the full-bandit setting, any orthogonal algorithm satisfies*

$$\mathcal{R} \geq \frac{m^{3/2}\sqrt{dKT}}{16(192 + 96\log(T))}$$

Proof. As in the proof of Theorem 54 the proof heavily relies on the following. In the construction of the lower bound we consider sequences of losses that are independent of the actions of the learner and contexts. The context are basis vectors sampled uniformly at random. For any sequence of (randomized) context to action mapping π_t we have that

$$\begin{aligned}\mathcal{R} &= \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{\Theta}_t \pi_t(\boldsymbol{e}_i) - \min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{\Theta}_t \boldsymbol{a}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} \pi_t(\boldsymbol{e}_i) - \min_{\boldsymbol{a}\in\mathcal{A}} \sum_{t=1}^{T} (\boldsymbol{\Theta}_t)_{i,\cdot} \boldsymbol{a}\right].\end{aligned}$$

Note that introduced the ghost sample $\boldsymbol{x}_0 = \boldsymbol{e}_i$ as in Lemma 3. Since we assume that π_t is an orthogonal algorithm, it does not use information from rounds in which $X_\tau \neq \boldsymbol{e}_i$ for $\tau < t$ to compute π_t. For simplicity we write $\boldsymbol{a}_t^i = \pi_t(\boldsymbol{e}_i)$ and assume that $n = K/m$ is an integer. The set of actions we consider is

$$\mathcal{A} = \left\{\boldsymbol{a} \in \{0,1\}^K : \forall j \in [m] \sum_{k=(j-1)n+1}^{jn} \boldsymbol{a}_k = 1\right\}.$$

In other words, we consider m instances of the n-armed bandit problem.

As in to the proof of Theorem 54, to prove the lower bound we use Yao's minimax principle. The sequence of stochastic losses that we use is almost exactly the same as the sequence of stochastic losses chosen by Cohen, Hazan, and Koren (2017).

As in the proof of the lower bound by Cohen, Hazan, and Koren (2017), we first construct an environment which generates unbounded losses, after which we adapt the lower bound to the bounded loss setting.

Let $\varepsilon = \sigma\sqrt{dmK/4T}$ for some $\sigma > 0$. In each of the m bandit problems, the environment samples the best action uniformly at random. Denote by $\boldsymbol{a}^\star \in \mathcal{A}$ the vector of the best composite action. In every round t, the environment samples $\zeta_t \sim \mathcal{N}(0, \sigma)$ and sets $(\boldsymbol{\Theta}_t)_{i,k} = \frac{1}{2} - \varepsilon(\boldsymbol{a}^\star)_k + \zeta_t$.

We now follow the proof by Cohen, Hazan, and Koren (2017, Lemma 4). We denote by $a_1^\star, \ldots, a_n^\star$ the locations of the non-zero coordinates of $\boldsymbol{a}^\star$, arranged in increasing order. Random variables $T_1, \ldots, T_m$ denote the number of times that the learner played an $\boldsymbol{a}_t^i$ such that $(\boldsymbol{a}_t^i)_{a_j^\star} = 1$. For each $\boldsymbol{a} \in \mathcal{A}$ we introduce notations $\mathbb{P}_{\boldsymbol{a}}$ and $\mathbb{E}_{\boldsymbol{a}}$ to denote the probability and expectation with respect to the

marginal distributions under which $\boldsymbol{a} = \boldsymbol{a}^\star$. By Cohen, Hazan, and Koren (2017, Equation (5)) we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t^i - \boldsymbol{a}^\star)(\boldsymbol{\Theta}_t)_{\cdot,i}\right] = \varepsilon\left(mT - \sum_{j=1}^{m}\frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{a}}[T_j]\right) . \tag{3.24}$$

For every $\boldsymbol{a} \in \mathcal{A}$ we also define the auxillary distribution $\mathbb{P}_{\boldsymbol{a},-j}$ and corresponding expectation $\mathbb{E}_{\boldsymbol{a},-j}$. This is the same distribution as $\mathbb{P}_{\boldsymbol{a}}$ except with $(\boldsymbol{\Theta}_t)_{i,k} = \frac{1}{2} + \zeta_t$. Denote by λ_t the loss observed in round t and by $\lambda^t = (\lambda_1, \ldots, \lambda_t)$ the tuple of losses observed up to and including round t. Crucially, λ_t might be empty since the learner does not observe $(\boldsymbol{\Theta}_t)_{\cdot,i}$ whenever $\boldsymbol{x}_t \neq \boldsymbol{e}_i$. Since the sequence λ^T determines the actions of the algorithm over the game we have that by Pinsker's inequality

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}}[T_j] - \mathbb{E}_{\boldsymbol{a},-j}[T_j] &\leq T\sqrt{\tfrac{1}{2}\mathrm{KL}\Big(\mathbb{P}_{\boldsymbol{a}}[\lambda^T] \| \mathbb{P}_{\boldsymbol{a},-j}[\lambda^T]\Big)} \\ &= T\sqrt{\tfrac{1}{2}\mathbb{E}_{\lambda^{t-1}\sim\mathbb{P}_{\boldsymbol{a},-j}}\Big[\mathrm{KL}\Big(\mathbb{P}_{\boldsymbol{a}}[\lambda_t|\lambda^{t-1}] \| \mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}]\Big)\Big]} \end{aligned} \tag{3.25}$$

We now shift our attention to the single terms in the sum. If $\boldsymbol{x}_t \neq \boldsymbol{e}_i$ then λ_t is the same under $\mathbb{P}_{\boldsymbol{a}}$ and $\mathbb{P}_{\boldsymbol{a},-j}$ irrespective of $\boldsymbol{a}_t^i$ and thus the KL divergence is 0. Similarly, if $(\boldsymbol{a}_t^i)_{a_j^\star} = 0$ then λ_t is the same under $\mathbb{P}_{\boldsymbol{a}}$ and $\mathbb{P}_{\boldsymbol{a},-j}$. Otherwise, if $\boldsymbol{x}_t \neq \boldsymbol{e}_i$ and $(\boldsymbol{a}_t^i)_{a_j^\star} = 1$ then $\mathbb{P}_{\boldsymbol{a}}$ and $\mathbb{P}_{\boldsymbol{a},-j}$ are Gaussian distributions with the same variance $\sigma^2 m^2$ whose means are ε apart. Therefore, by the log-sum inequality we have that

$$\begin{aligned}&\mathrm{KL}\Big(\mathbb{P}_{\boldsymbol{a}}[\lambda_t|\lambda^{t-1}] \| \mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}]\Big) \\ &= \mathrm{KL}\Big((1-\tfrac{1}{d})\,\mathbb{P}_{\boldsymbol{a}}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t \neq \boldsymbol{e}_i] + \tfrac{1}{d}\,\mathbb{P}_{\boldsymbol{a}}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t = \boldsymbol{e}_i] \,\|\, (1-\tfrac{1}{d})\,\mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t \neq \boldsymbol{e}_i] \\ &\qquad\qquad + \tfrac{1}{d}\,\mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t = \boldsymbol{e}_i]\Big) \\ &\leq \frac{1}{d}\mathrm{KL}\Big(\mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t = \boldsymbol{e}_i] \,\|\, \mathbb{P}_{\boldsymbol{a},-j}[\lambda_t|\lambda^{t-1}, \boldsymbol{x}_t = \boldsymbol{e}_i]\Big) \\ &= \frac{\varepsilon^2}{d 2m^2\sigma^2}\end{aligned}$$

Using the above inequality in Equation (3.25) we can see that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}}[T_j] &\leq \mathbb{E}_{\boldsymbol{a},-j}[T_j] + \frac{\varepsilon T}{2m\sigma^2}\sqrt{\tfrac{1}{d}\sum_{t=1}^{T}\mathbb{P}_{\boldsymbol{a},-j}[(\boldsymbol{a}_t^i)_{a_j^\star} = 1]} \\ &= \mathbb{E}_{\boldsymbol{a},-j}[T_j] + \frac{\varepsilon T}{2m\sigma^2}\sqrt{\tfrac{1}{d}\mathbb{E}_{\boldsymbol{a},-j}[T_j]}.\end{aligned}$$

Thus, by Jensen's inequality we have that

$$\begin{aligned}\frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{a}}[T_j] &\leq \frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{a},-j}[T_j] + \frac{\varepsilon T}{2m\sigma^2}\frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\sqrt{\tfrac{1}{d}\mathbb{E}_{\boldsymbol{a},-j}[T_j]} \\ &\leq \frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{a},-j}[T_j] + \frac{\varepsilon T}{2m\sigma^2}\frac{1}{n^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\sqrt{\frac{1}{dn^m}\sum_{\boldsymbol{a}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{a},-j}[T_j]} \\ &\leq \frac{T}{2} + \frac{\varepsilon T}{2\sigma}\sqrt{\frac{T}{dmK}}\end{aligned}$$

where the last inequality follows from Lemma 7 by Cohen, Hazan, and Koren (2017) and the assumption that $K \geq 2m$.

Returning to Equation (3.24) we can see that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\boldsymbol{a}_t^i - \boldsymbol{a}^\star)(\boldsymbol{\Theta}_t)_{\cdot,i}\right] \geq \varepsilon m\left(\tfrac{1}{2} - \frac{\varepsilon}{2\sigma}\sqrt{\frac{T}{dmK}}\right) = \frac{\sigma}{8}m^{3/2}\sqrt{dKT}$$

where the equality follows from $\varepsilon = \sigma\sqrt{dmK/4T}$.

As a final step we have to convert the regret on the unconstrained sequence of losses to the regret on a constrained sequence of losses. Luckily the steps in the proof of Cohen, Hazan, and Koren (2017, Theorem 5) apply to our setting too, and we can see that as long as $T \geq dmK$ we can simply set $\sigma^2 = \frac{1}{192+96\log(T)}$ and choose losses $(\boldsymbol{\Theta}_t)'_{i,k} = \max\Big\{\min\{(\boldsymbol{\Theta}_t)_{i,k}, 1\}, 0\Big\}$ to show that

$$\mathcal{R} \geq \frac{m^{3/2}\sqrt{dKT}}{16(192 + 96\log(T))}$$

which completes the proof. □

3.F Experiments - Additional Graphics

3.F.1 Full-Bandits

Figure 3.F.1: Boxplots over 10 repetitions of the regret (in thousands) of the algorithms in the full-bandit setting using theoretical tuning (lower is better).

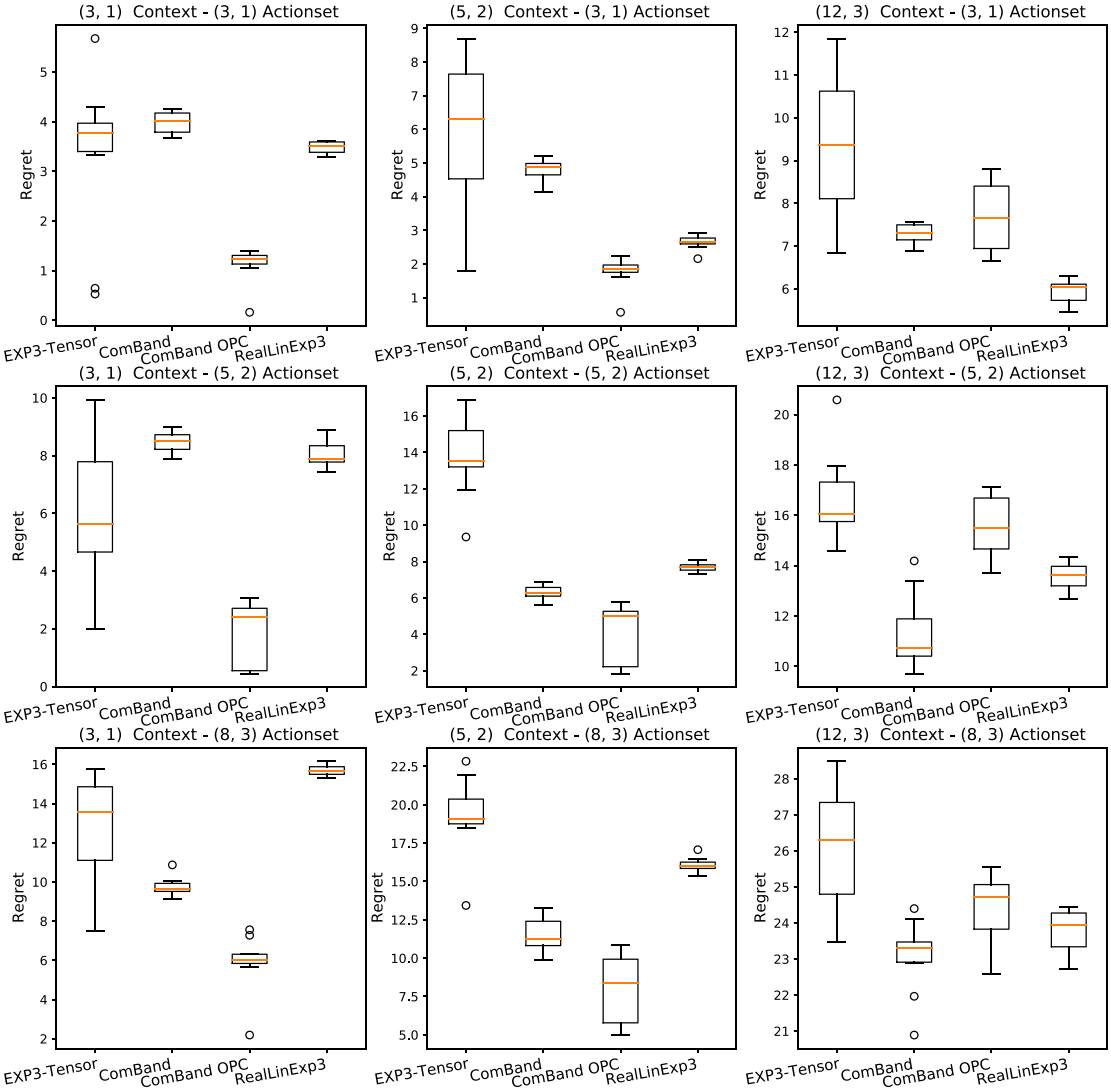
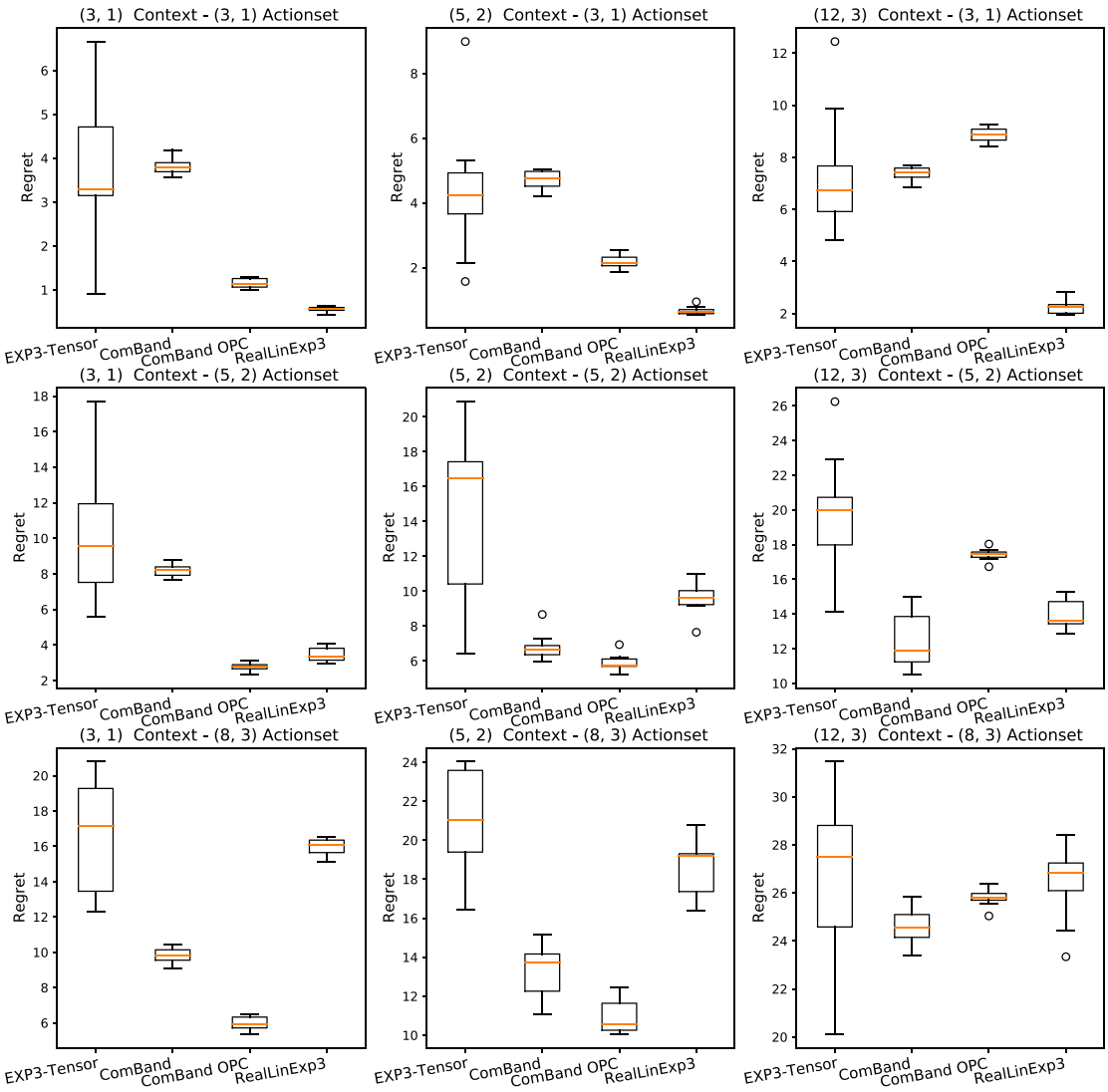


Figure 3.F.2: Boxplots over 10 repetitions of the regret (in thousands) of the algorithms in the full-bandit setting using $1/\sqrt{T}$ tuning (lower is better).



3.F.2 Semi-Bandits

Figure 3.F.3: Boxplots over 10 repetitions of the regret (in thousands) of the algorithms in the semi-bandit setting using theoretical tuning (lower is better).

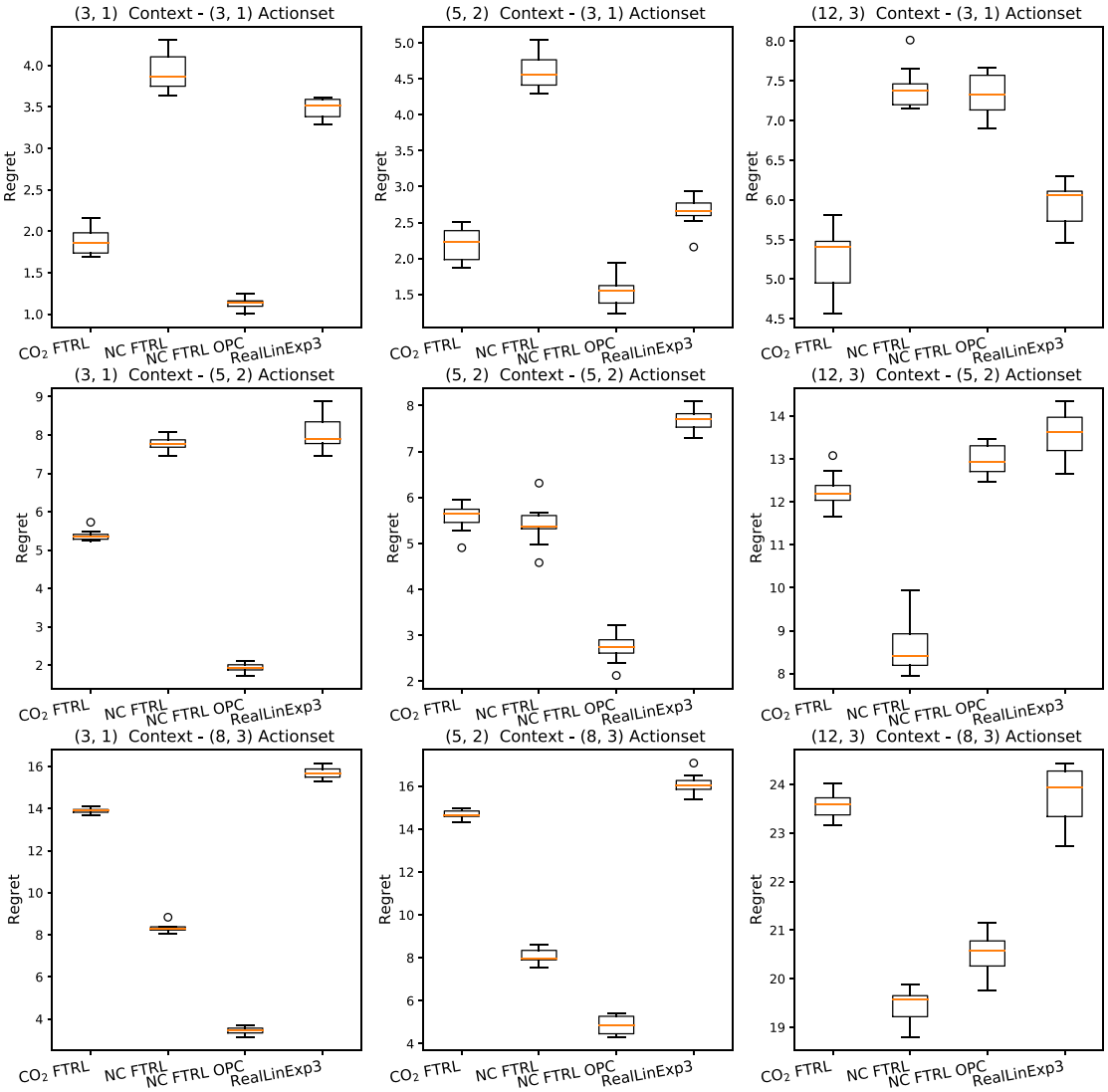
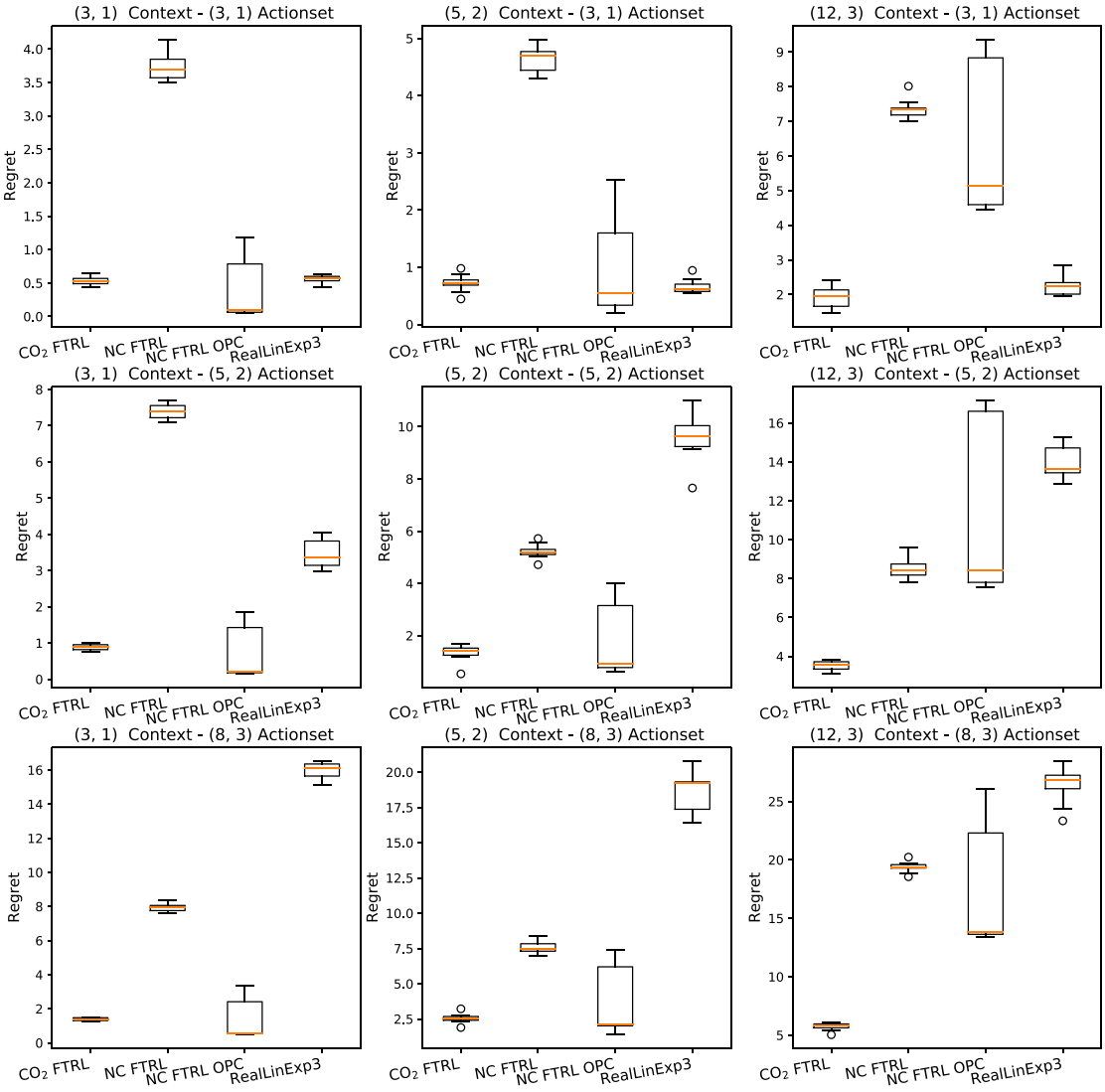


Figure 3.F.4: Boxplots over 10 repetitions of the regret (in thousands) of the algorithms in the semi-bandit setting using $1/\sqrt{T}$ (lower is better).



Chapter 4

Delay Bandits with a Linear Loss

This chapter is based on

van der Hoeven, D., Zierahn, L., Lancewicki, T., Rosenberg, A. & Cesa-Bianchi, N.. (2023). A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs. *Proceedings of Thirty Sixth Conference on Learning Theory*, in *Proceedings of Machine Learning Research* 195:1285-1321 Available from https://proceedings.mlr.press/v195/hoeven23a.html

which was later extend to

Zierahn, L., van der Hoeven, D., Lancewicki, T., Rosenberg, A. & Cesa-Bianchi, N.. (2025). A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs. *Journal of Machine Learning Research (to appear)*.

The author of this dissertation co-derived the theoretical results, co-wrote the paper, and performed the combinatorial bandit experiments.

4.1 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.

In this chapter we present a novel way of analyzing delay. Just like in the previous chapters, the results are based on Follow The Regularized Leader (FTRL), and we show regret bounds for combinatorial bandits semi-bandits, which we have seen in Chapters 2 and 3 already, as well as linear bandits and Markov Decision Processes (MDPs).

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, György, and Szepesvári, 2013; Quanrud and Khashabi, 2015; Joulani, György, and Szepesvári, 2016; Flaspohler, Orabona, Cohen, Mouatadid, Oprescu, Orenstein, and Mackey, 2021) and bandit feedback (Cesa-Bianchi, Gentile, Mansour, and Minora, 2016; Thune, Cesa-Bianchi, and Seldin, 2019; Bistritz, Zhou, Chen, Bambos, and Blanchet, 2019; Zimmert and Seldin, 2020; Ito, Hatano, Sumita, Takemura, Fukunaga, Kakimura, and Kawarabayashi, 2020; György and Joulani, 2021; Van der Hoeven and Cesa-Bianchi, 2022; Masoudian, Zimmert, and Seldin, 2022). Our analysis, unifies previous analyses and sheds light on the impact of delayed bandit feedback in online learning. Our main insight is that one can separate the cost of delayed feedback and bandit feedback through a novel decomposition of the FTRL regret, which allows to separately bound these different regret components. This insight leads to new results in all of the settings we consider. We prove the first regret bounds for combinatorial semi-bandits with delays, which also turn out to be optimal for sufficiently large T (throughout the chapter, by optimal we always mean optimal for sufficiently large T). We also prove the first regret bounds for adversarial MDPs with delays and known transitions, which are again optimal. Finally, we derive a computationally efficient algorithm for linear bandits, whose regret has an optimal dependence on delays.

We now formally introduce the setting of online learning with delayed bandit feedback studied in this chapter. 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 $\boldsymbol{a}_t$ of dimension K from an action set, $\boldsymbol{a}_t \in \mathcal{A} \subseteq \mathbb{R}^K$. Then the learner suffers loss $\boldsymbol{a}_t^\top \boldsymbol{\ell}_t$, where $\boldsymbol{\ell}_t \in \mathbb{R}^K$ is bounded in some suitably chosen norm, and observes $\{\mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau) : \tau + d_\tau = t\}$, where $d_1, \dots, d_T$ is an unknown sequence of delays and $\mathcal{L}$ is an application-specific (possibly randomized) feedback function, encoding which information about $\boldsymbol{\ell}_\tau$ the learner sees based on the action $\boldsymbol{a}_\tau$.

For example, in the combinatorial semi-bandit setting the learner observes all loss components corresponding to the non-zero elements of the action, whereas in the linear bandit setting the learner only observes the scalar $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$. We assume that delays $d_1, \dots, d_T$ and losses $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$ are both generated by an oblivious adversary.

4.1.1 Contributions

This work and the work it extends (Van der Hoeven, Zierahn, Lancewicki, Rosenberg, and Cesa-Bianchi, 2023) has the following main contributions:

New analysis. In section 4.3 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, Dai, and Huang (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 $\max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_1 \leq m$, then the regret after T rounds is of order $\sqrt{m(KT + mD)\log(K)}$, 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 $\Omega(\sqrt{mT(K + md)})$ regret.

Linear bandits. In the linear bandit setting, Ito, Hatano, Sumita, Takemura, Fukunaga, Kakimura, and Kawarabayashi (2020) 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 $\tilde{O}(K\sqrt{T} + \sqrt{dT})$ 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, Hazan, and Rakhlin, 2008), we derive an algorithm that achieves a slightly suboptimal $\tilde{O}(K^{3/2}\sqrt{T} + \sqrt{D})$ 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, Dai, and Huang (2023) show an algorithm with a similar running time, but with a worse regret bound of $\tilde{O}(K^{3/2}\sqrt{T} + K^2\sqrt{D})$.

Adversarial Markov Decision Processes. Delayed feedback in adversarial (finite-horizon and episodic) MDPs was first studied by Lancewicki, Rosenberg, and Mansour, 2022b. Under full-information feedback, where the agent observes the entire cost function at the end of the episode, they achieve the optimal regret bound: $\tilde{O}(H\sqrt{T+D})$, 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 $T^{2/3}+D^{2/3}$, which is far from optimal. The current state-of-the-art guarantees under delayed bandit feedback are by Jin, Lancewicki, Luo, Mansour, and Rosenberg, 2022 and Lancewicki, Rosenberg, and Sotnikov, 2023 who achieve a regret bound of $\tilde{O}(H\sqrt{SAT}+H(HSA)^{1/4}\sqrt{D})$ and $\tilde{O}(H^2\sqrt{SAT}+H^3\sqrt{D})$ in the known transition setting, and a regret bound of $\tilde{O}(H^2S\sqrt{AT}+H(HSA)^{1/4}\sqrt{D})$ and $\tilde{O}(H^3S\sqrt{AT}+H^3\sqrt{D})$ 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, Rosenberg, and Mansour (2022b). Remarkably, the application of our FTRL analysis to adversarial MDPs allows us to close this gap and achieve the first optimal regret bound of $\tilde{O}(H\sqrt{SAT}+H\sqrt{D})$ for the case of known transitions. Moreover, our bound of $\tilde{O}(H^2S\sqrt{AT}+H\sqrt{D})$ 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.

4.1.2 Additional related work

Combinatorial semi-bandits with delayed feedback. Stochastic combinatorial semi-bandits have first been introduced by Gai, Krishnamachari, and Jain (2012) but featured an undesirable dependency on the reciprocal of the square of the smallest gap between arms, which was improved by Chen, Wang, and Yuan (2013) by removing the square. The first matching upper and lower bounds are due to Kveton, Wen, Ashkan, and Szepesvari (2015) by using an upper confidence bound (UCB) based approach, though bounds using Thompson sampling are also known (Wen, Kveton, and Ashkan, 2015). Special cases 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, Hsu, Kale, Karampatziakis, Langford, Reyzin, and Zhang, 2011; Agarwal and Duchi, 2012; Pike-Burke, Agrawal, Szepesvari, and Grunewalder, 2018; Zhou, Xu, and Blanchet, 2019; Gael, Vernade, Carpentier, and Valko, 2020; Lancewicki, Segal, Koren, and Mansour, 2021; Cohen, Daniely, Drori, Koren, and Schain, 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, Linder, Lugosi, and Ottucsák (2007a) achieving a sub-optimal upper bound, an optimal upper bound for m-sets is due to Kale, Reyzin, and Schapire (2010a) and Uchiya, Nakamura, and Kudo (2010). The

optimal bound for semi-bandits in general, which we recover in the non-delayed setting, is due to Audibert, Bubeck, and Lugosi (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, György, Szepesvári, and Antos (2010) and Neu, György, Szepesvári, and Antos (2014) were among the first ones to study the impact of delayed feedback in the nonstochastic setting. Subsequently, Cesa-Bianchi, Gentile, and Mansour (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, Cesa-Bianchi, and Seldin, 2019; Bistritz, Zhou, Chen, Bambos, and Blanchet, 2019; György 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, Linder, Lugosi, and Ottucsák, 2007b; Kale, Reyzin, and Schapire, 2010a; Uchiya, Nakamura, and Kudo, 2010; Cesa-Bianchi and Lugosi, 2012b; Audibert, Bubeck, and Lugosi, 2014; Combes, Talebi Mazraeh Shahi, and Proutiere, 2015; Lattimore, Kveton, Li, and Szepesvari, 2018; Zimmert, Luo, and Wei, 2019).

Adversarial Markov Decision Processes. There is a rich literature on regret minimization in MDPs with non-delayed feedback (Even-Dar, Kakade, and Mansour, 2009; Jaksch, Ortner, and Auer, 2010; Zimin and Neu, 2013; Dick, György, and Szepesvari, 2014; Rosenberg and Mansour, 2019a; Rosenberg and Mansour, 2019b; Rosenberg and Mansour, 2021; Jin, Jin, Luo, Sra, and Yu, 2020; Shani, Efroni, Rosenberg, and Mannor, 2020; Luo, Wei, and Lee, 2021). Under delayed feedback, apart from the literature mentioned earlier, Dai, Luo, and Chen (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, Lancewicki, Luo, Mansour, and Rosenberg (2022) mentioned earlier. Finally, a different line of work (Katsikopoulos and Engelbrecht, 2003; Walsh, Nouri, Li, and Littman, 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, Rosenberg, and Mansour (2022b). A stochastic version of MDPs with delayed feedback has been studied by (Howson, Pike-Burke, and Filippi, 2023b).

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, Kakade, and Hayes (2007) and Abernethy, Hazan, and Rakhlin (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, Hirahara, Soma, and Yoshida, 2020; Zimmert and Lattimore,

2022) obtained the optimal $O(K\sqrt{T})$ regret bound. Stochastic linear bandits with delayed feedback has been studied by Vernade, Carpentier, Lattimore, Zappella, Ermis, and Brueckner (2020) and Howson, Pike-Burke, and Filippi (2023a).

4.2 Preliminaries

We denote by $\hat{\boldsymbol{\ell}}_t \in \mathbb{R}^K$ 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 $\hat{\boldsymbol{L}}$) as follows,

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

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 $\boldsymbol{w}_t(\hat{\boldsymbol{L}})$ do not necessarily coincide. We will specify the relationship between $\boldsymbol{a}_t$ and $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$ in each application.

If we define $\tilde{o}_t = \{\tau : \tau + d_\tau < t\}$ as the set of all losses available at the beginning of round t, then FTRL predicts $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$, where $\hat{\boldsymbol{L}}_t = \sum_{\tau\in\tilde{o}_t} \hat{\boldsymbol{\ell}}_\tau$. We then use the notation $[N] = \{1, \dots, N\}$ and define $\tilde{m}_t = [t-1] \setminus \tilde{o}_t$ 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, Zhou, Chen, Bambos, and Blanchet, 2019; Lancewicki, Rosenberg, and Mansour, 2022b), see also Appendix 4.E.

Additional notations. We denote by $\boldsymbol{w}_{t,i}(\boldsymbol{L})$ 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 $\tilde{O}(\cdot)$ 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

$$\boldsymbol{w}_t(\boldsymbol{L}) = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{W}_t} \boldsymbol{L}^\top \boldsymbol{v} + R_t(\boldsymbol{v}), \tag{4.1}$$

where we require all $\mathcal{W}_t$ to be compact convex sets. Our analysis requires that if we observe the feedback from round τ in timestep t, then the corresponding iterate of FTRL, $\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)$, must be in the same Dikin ellipsoid as the current iterate $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$. 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 $\bar{\Lambda} = [T] \setminus \Lambda$ be the rounds that we do not skip. Since we choose 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: $\mathcal{O}_t = \tilde{o}_t \cap \bar{\Lambda}$, $\mathcal{M}_t = \tilde{m}_t \cap \bar{\Lambda}$. When we observe the loss of round τ, we know whether we have changed the domain since τ and thus $\mathcal{O}_t$ is well defined and non-random given the history $\mathcal{F}_t$. The same is not true for $\mathcal{M}_t$, which can depend on future rounds. This is not a problem for the algorithms considered here, as $\mathcal{M}_t$ is a quantity only used in the analysis and for tuning the learning rates, where $|\tilde{m}_t|$ can be used as an upper bound for $|\mathcal{M}_t|$. The constraints that must be fulfilled to use changing domains are formalized in Assumption 56.

Assumption 56. *For all $t \in [T]$ we assume that $\mathcal{O}_t$ is non-random given the history $\mathcal{F}_t$ and that $\mathcal{W}_t = \mathcal{W}_\tau$ for all $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$. 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 56 reduces to the standard assumption that $\mathcal{W}$ is compact, convex, and non-empty as in that case $\mathcal{O}_t = \tilde{o}_t$ and $\mathcal{M}_t = \tilde{m}_t$.

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

$$\hat{\boldsymbol{L}}_t = \sum_{\tau\in\mathcal{O}_t} \hat{\boldsymbol{\ell}}_\tau\,, \qquad \hat{\boldsymbol{L}}_t^{\mathcal{M}} = \hat{\boldsymbol{L}}_t + \sum_{\tau\in\mathcal{M}_t} \hat{\boldsymbol{\ell}}_\tau\,, \qquad \hat{\boldsymbol{L}}_t^{\star} = \sum_{\tau\in[t]} \hat{\boldsymbol{\ell}}_\tau\,.$$

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

Assumption 57. *Let R_t be the regularizer associated with Equation (4.1) 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) $\kappa \left(\nabla R_t(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)) - \nabla R_{t+\delta}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\right)^\top \boldsymbol{x} \le \frac{\sqrt{\kappa}}{32}\sqrt{\boldsymbol{x}^\top \nabla^2 R_{t+\delta}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\boldsymbol{x}}$ *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 57(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 chapter are approximately self-concordant, Assumption 57(a) is almost automatically satisfied (Nemirovski, 2004), see also Equation 4.18. 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 57(c) is a technical assumption and is satisfied by almost all standard regularizers, including those that we use in this chapter.

4.3 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.

4.3.1 Overview

We build on the analysis of Flaspohler, Orabona, Cohen, Mouatadid, Oprescu, Orenstein, and Mackey (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 $\mathbb{E}[\hat{\boldsymbol{\ell}}_t|\mathcal{F}_t] = \boldsymbol{\ell}_t + \boldsymbol{b}_t$, 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

$$\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right]=\underbrace{\sum_{t=1}^{T}-\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})-\boldsymbol{u})^\top\boldsymbol{b}_t\right]}_{\text{bias}}+\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star})-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\right]}_{\text{cheating regret}} \tag{4.2}$$

$$+\sum_{t=1}^{T}\left(\mathbb{E}\left[\underbrace{(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}))^\top\boldsymbol{\ell}_t}_{H_1}\right]+\mathbb{E}\left[\underbrace{(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star}))^\top\hat{\boldsymbol{\ell}}_t}_{H_2}\right]\right).$$

If $\hat{\boldsymbol{\ell}}_t$ 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 (György and Joulani, 2021, Equation 4)—and can be bounded using the standard be-the-leader lemma (Lemma 73 in Appendix 4.F), see also (Joulani, György, and Szepesvári, 2020, Theorem 3). Now we focus on the second line of Equation (4.2). Typically, H_1 and H_2 are analyzed simultaneously and referred to as "drift", for example, see (György 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 60 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 $\|\boldsymbol{w}_t(\boldsymbol{L})-\boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}}\le 8\|\boldsymbol{L}'-\boldsymbol{L}\|_{R_t,\boldsymbol{v}}$. In order to apply this result for different $\boldsymbol{w}_t(\cdot)$, we require them to lie in the same Dikin ellipsoid, and Lemma 61 establishes machinery to allow us to determine when that is the case. Specifically, if $\boldsymbol{L}'=\boldsymbol{L}+\sum_{\tau\in z}\hat{\boldsymbol{\ell}}_\tau$ for some finite set z and $\sum_{\tau\in z}\kappa\|\hat{\boldsymbol{\ell}}_\tau\|_{R_{t'},\boldsymbol{w}_t(\boldsymbol{L})}\le\frac{1}{32}$, 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 62 to establish that $\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)\in\mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t),\frac{1}{2\sqrt{\kappa}})$ for all $\tau\in\mathcal{M}_t$, which in turn allows us to conclude that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}),\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star})\in\mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t),\frac{1}{2\sqrt{\kappa}})$. Thus, we can repeatedly apply $\|\boldsymbol{w}_t(\boldsymbol{L})-\boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}}\le 8\|\boldsymbol{L}'-\boldsymbol{L}\|_{R_t,\boldsymbol{v}}$ due to Lemma 60, which leads to Lemma 58.

Lemma 58. *Suppose that $\mathbb{E}[\hat{\boldsymbol{\ell}}_t|\mathcal{F}_t]=\boldsymbol{\ell}_t+\boldsymbol{b}_t$ and suppose that Assumption 56 and Assumption 57 hold. Let $t\in[T]$ and $\tau\in\mathcal{M}_t\cup\{t\}$. Suppose that $\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\le\alpha_t$ and $\mathbb{E}\left[\|\hat{\boldsymbol{\ell}}_\tau\|^2_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)}\right]\le\beta_t^2$. Suppose that for all $t,t'\in[T]$ $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\le$*

$\frac{1}{128 d_{\max}}$. *Then for all* $\boldsymbol{u} \in \mathcal{W}_T$,

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

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 58, 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 $\boldsymbol{u} = \left(1-\frac{1}{T}\right)\tilde{\boldsymbol{u}} + \frac{1}{T}\arg\min_{v\in\mathcal{W}} R(\boldsymbol{v})$, we have that the expected regret against $\tilde{\boldsymbol{u}}$ is of order

$$\frac{1}{\eta}\log(K) + d_{\max} K \ln(T) + \eta(KT+D) + \sqrt{\eta}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}(\boldsymbol{\ell}_\tau-\hat{\boldsymbol{\ell}}_\tau)\right\|_{R,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right], \tag{4.3}$$

where we used $\sum_{t=1}^{T}|\mathcal{M}_t| = D$. 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 (4.3). However, due to the delay, if $\tau, \tau' \in \mathcal{M}_t$, then $\hat{\boldsymbol{\ell}}_\tau$ and $\hat{\boldsymbol{\ell}}_{\tau'}$ 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

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

where the second inequality is due to Lemma 62, 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 $\sqrt{\eta|\mathcal{M}_t|\eta K} \le \frac{1}{2}(\eta|\mathcal{M}_t|+\eta K)$ we can see that (4.3) 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 $\hat{\boldsymbol{\ell}}_\tau$ and $\hat{\boldsymbol{\ell}}_{\tau'}$ are independent for $\tau,\tau'\in\mathcal{M}_t$, Lemma 58 implies that we have effectively split the cost of delayed feedback and bandit feedback. We formalize the above in Corollary 59, whose proof can be found in Section 4.3.2.

Corollary 59. *Under the same assumptions as in Lemma 58, suppose that* $\mathbb{E}[\hat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t] = \boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau)^\top\Big(\nabla^2 R_t(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\Big)^{-1}(\hat{\boldsymbol{\ell}}_{\tau'}-\boldsymbol{\ell}_{\tau'})\,\middle|\,\mathcal{F}_t\right] = 0$ *for all* $t\in[T]$ *and all* $\tau,\tau'\in\mathcal{M}_t$ *where* $\tau'\neq\tau$. *Let* $\Lambda=\emptyset$ *and let* $\mathcal{W}_t=\mathcal{W}$. *Then for all* $\boldsymbol{u}\in\mathcal{W}$

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

4.3.2 Analysis Details

In this section we present the proofs of Lemma 58 and Corollary 59. We start by developing the necessary tools in Lemmas 60, 61, and 62. Beginning with the former two, both of which are standard and can be found in various forms in the literature.

Lemma 60. *Suppose that Assumption 57 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* $\|\boldsymbol{w}_t(\boldsymbol{L})-\boldsymbol{w}_t(\boldsymbol{L}')\|^\star_{R_t,\boldsymbol{v}} \le 8\|\boldsymbol{L}'-\boldsymbol{L}\|_{R_t,\boldsymbol{v}}$.

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 Taylor's theorem 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 61. *Suppose that Assumption 57 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 $\mathcal{D}_{R_{t'}}(\boldsymbol{w}_t(\boldsymbol{L}), \frac{1}{2\sqrt{\kappa}})$

$$\boldsymbol{L}'^{\top}\boldsymbol{x} + R_{t'}(\boldsymbol{x}) \geq \boldsymbol{L}'^{\top}\boldsymbol{w}_t(\boldsymbol{L}) + R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})). \tag{4.4}$$

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 (4.4). 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 (4.4) 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 (4.4), 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 (4.4). Let $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{w}_t(\boldsymbol{L})$. Note that

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

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 57(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

$$\begin{aligned}
&\kappa\boldsymbol{L}'^{\top}\boldsymbol{x} + \kappa R_{t'}(\boldsymbol{x}) - \kappa\boldsymbol{L}'^{\top}\boldsymbol{w}_t(\boldsymbol{L}) - \kappa R_{t'}(\boldsymbol{w}_t(\boldsymbol{L})) \\
&= \kappa(\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + \Big(\kappa\boldsymbol{L} + \kappa\nabla R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\Big)^{\top}\boldsymbol{h} + \frac{\kappa}{2}\boldsymbol{h}^{\top}\nabla^2 R_{t'}(\zeta)\boldsymbol{h} \\
&\geq \kappa(\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} - \frac{1}{64} + \frac{1}{2}\boldsymbol{h}^{\top}\nabla^2 R_{t'}(\zeta)\boldsymbol{h} \qquad (4.5) \\
&\geq \kappa(\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} - \frac{1}{64} + \frac{\kappa}{8}\boldsymbol{h}^{\top}\nabla^2 R_{t'}(\boldsymbol{w}_t(\boldsymbol{L}))\boldsymbol{h} \\
&= \kappa(\boldsymbol{L}' - \boldsymbol{L})^{\top}\boldsymbol{h} + \frac{1}{64}
\end{aligned}$$

where we also used Assumption 57(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, $\|\hat{\boldsymbol{\ell}}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}$, 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 58. This lemma might be of independent interest.

Lemma 62. *Suppose that Assumption 56 and Assumption 57 hold. Also suppose that for all $t, t' \in [T]$, $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R_{t'},\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$. Then, for all $t \in [T]$ and all $\tau \in \mathcal{M}_t$ we have that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$.*

Proof. We will prove the statement by induction. Assume that there exists a $t \in [T]$ such that for all $\tau < t$ and all $s \in \mathcal{M}_\tau$, it holds that $\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau) \in \mathcal{D}_{R_\tau}(\boldsymbol{w}_s(\hat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$. Now pick any $s \in \mathcal{M}_t$. For the induction step we need to show that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_s(\hat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$. The goal is to apply Lemma 61 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,

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

For the inequality, we are applying Lemma 72 using the fact that $\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau) \in \mathcal{D}_{R_\tau}(\boldsymbol{w}_s(\hat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$ for $\tau \in \mathcal{M}_s$ and $\boldsymbol{w}_s(\hat{\boldsymbol{L}}_s) \in \mathcal{D}_{R_s}(\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$ for $\tau \geq s$ (where we follow $s \in \mathcal{M}_\tau$ which follows from $s \in \mathcal{M}_t$ and $t \geq \tau$), both of which hold by the inductive assumption. We continue:

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

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

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

Since we have now established that $\sum_{\tau \in o_t \setminus o_s} \sqrt{\kappa} \|\hat{\boldsymbol{\ell}}_\tau\|_{R_t, \boldsymbol{w}_s(\hat{\boldsymbol{L}}_s)} \leq \frac{1}{128 d_{\max}}$ we can apply Lemma 61 to conclude that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_s(\hat{\boldsymbol{L}}_s), \frac{1}{2\sqrt{\kappa}})$ as $\mathcal{W}_t = \mathcal{W}_s$, which holds by Assumption 56. That completes the induction step as we have chosen s arbitrarily. For the basis of induction it is sufficient to note that $\boldsymbol{w}_1(\hat{\boldsymbol{L}}_1) \in \mathcal{D}_{R_1}(\boldsymbol{w}_1(\hat{\boldsymbol{L}}_1), \frac{1}{2\sqrt{\kappa}})$ holds trivially. □

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

Lemma 58 (RESTATED)**.** *Suppose that* $\mathbb{E}[\hat{\boldsymbol{\ell}}_t | \mathcal{F}_t] = \boldsymbol{\ell}_t + \boldsymbol{b}_t$ *and suppose that Assumption 56 and Assumption 57 hold. Let* $t \in [T]$ *and* $\tau \in \mathcal{M}_t \cup \{t\}$*. Suppose that* $\|\boldsymbol{\ell}_\tau\|_{R_t, \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \leq \alpha_t$ *and* $\mathbb{E}\left[\|\hat{\boldsymbol{\ell}}_\tau\|^2_{R_t, \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)}\right] \leq \beta_t^2$*. Suppose that for all* $t, t' \in [T]$ $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$*. Then for all* $\boldsymbol{u} \in \mathcal{W}_T$*,*

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

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

Since $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \leq \frac{1}{128 d_{\max}}$, by Lemma 62 we may conclude that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau), \frac{1}{2\sqrt{\kappa}})$ for all $\tau \in \mathcal{M}_t$ and all t, which is also a prerequisite for Lemma 72. We also note that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\kappa}})$ holds trivially. Now we can conclude that

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

where we used Lemma 72 in the second inequality and the assumption on $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R_{t'}, \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}$ alongside the fact that $|\mathcal{M}_t \cup \{t\}| \leq d_{\max} + 1 \leq 2d_{\max}$ in the third inequality. By Lemma 61 and Assumption 56 we now know that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}), \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star}) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \frac{1}{2})$.

We decompose the regret as follows

$$\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right]=\underbrace{\mathbb{E}\left[\sum_{t\in\Lambda}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{u})^\top\boldsymbol{\ell}_t\right]}_{\text{skipped rounds}}+\underbrace{\sum_{t\in\bar{\Lambda}}-\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})-\boldsymbol{u})^\top\boldsymbol{b}_t\right]}_{\text{bias}} \tag{4.6}$$

$$+\underbrace{\sum_{t\in\bar{\Lambda}}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star})-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\right]}_{\text{cheating regret}}+\sum_{t\in\bar{\Lambda}}\left(\mathbb{E}\left[\underbrace{(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}))^\top\boldsymbol{\ell}_t}_{H_1}\right]+\mathbb{E}\left[\underbrace{(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star}))^\top\hat{\boldsymbol{\ell}}_t}_{H_2}\right]\right).$$

By Hölder's inequality and Lemma 60

$$\begin{aligned}
H_1&=\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}))^\top\boldsymbol{\ell}_t\right]\\
&\le\mathbb{E}\left[\|\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})\|^{\star}_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\\
&\le\mathbb{E}\left[8\|\hat{\boldsymbol{L}}_t-\hat{\boldsymbol{L}}_t^{\mathcal{M}}\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\\
&=\mathbb{E}\left[8\left\|\sum_{\tau\in\mathcal{M}_t}\hat{\boldsymbol{\ell}}_\tau\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\\
&=\mathbb{E}\left[8\left\|\sum_{\tau\in\mathcal{M}_t}\left(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau\right)+\sum_{\tau\in\mathcal{M}_t}\boldsymbol{\ell}_\tau\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\\
&\le\mathbb{E}\left[8\left(\left\|\sum_{\tau\in\mathcal{M}_t}\left(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau\right)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}+\left\|\sum_{\tau\in\mathcal{M}_t}\boldsymbol{\ell}_\tau\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right)\|\boldsymbol{\ell}_t\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\\
&\le 8\alpha_t\,\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}\left(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau\right)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]+8\alpha_t^2|\mathcal{M}_t|\,,
\end{aligned} \tag{4.7}$$

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

$$H_2=\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})-\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star}))^\top\hat{\boldsymbol{\ell}}_t\right]\le 8\beta_t^2\,. \tag{4.8}$$

By Lemma 73 we have that

$$\text{cheating regret}=\sum_{t\in\bar{\Lambda}}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\star})-\boldsymbol{u})^\top\hat{\boldsymbol{\ell}}_t\le R_T(\boldsymbol{u})-\min_{\boldsymbol{v}\in\mathcal{W}_1}R_1(\boldsymbol{v}). \tag{4.9}$$

By combining equations (4.7), (4.8), and (4.9) with the regret decomposition and

leaving the skipped rounds and bias untouched we find

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

which concludes the proof. □

We conclude this section with the proof of Corollary 59.

Corollary 59 (RESTATED)**.** *Under the same assumptions as in Lemma 58, suppose that* $\mathbb{E}[\hat{\boldsymbol{\ell}}_\tau|\mathcal{F}_t]=\boldsymbol{\ell}_\tau$ *and that* $\mathbb{E}\left[(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau)^\top\left(\nabla^2R_t(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\right)^{-1}(\hat{\boldsymbol{\ell}}_{\tau'}-\boldsymbol{\ell}_{\tau'})\,\middle|\,\mathcal{F}_t\right]=0$ *for all* $t\in[T]$ *and all* $\tau,\tau'\in\mathcal{M}_t$ *where* $\tau'\neq\tau$*. Let* $\Lambda=\emptyset$ *and let* $\mathcal{W}_t=\mathcal{W}$*. Then for all* $\boldsymbol{u}\in\mathcal{W}$

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

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

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

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

above together with Jensen's inequality implies that

$$\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}(\hat{\boldsymbol{\ell}}_\tau-\boldsymbol{\ell}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\leq\sqrt{\sum_{\tau\in\mathcal{M}_t}\mathbb{E}\left[\left\|\hat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]}$$
$$\leq\sqrt{4\sum_{\tau\in\mathcal{M}_t}\mathbb{E}\left[\left\|\hat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)}\right]}\leq\sqrt{4|\mathcal{M}_t|\beta_t^2}\,, \quad (4.10)$$

where in the second inequality we used Lemma 62 together with Lemma 72. Finally, the third inequality of (4.10) is due to the assumptions of Lemma 58. We conclude by substituting this bound into the results of Lemma 58,

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

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

4.4 Combinatorial Bandits

In this section, we demonstrate how to apply our generic FTRL approach to combinatorial bandits (CMAB) with delayed feedback. As outlined in the introduction, combinatorial bandits extend multi-armed bandits to be able to efficiently deal with combinatorial decision spaces and have been used in portfolio management (Ni, Xu, Ma, and Fan, 2023) and recommendation systems (Louëdec, Chevalier, Mothe, Garivier, and Gerchinovitz, 2015) among others. In combinatorial 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 $\boldsymbol{a}_t^\top\boldsymbol{\ell}_t$ for an $\boldsymbol{\ell}_t\in[-1,1]^K$ but may not be observed immediately due to the delay. The feedback function is either $\mathcal{L}(\boldsymbol{\ell}_\tau,\boldsymbol{a}_\tau)=\boldsymbol{a}_\tau\odot\boldsymbol{\ell}_\tau$, where $\odot$ is the Hadamard (elementwise) vector product, in the so called semi-bandit setting or $\mathcal{L}(\boldsymbol{\ell}_\tau,\boldsymbol{a}_\tau)=\boldsymbol{a}_\tau^\top\boldsymbol{\ell}_\tau$ in the full-bandit setting. 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.

Input: Regularizers $\{R_t\}_{t\geq 1}$ defined in (4.11), including hyperparameters $\gamma \in (0,1)$ and $\{\eta_t\}_{t\geq 1}$.
for $t \in [T]$ **do**
  Observe $\boldsymbol{a}_\tau \odot \boldsymbol{\ell}_\tau$ for $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Find loss estimators $\hat{\boldsymbol{\ell}}_\tau(i) = \frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)}$ for new observations $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v}\in\mathcal{W}} \hat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$.
  Find probability distribution $\boldsymbol{p}_t$ such that $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$.
  Draw and play $\boldsymbol{a}_t \sim \boldsymbol{p}_t$.
end for

Algorithm 8: Delayed FTRL for combinatorial bandits

In this work we will focus on the semi-bandit setting exclusively. 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 8 is inspired by the algorithm of Audibert, Bubeck, and Lugosi (2014). In any given round t, Algorithm 8 first computes $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$, the solution of the FTRL optimization problem of Eq. (4.1) 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. $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$ can be computed efficiently using standard methods from convex optimization 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 $\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{p}_t}[\boldsymbol{a}] = \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$. 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, Warmuth, and Kivinen, 2010). For a more complete discussion on the computational efficiency of FTRL style combinatorial bandit algorithms and for which actionsets $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$ and $\boldsymbol{p}_t$ can be obtained efficiently we refer to Koolen, Warmuth, and Kivinen (2010), Cesa-Bianchi and Lugosi (2012b), and Audibert, Bubeck, and Lugosi, 2014. The estimator of loss is given by $\hat{\boldsymbol{\ell}}_t(i) = \frac{\boldsymbol{a}_t(i)\boldsymbol{\ell}_t(i)}{\boldsymbol{w}_t(\hat{\boldsymbol{L}}, i)}$, which is unbiased. We use the regularizer

$$R_t(\boldsymbol{v}) = \sum_{i=1}^{K} \left(\frac{1}{\eta_t}\boldsymbol{v}(i)\log(\boldsymbol{v}(i)) - \frac{1}{\gamma}\log(\boldsymbol{v}(i))\right), \tag{4.11}$$

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

Main Result and Discussion The main result of this section is Theorem 63.

Theorem 63. *Suppose that* $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_1 \leq m$. *Algorithm 8 with*

$$\eta_t = \min\left\{\sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{16(m\sum_{t=1}^{T}|\mathcal{M}_t|+Kt)}}, \frac{m^2\left(1+\log\left(\frac{K}{m}\right)\right)}{128K(md_{\max}+K)}\right\}, \quad \gamma = \frac{1}{128\sqrt{m}d_{\max}},$$

guarantees that

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

The result is based on Corollary 59. After confirming the conditions on the regularizer R_t, the proof finds $\alpha_t = \sqrt{\eta_t m}$ 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 59. 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 4.A.

Theorem 64 shows that our results are optimal up to log-factors.

Theorem 64. *Suppose that* $d_t = d$ *for all* t *and that* $m \leq K/2$. *Then for any algorithm there exists a sequence of losses such that*

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

The proof for Theorem 64 can be found in Appendix 4.A. When using an actionset 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 $O(\sqrt{m\left(1+\log\left(\frac{K}{m}\right)\right)KT})$, which also matches the optimal upper bound of order by Audibert, Bubeck, and Lugosi (2014) up to constants.

4.5 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.

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

Algorithm 9: Delayed FTRL for linear bandits

Algorithm Our algorithm for the linear bandit setting is inspired by Abernethy, Hazan, and Rakhlin (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 $R_t(\boldsymbol{v}) = \frac{1}{\eta_t}\|\boldsymbol{v}\|_2^2 + \frac{1}{\gamma_t}\tilde{\Psi}(\boldsymbol{v})$, where $\tilde{\Psi}(\boldsymbol{v}) = \Psi(\boldsymbol{v}) - \min_{\boldsymbol{v}'\in\mathcal{W}} \Psi(\boldsymbol{v}')$ 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 4.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 57(a) and 57(b). As detailed in Appendix 4.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

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

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 $\boldsymbol{a}_t = \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) + \Big(\nabla^2\Psi(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\Big)^{-1/2}\boldsymbol{v}_t$, 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 4.B). Since $\|\boldsymbol{a}_t - \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)\|_{\Psi,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} = 1$, we have that $\boldsymbol{a}_t \in \mathcal{W}$. As for the loss estimate, we use $\hat{\boldsymbol{\ell}}_t = K\boldsymbol{\ell}_t^\top \boldsymbol{a}_t \Big(\nabla^2\Psi(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\Big)^{1/2}\boldsymbol{v}_t$, 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 9: the computation of $w_t(\hat{\boldsymbol{L}}_t)$ and the computation of $\left(\nabla^2\Psi(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\right)^{1/2}$ and its inverse. $\left(\nabla^2\Psi(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\right)^{1/2}$ and its inverse can be computed through an eigenvalue decomposition, which can be done in $O(K^3)$. Abernethy, Hazan, and Rakhlin (2008) show that an approximation of $w_t(\hat{\boldsymbol{L}}_t)$ 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 4.B.1.

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

Theorem 65. *Suppose that $T > 100$ and $B \geq 1$. Algorithm 9, run with a ν-self-concordant barrier Ψ and with*

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

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 65 can be found in Appendix 4.B. It follows from an application of Corollary 59 and carefully tuning the learning rates.

Let us put Theorem 65 in perspective. For arbitrary $\mathcal{W}$ we can use the entropic barrier as the regularizer, which means $\nu = d$ and thus algorithm 9 has a $\tilde{O}(K^{2/3}\sqrt{T} + \sqrt{D})$ regret bound. For constant delay, i.e. $d_t = d$, Ito, Hatano, Sumita, Takemura, Fukunaga, Kakimura, and Kawarabayashi (2020) show that continuous exponential weights obtains a $\tilde{O}(K\sqrt{T} + \sqrt{dT})$ 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 9 can be computed in $O(K^3)$ time, with only a slightly worse regret bound. Huang, Dai, and Huang (2023) provide an algorithm with similar computational complexity as algorithm 9, but their regret bound is $\tilde{O}(K^{2/3}\sqrt{T} + K^2\sqrt{D})$, 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, Dai, and Huang (2023) can be improved

to $\tilde{O}(K\sqrt{\nu T} + K\sqrt{\nu D})$. In their terminology: Banker-BOLO is $(O(\nu \log(T), K^2))$-stable, which together with Theorem 4.6 of Huang, Dai, and Huang (2023) leads to a $\tilde{O}(K\sqrt{\nu T} + K\sqrt{\nu D})$ regret bound. Still, the unnecessary dependence on the dimension K in the delay term of the regret bound remains.

4.6 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

$$\mathscr{M} = (\mathcal{S}, \mathcal{A}_{\mathscr{M}}, H, p, \{\boldsymbol{\ell}_t\}_{t=1}^T, s_{\text{init}}) \ ,$$

where $\mathcal{S}$ and $\mathcal{A}_{\mathscr{M}}$ 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 $p : [H] \times \mathcal{S} \times \mathcal{A}_{\mathscr{M}} \to \Delta_{\mathcal{S}}$, 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 $\pi_t = [H] \times \mathcal{S} \to \Delta_{\mathcal{A}_{\mathscr{M}}}$ 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 $\mathscr{M}$.

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

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

Input: Regularizers $\{R_t\}_{t\geq 1}$ defined in (4.12).
for $t = 1, ..., T$ **do**
  Observe feedback $\boldsymbol{\ell}_t(h, s_{\tau,h}, a_{\tau,h})$ for $h \in [H], \tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute upper occupancy bounds $\boldsymbol{q}_\tau^{\max}(h,s,a) = \max_{\hat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi_\tau,\hat{p}}(h,s,a)$.
  Compute $\hat{\boldsymbol{\ell}}_\tau(h,s,a) = \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}\boldsymbol{\ell}_\tau(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)}$ for $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v}\in\mathcal{W}} \hat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$ and policy
  $\pi_t(a \mid h, s) = \frac{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a)}{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s)}$.
  Play episode t with policy π_t
end for

Algorithm 10: Delayed FTRL for adversarial MDPs

By Rosenberg and Mansour, 2019a—see also Zimin and Neu (2013) and Dick, György, and Szepesvari (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 $\mathscr{M}$ is denoted by $\Delta(\mathscr{M})$, and the set of all policies by $\Pi = \{\pi : [H] \times \mathcal{S} \to \Delta_{\mathcal{A}_\mathscr{M}}\}$. 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

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

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 $\boldsymbol{\ell}_t \in [0,1]^{HSA}$, $\mathcal{A} = \Delta(\mathscr{M})$ 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 10 is based on the general framework presented in Section 4.3. To satisfy the stability conditions required for Lemma 58, we employ a

hybrid regularization of negative entropy and log-barrier just like in the combinatorial bandit case:

$$R_t(\boldsymbol{v}) = \frac{1}{\eta_t} \sum_{h,s,a,s'} \boldsymbol{v}(h,s,a,s') \log \boldsymbol{v}(h,s,a,s') - \frac{1}{\gamma} \sum_{h,s,a,s'} \log \boldsymbol{v}(h,s,a,s'). \tag{4.12}$$

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

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

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 66. *$\mathcal{W}$ satisfies the following:*

1. *For any $\boldsymbol{q} \in \Delta(\mathscr{M})$, there exists $\tilde{\boldsymbol{q}} \in \mathcal{W}$ such that $\min_{h,s,a,s'} \tilde{\boldsymbol{q}}(h,s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$ and $\|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|_1 \leq \frac{2H}{T}$.*
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 $\boldsymbol{q}^{max}(h,s,a) = \max_{\hat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi,\hat{p}}(h,s,a)$. 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 4.C. The importance-weighted loss estimator for Algorithm 10 is inspired by Jin, Jin, Luo, Sra, and Yu (2020),

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

where $\boldsymbol{q}_\tau^{\max}(h,s,a) = \max_{\hat{p}\in\mathcal{P}} \boldsymbol{q}^{\pi_\tau,\hat{p}}(h,s,a)$ is an upper bound on the occupancy measure for each state h,s,a when following policy π_τ. That means that $\hat{\boldsymbol{\ell}}_\tau$ 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, Hazan, and Rakhlin, 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, Jin, Luo, Sra, and Yu, 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 Theorem67.

Theorem 67. *Suppose that $T \geq H$. Algorithm 10 with*

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

guarantees

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

The proof relies on yet another regret decomposition given by

$$\mathcal{R}_T = \sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle = \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle}_{\textsc{Shift-Penalty}} \ .$$

$\Delta(\mathcal{P})$ is only slightly larger than $\Delta(\mathscr{M})$, and we can easily bound Error using the first property in Lemma 66. Since $R_T(\boldsymbol{u})$ can be arbitrarily large near the boundary of the domain, we slightly shift $\boldsymbol{u}$ to $\tilde{\boldsymbol{u}}$ using the first property in Lemma 66 to ensure that (i) $R_T(\tilde{\boldsymbol{u}}) \leq \tilde{O}\left(\frac{HS^2A}{\gamma}\right)$, and (ii) Shift-Penalty is bounded by $2H$. We can not apply Corollary 59 to bound Reg because of the bias in our estimator. We apply Lemma 58 instead. By Lemma 80 we can show that R_t satisfies Assumption 57 and that $R_T(\tilde{\boldsymbol{u}}) - \min_{\boldsymbol{v}\in\mathcal{W}_1} R_1(\boldsymbol{v})$ is bounded by $\tilde{O}\left(\frac{H}{\eta_T} + \frac{HS^2A}{\gamma}\right)$.

The fact that $\boldsymbol{q}_t^{\max}(h,s,a)$ upper bounds both $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t,h,s,a)$ 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 66, 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 $\|\sum_{\tau\in\mathcal{M}_t}(\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}$. This term, which arises because we are using biased estimators, is not present when applying Corollary 59. The full proof can be found in Appendix 4.C.

The algorithm is optimal, matching the lower bound of Lancewicki, Rosenberg, and Mansour (2022b) up to log-factors and improves on previous state-of-the-art regret bounds $\tilde{O}\left(H^2S\sqrt{AT} + H(HSA)^{1/4}\sqrt{D}\right)$ by Jin, Lancewicki, Luo, Mansour, and Rosenberg (2022) and $\tilde{O}\left(H^2\sqrt{SAT} + H^3\sqrt{D}\right)$ by Lancewicki, Rosenberg, and Sotnikov (2023).

4.7 Adversarial MDPs with Unknown Transitions

In this section, we apply our FTRL approach to adversarial Markov Decision Processes (MDPs) setting detailed in Section 4.6, 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.

Initialize $j = 1$, $\hat{\mathcal{P}}_1$ as the set of all transition functions, $\mathcal{W}_0 = \Delta(\hat{\mathcal{P}}_1)$.
For all $(h, s, a, s') \in [H] \times \mathcal{S} \times \mathcal{A}_{\mathscr{M}} \times \mathcal{S}$ 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 $h \in [H], \tau \in \tilde{o}_t \setminus \tilde{o}_{t-1}$.
  Update counters: $N_j(s_{\tau,h+1}|h, s_{\tau,h}, a_{\tau,h}) \mathrel{+}= 1$ for $h \in [H], \tau \in \tilde{o}_t \setminus \tilde{o}_{t-1}$.
  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**
    $j \mathrel{+}= 1$
    For all $(h, s, a, s') \in \mathcal{S} \times \mathcal{A}_{\mathscr{M}} \times \mathcal{S}$, set $N_j(s'|h, s, a) = N_{j-1}(s'|h, s, a)$.
    Update set $\hat{\mathcal{P}}_j$ as in Equation (4.13).
    Set $\mathcal{W}_t = \bigcap_{j'=1}^{j} \Delta(\hat{\mathcal{P}}_{j'})$. If $\mathcal{W}_t = \emptyset$ then set $\mathcal{W}_t = \Delta(\hat{\mathcal{P}}_j)$.
    Skip all rounds that are missing by adding all elements in $\tilde{m}_t$ 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 $h \in [H], \tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute upper occupancy bounds $\boldsymbol{q}_\tau^{\max}(h, s, a) = \max_{\hat{p} \in \hat{\mathcal{P}}_j} \boldsymbol{q}^{\pi_\tau, \hat{p}}(h, s, a)$.
  Compute $\hat{\boldsymbol{\ell}}_\tau(h, s, a) = \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}\boldsymbol{\ell}_\tau(h,s,a)}{\boldsymbol{q}_\tau^{\max}(h,s,a)+\xi}$ for new observations $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) = \arg\min_{\boldsymbol{v} \in \mathcal{W}_t} \hat{\boldsymbol{L}}_t^\top \boldsymbol{v} + R_t(\boldsymbol{v})$ and policy
  $\pi_t(a \mid h, s) = \frac{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a)}{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s)}$.
  Play episode t with policy π_t
end for

Algorithm 11: Delayed FTRL for adversarial MDPs with unknown transitions

Algorithm Algorithm 11 is very similar to the one presented in Section 4.6 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, Jin, Luo, Sra, and Yu (2020). The confidence sets are updated in epochs. Specifically, the algorithm maintain 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

$$\hat{\mathcal{P}}_j = \left\{ \hat{p} : \left| \hat{p}(s'|h, s, a) - \bar{p}_j(s'|h, s, a) \right| \leq \epsilon_j(s'|h, s, a)\,,\ \forall (h, s', a, s) \in [H] \times \mathcal{S} \times \mathcal{A}_{\mathscr{M}} \times \mathcal{S} \right\}, \tag{4.13}$$

where

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

for $\bar{p}_j(s'|h,s,a) = \frac{N_j(s'|h,s,a)}{N_j(h,s,a)}$ 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 $\Delta(\hat{\mathcal{P}})$. That is, if round t is in epoch j, then $\mathcal{W}_t = \bigcap_{j'=1}^{j} \Delta(\hat{\mathcal{P}}_{j'})$. Lemma 68 below shows that the true transition function is contained in our confidence set with high probability.

Lemma 68. *With probability at least $1 - 4/T^2$, we have $p \in \hat{\mathcal{P}}_j$ for all j.*

Proof. The proof is a straightforward modification of the proof of Lemma 2 of Jin, Jin, Luo, Sra, and Yu (2020). □

As a consequence of Lemma 68, 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 $\mathcal{W}_{t+1} \subseteq \mathcal{W}_t$. This will later be crucial in the analysis to apply the be-the-leader lemma (Lemma 73 in Appendix 4.F). In order to ensure that Assumption 56, specifically the fact that $\mathcal{W}_t = \mathcal{W}_\tau$ for all outstanding observations $\tau \in \mathcal{M}_t$, 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 4.6), 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 $O(\mathrm{poly}(H,S,A,\log T))$. 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 69. 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 4.6.

Theorem 69. *Algorithm 11 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,*

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

The proof relies on the same regret decomposition as in Section 4.6.

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

The Error and Shift-Penalty terms are bounded using standard arguments (see Lemma 84 in the appendix). To bound the Reg term we will apply Lemma 58 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 $\tau \in \mathcal{O}_t \backslash \mathcal{O}_{t-1}$, 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 58, 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 4.D.

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

4.8 Experiments

In this section we are evaluating the performance of Algorithm 8 and Algorithm 9 on synthetic experiments. The full code for the experiments can be found here[1]

4.8.1 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 $\mathcal{M} = 3$ and $K = 10$, the losses 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, that is the oscillating arms are the good arms and the constants are the bad arms. We use $\mathcal{M} = 3$ oscillating and 7 fixed arms. As mentioned earlier, Algorithm 8 is the first algorithm for delayed adversarial

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

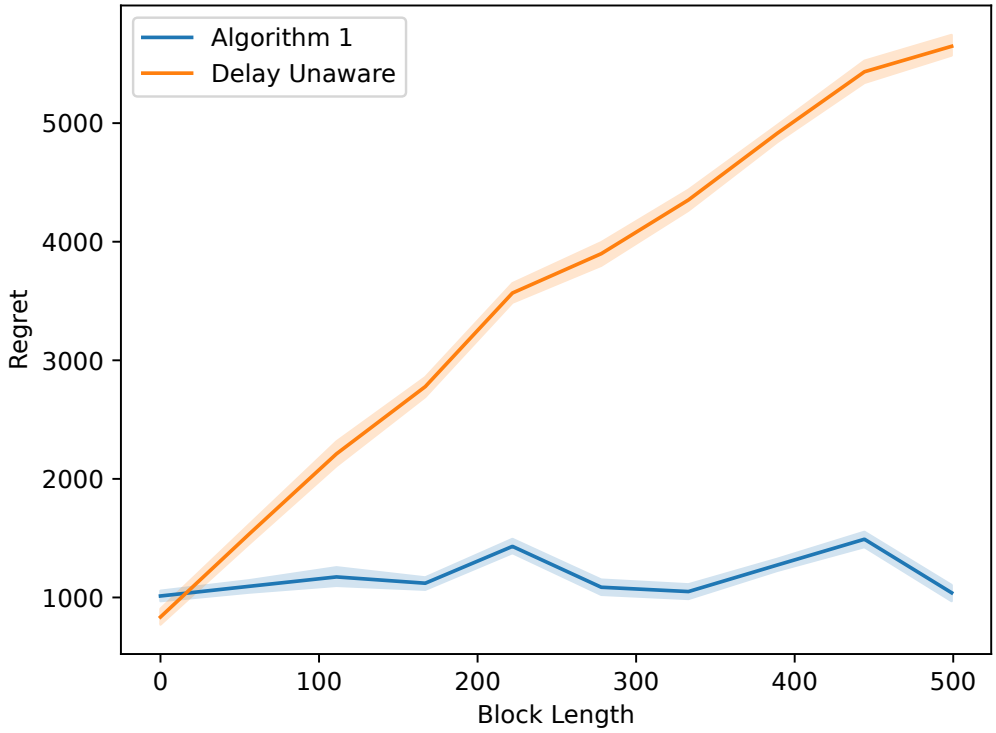


Figure 4.8.1: Regret with 95% confidence intervals over 20 repetitions over $T = 10000$ rounds.

combinatorial bandits thus we will compare it to a standard algorithm not adapted to delay. Namely, an FTRL based version of the algorithm presented in Figure 3 of Audibert, Bubeck, and Lugosi (2014), which is the same as running Algorithm 8 with regularizer $R_t(\boldsymbol{v}) = \sum_{i=1}^{K} \frac{1}{\eta_t} \boldsymbol{v}(i) \log(\boldsymbol{v}(i))$ and learning rate $\eta_t = \sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{16Kt}}$.

The results of experiments for varying block sizes can be found in Figure 4.8.1. 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-regularizing. But even for small delays, the delay aware tuning outperforms the non-delayed tuning significantly.

4.8.2 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 9, Algorithm 12, and Banker-BOLO (Huang, Dai, and Huang, 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 4.5. This version of Banker-BOLO is denoted by Banker-BOLO-V2. A fifth possible algorithm to compare with is the algorithm of Ito, Hatano, Sumita, Takemura, Fukunaga, Kakimura, and Kawarabayashi (2020). This is an instance of continuous exponential weights, 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 4.8.2 in the main body, and Figures 4.G.1 and 4.G.2 in Appendix 4.G. As predicted by theory, the regret grows with the square root of the block size for all algorithms. 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 a $\sqrt{D}$ term in the regret bound for the cost of delay, given that the algorithm is appropriately tuned. We do see that our algorithms consistently outperform both versions of Banker-BOLO. However, we do believe that with the right tuning a version of OMD will perform similarly to our algorithms. We do see that Banker-BOLO-V2 has better performance than Banker-BOLO, which is predicted by theory. We also see that the performance of Algorithms 9 and 12 hardly differs, which is also predicted by theory.

4.9 Conclusion

In sections 4.4, 4.6, and 4.7 we have shown that FTRL leads to optimal regret bounds under delayed feedback in combinatorial semi-bandits, MDPs with known transitions, and MDPs with unknown transitions. Furthermore, in section 4.5 we have provided an efficient algorithm with nearly optimal regret for linear bandits.

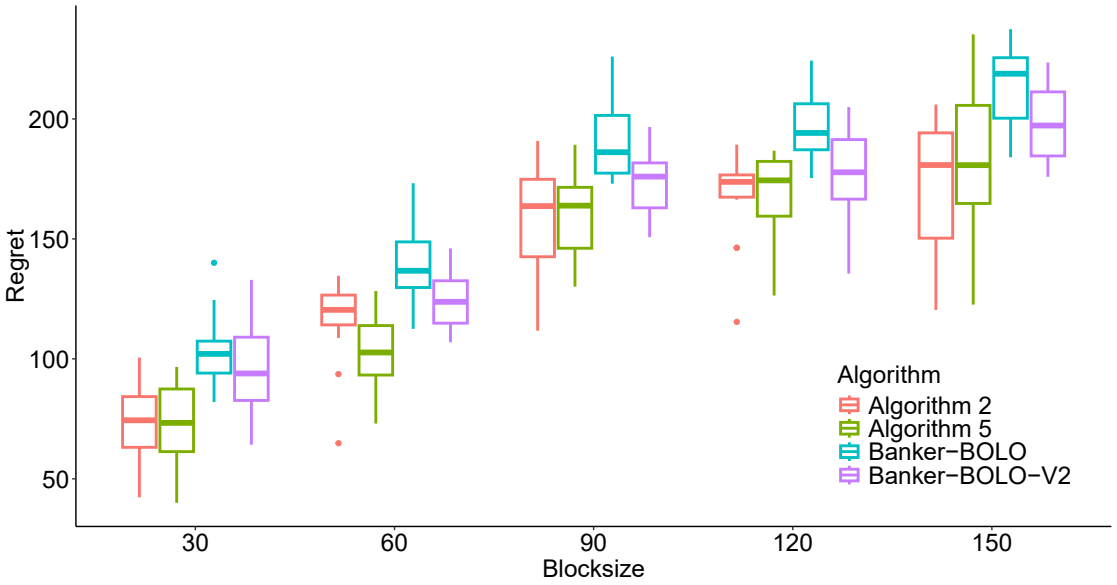


Figure 4.8.2: Boxplot of the regret over 20 repetitions over $T = 10000$ rounds with $K = 20$.

In section 4.8 we have shown that Algorithm 8 and Algorithm 9 outperform delay-unaware and previous algorithms respectively on our synthetic datasets.

Appendix

4.A Combinatorial Bandits

In this appendix we proof the main results of Section 4.4.

Theorem 63 (RESTATED)**.** *Suppose that* $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_1 \leq m$*. Algorithm 8 with*

$$\eta_t = \min\left\{\sqrt{\frac{m\left(1+\log\left(\frac{K}{m}\right)\right)}{16(m\sum_{t=1}^{T}|\mathcal{M}_t|+Kt)}}, \frac{m^2\left(1+\log\left(\frac{K}{m}\right)\right)}{128K(md_{\max}+K)}\right\}\ ,\quad \gamma = \frac{1}{128\sqrt{m}d_{\max}}\ ,$$

guarantees that

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

Proof. We start by verifying the conditions of Corollary 59 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 56 holds. Next, note that R_t as specified in (4.11) does not satisfy Assumption 57(c) because $R_t(\boldsymbol{v}) \leq R_{t-1}(\boldsymbol{v})$. However, by using regularizer $\tilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathcal{W}} R_t(\boldsymbol{v})$, and running the algorithm with this regularizer instead, we can see that Assumption 57(c) is satisfied and, crucially, the algorithm produces the same iterates. Note also that the gradients and Hessians of R_t and $\tilde{R}_t$ are equivalent. We continue the analysis as if the algorithm is run with regularizer $\tilde{R}_t$. By Lemma 81

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

Thus, using Lemma 80 and plugging in γ gives,

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

for all t, $\delta \in [d_{\max}]$, $\boldsymbol{v} \in \mathcal{W}$ and all $\boldsymbol{y} \in \mathbb{R}^K$, which verifies Assumption 57(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 80, showing that Assumptions 57(a) holds for $\kappa = \gamma$.

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

$$\|\boldsymbol{\ell}_\tau\|_{R_t, \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} = \sqrt{\sum_{i=1}^K \boldsymbol{\ell}_\tau(i)^2 \frac{\eta_t \gamma \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, i)^2}{\eta_t + \gamma \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, i)}} \leq \underbrace{\sqrt{\eta_t m}}_{\alpha_t} . \tag{4.14}$$

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

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

where $\mathcal{F}_\tau$ is a filtration over all random events observed by the learner as defined in Section 4.2. Let us consider $\mathbb{E}_{\boldsymbol{a}_\tau}\left[\|\hat{\boldsymbol{\ell}}_\tau\|^2_{R_t, \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)} | \mathcal{F}_\tau\right]$ in isolation:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{a}_\tau}\left[\|\hat{\boldsymbol{\ell}}_\tau\|^2_{R_t, \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)} | \mathcal{F}_\tau\right] &= \mathop{\mathbb{E}}_{\boldsymbol{a}_\tau}\left[\sum_{i=1}^K \left(\frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)}\right)^2 \left(\nabla^2 R_t(\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau))\right)^{-1}(i,i) | \mathcal{F}_\tau\right] \\
&= \sum_{i=1}^K \frac{\boldsymbol{\ell}_\tau(i)^2}{\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)} \frac{\eta_t \gamma \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)^2}{\eta_t + \gamma \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)} \leq \underbrace{\eta_t K}_{\beta_t^2} ,
\end{aligned} \tag{4.15}$$

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

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

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

The last requirement is to show that $\hat{\boldsymbol{\ell}}_\tau$ and $\hat{\boldsymbol{\ell}}_{\tau'}$ are independent for all $\tau, \tau' \in \mathcal{M}_t$ where $\tau' \neq \tau$. Recall that $\hat{\boldsymbol{\ell}}_\tau(i) = \frac{\boldsymbol{a}_\tau(i)\boldsymbol{\ell}_\tau(i)}{\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau, i)}$, for all i. Conditioned on the observed history $\mathcal{F}_t$, the only random element of $\hat{\boldsymbol{\ell}}_\tau$ is $\boldsymbol{a}_\tau \sim \boldsymbol{p}_\tau$. Since $\hat{\boldsymbol{\ell}}'_\tau$ can not have been used in round τ to compute $\boldsymbol{p}_\tau$ (and vice versa) we have that $\hat{\boldsymbol{\ell}}_{\tau'}$ and $\hat{\boldsymbol{\ell}}_\tau$ are independent. We conclude that

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

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

We are now in a position to apply Corollary 59. By using α_t from Equation (4.14) and β_t^2 from Equation (4.15) it follows that for any $\boldsymbol{u} \in \mathcal{W}$

$$\mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{w}_t - \boldsymbol{u})^\top \boldsymbol{\ell}_t\right] \leq R_T(\boldsymbol{u}) + 16 \sum_{t=1}^T \beta_t^2 + 16 \sum_{t=1}^T \alpha_t^2 |\mathcal{M}_t| . \tag{4.16}$$

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 $\boldsymbol{u} = \arg\min_{\boldsymbol{v}\in\tilde{\mathcal{W}}} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{v}$ where $\tilde{\mathcal{W}} = \mathcal{W} \cap \{\boldsymbol{x} \in \mathbb{R}_+ : \forall i \in [K]\ \boldsymbol{x}(i) \geq \theta\}$ 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 $\|\xi\|_\infty \leq 1$.

By Lemma 80, we have

$$R_T(\boldsymbol{u}) \leq \frac{B\left(1 + \log\left(\frac{K}{B}\right)\right)}{\eta_T} + \frac{K \log\left(\frac{1}{\theta}\right)}{\gamma}\,. \tag{4.17}$$

To finish the proof, we start from the regret

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

where we bound $\xi^\top \boldsymbol{\ell}_t \leq K$ in the inequality. We continue by using (4.16)

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

where we first substituted in (4.17), with $\theta = \frac{1}{T}$, $\alpha_t^2 = \eta_t m$ and $\beta_t^2 = \eta_t K$, and applied Lemma 79. Using the last two inequalities and substituting

$$\begin{aligned}\eta_t &= \min\left\{\sqrt{\frac{m\left(1 + \log\left(\frac{K}{m}\right)\right)}{16(m\sum_{t=1}^{T}|\mathcal{M}_t| + Kt)}}, \frac{m^2\left(1 + \log\left(\frac{K}{m}\right)\right)}{128K(md_{\max} + K)}\right\} \\ \gamma &= \frac{1}{128\sqrt{m}d_{\max}}\end{aligned}$$

and doing some simplifications yields

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

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

Theorem 64 (RESTATED)**.** *Suppose that $d_t = d$ for all t and that $m \leq K/2$. Then for any algorithm there exists a sequence of losses such that*

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

Proof. By Audibert, Bubeck, and Lugosi (2014), we have that any algorithm without delay must suffer at least $\Omega(\sqrt{mKT})$ 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, Smola, and Zinkevich (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 m-sets in (Koolen, Warmuth, and Kivinen, 2010, Section 4), which states that any algorithm in the full information setting must suffer at least $\Omega(m\sqrt{T'})$ regret after T' rounds. We take the loss of the first round of the lower bound (Koolen, Warmuth, and Kivinen, 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 $\Omega(dm\sqrt{T/d}) = \Omega(m\sqrt{dT})$ regret, which completes the proof. □

4.B Linear Bandits

Recall that a thrice-differentiable function Ψ is called self-concordant if it is convex and satisfies $|\nabla^3\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h},\boldsymbol{h}]| \leq 2\left(\nabla^2\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h}]\right)^{3/2}$, 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 $|\nabla\Psi(\boldsymbol{v})[\boldsymbol{h}]| \leq \sqrt{\nu\nabla^2\Psi(\boldsymbol{v})[\boldsymbol{h},\boldsymbol{h}]}$. The following property allows us to satisfy the stability condition of the Hessian in Assumption 57(a): for $\boldsymbol{v},\boldsymbol{v}' \in \mathcal{W}$, if $\|\boldsymbol{v}-\boldsymbol{v}'\|^\star_{\Psi,\boldsymbol{v}} < 1$, then

$$\left(1-\|\boldsymbol{v}-\boldsymbol{v}'\|^\star_{\Psi,\boldsymbol{v}}\right)^2\nabla^2\Psi(\boldsymbol{v}) \preceq \Psi(\boldsymbol{v}') \preceq \left(1-\|\boldsymbol{v}-\boldsymbol{v}'\|^\star_{\Psi,\boldsymbol{v}}\right)^{-2}\nabla^2\Psi(\boldsymbol{v}) \ . \tag{4.18}$$

Next, given $\boldsymbol{y} \in \mathcal{W}$ denote by $\pi_{\boldsymbol{y}}(\boldsymbol{x}) = \inf\{z \geq 0 : \boldsymbol{y} + z^{-1}(\boldsymbol{x} - \boldsymbol{y}) \in \mathcal{W}\}$ the Minkowsky function. We denote by $\mathcal{W}_\delta = \{\boldsymbol{v} : \pi_{\boldsymbol{v}^+}(\boldsymbol{v}) \leq (1+\delta)^{-1}\}$, 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 $\boldsymbol{v} \in \mathcal{W}_\delta$

$$\Psi(\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathcal{W}} \Psi(\boldsymbol{v}) \leq \nu \ln\left((1+\delta)\delta^{-1}\right). \tag{4.19}$$

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

Theorem 65 (RESTATED)**.** *Suppose that $T > 100$ and $B \geq 1$. Algorithm 9, run with a ν-self-concordant barrier Ψ and with*

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

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

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

Proof. We start by verifying the assumptions of Corollary 59. Because we are not skipping rounds and have a constant actionset of $\mathcal{W} = \mathrm{Conv}(\mathcal{A})$, we have that Assumption 56 holds. Using $\mathbb{E}[\boldsymbol{v}_t] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_t^\top] = \frac{1}{K}\boldsymbol{I}$ we see that $\mathbb{E}[\hat{\boldsymbol{\ell}}_t] = \boldsymbol{\ell}_t$. For $\tau \leq t$, observe that the distribution of $\hat{\boldsymbol{\ell}}_{\tau'}$ is fully determined given $\mathcal{F}_t$ because $\mathcal{F}_{\tau'} \subseteq \mathcal{F}_t$. Furthermore, since $\hat{\boldsymbol{\ell}}_\tau$ can not be used in round τ' because τ is not available in round t due to the delay, we must have that $\hat{\boldsymbol{\ell}}_{\tau'}$ is independent of $\hat{\boldsymbol{\ell}}_\tau$. Thus,

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

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

We now turn to verifying that Assumption 57 holds. Assumption 57(c) holds by definition of η_t and γ_t. Because $\tilde{\Psi}$ is a self-concordant, if we choose $\kappa = \frac{1}{256BKd_{\max}}$, $\frac{\kappa}{\gamma_t}\tilde{\Psi}$ 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 (4.18), 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 57(a).

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \leq \underbrace{\sqrt{\eta_t}}_{\alpha_t} \cdot$$

We have fulfilled all requirements for Corollary 59. Let $\tilde{\boldsymbol{u}} = \frac{\boldsymbol{u}-\boldsymbol{v}^+}{1+\frac{1}{\sqrt{T}}} + \boldsymbol{v}^+ \in \mathcal{W}_{\frac{1}{\sqrt{T}}}$, with $\boldsymbol{v}^+ = \arg\min_{\boldsymbol{v}\in\mathcal{W}} R_1(\boldsymbol{v})$. By Equation 4.19 and $\mathcal{W} \subseteq \mathcal{B}(B)$ we have that

$$\begin{aligned}R_T(\tilde{\boldsymbol{u}}) - R_1(\boldsymbol{v}^+) &\leq \frac{B^2}{\eta_T} + \frac{1}{\gamma_T}\tilde{\Psi}(\tilde{\boldsymbol{u}}) - \frac{1}{\gamma_1}\tilde{\Psi}(\boldsymbol{v}^+) \\ &\leq \frac{B^2}{\eta_T} + \frac{\nu}{\gamma_T}\log\left(\frac{1+\xi}{\xi}\right) \ , \end{aligned} \tag{4.20}$$

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

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

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 $t = 1, \dots, T$ **do**
  Observe $\boldsymbol{a}_\tau^\top \boldsymbol{\ell}_\tau$ for $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Find loss estimators $\hat{\boldsymbol{\ell}}_\tau = K \boldsymbol{\ell}_\tau^\top \boldsymbol{a}_\tau \left(\nabla^2 \Psi(\boldsymbol{z}_\tau)\right)^{1/2} \boldsymbol{v}_\tau$ for new observations $\tau \in \mathcal{O}_t \setminus \mathcal{O}_{t-1}$.
  Compute $\boldsymbol{z}_t = DN(\Psi_{t-1}, \boldsymbol{z}_{t-1})$
  Play $\boldsymbol{a}_t = \boldsymbol{z}_t + \left(\nabla^2 \Psi(\boldsymbol{z}_t)\right)^{-1/2} \boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is uniformly sampled from the unit sphere.
end for

Algorithm 12: Efficient implementation of delayed FTRL for linear bandits

Thus, by Corollary 59 with $\alpha_t^2 = \eta_t$ and $\beta_t^2 = \gamma_t B^2 K^2$ we obtain

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

where in the second inequality we used $\sum_{t=1}^T \gamma_t \le \frac{\sqrt{\nu T \log(1+\sqrt{T})}}{2BK}$ and $\sum_{t=1}^T \eta_t |\mathcal{M}_t| \le \frac{1}{2} B\sqrt{D}$, both of which follow from Lemma 79, and the last inequality follows from simplifications. □

4.B.1 Efficient Implementation

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

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

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

$$\lambda(\Phi_t, DN(\Phi_t, \boldsymbol{v})) \leq 2\lambda(\Phi_t, \boldsymbol{v})^2 \tag{4.21}$$

$$\|\boldsymbol{v} - \boldsymbol{z}_t^\star\|_{\Phi_{t-1}, \boldsymbol{z}_t^\star} \leq \frac{\lambda(\Phi_{t-1}, \boldsymbol{v})}{1 - 2\lambda(\Phi_{t-1}, \boldsymbol{v})} \qquad \text{if} \quad \lambda(\Phi_{t-1}, \boldsymbol{v}) < \tfrac{1}{2} \tag{4.22}$$

Algorithm 12 is a simple modification of Algorithm 2 in section 9 of Abernethy, Hazan, and Rakhlin (2008). Abernethy, Hazan, and Rakhlin (2008) show that in the non-delayed setting, given the previous iterate, it takes essentially one iteration of the damped Newton method to compute $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)$. 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, Hazan, and Rakhlin (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 chapter. 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, Hazan, and Rakhlin (2008). With Lemma 70 in hand, one can follow the steps provided by Abernethy, Hazan, and Rakhlin (2008) to see that the regret of Algorithm 12 is of the same order as that of Algorithm 9.

Lemma 70. *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

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

Now, with a minor modification of Lemma 62 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

$$\begin{aligned}
&2\gamma^2 \sum_{\tau \in \mathcal{M}_t} \hat{\boldsymbol{\ell}}_\tau^\top \left(\frac{\gamma}{\eta}\boldsymbol{I} + \nabla^2 \Psi(\boldsymbol{z}_t)\right)^{-1} \hat{\boldsymbol{\ell}}_\tau \\
&\le 2\gamma^2 \sum_{\tau \in \mathcal{M}_t} \hat{\boldsymbol{\ell}}_\tau^\top \left(\nabla^2 \Psi(\boldsymbol{z}_t)\right)^{-1} \hat{\boldsymbol{\ell}}_\tau \\
&\le 8\gamma^2 \sum_{\tau \in \mathcal{M}_t} \hat{\boldsymbol{\ell}}_\tau^\top \left(\nabla^2 \Psi(\boldsymbol{z}_\tau)\right)^{-1} \hat{\boldsymbol{\ell}}_\tau && \text{(Equation (4.18))} \\
&\le 8\gamma^2 K^2 |\mathcal{M}_t| \le 8\gamma^2 K^2 d_{\max} \ . && \text{(by def. of } \hat{\boldsymbol{\ell}}_\tau)
\end{aligned}$$

By Equation (4.21) and the induction assumption we have that

$$\lambda(\Phi_{t-1}, \boldsymbol{z}_t)^2 \le 2\lambda(\Phi_{t-1}, \boldsymbol{z}_{t-1})^4 \le 162\gamma^4 K^4 d_{\max}^2 \ . \tag{4.23}$$

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

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

after which we have proven the induction step for the first claim. For the second claim, we start with Equation (4.22) and then the fact that $\lambda(\Phi_{t-1}, \boldsymbol{z}_t)^2 \le \frac{1}{16}$ which follows by Equation (4.23) and the assumption that $\gamma^2 \le \frac{1}{162K^2 d_{\max}}$, then using Equation (4.21) and finally applying the first claim yields

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

□

4.C Adversarial Markov Decision Processes (MDPs)

Lemma 66 (RESTATED)**.** *$\mathcal{W}$ satisfies the following:*

1. *For any $\boldsymbol{q} \in \Delta(\mathscr{M})$, there exists $\tilde{\boldsymbol{q}} \in \mathcal{W}$ such that $\min_{h,s,a,s'} \tilde{\boldsymbol{q}}(h,s,a,s') \ge \frac{1}{T^3 H^2 S^4 A^2}$ and $\|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|_1 \le \frac{2H}{T}$.*
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 $\boldsymbol{q}^{max}(h,s,a) = \max_{\hat{p} \in \mathcal{P}} \boldsymbol{q}^{\pi,\hat{p}}(h,s,a)$. Then, $\|\boldsymbol{q}^\pi - \boldsymbol{v}\|_1 \le \frac{2H}{T}$ and $\|\boldsymbol{q}^{max} - \boldsymbol{v}\|_1 \le \frac{4H^2 S}{T}$.*

Proof. We start by proving the first claim. Define $\tilde{p} : [H] \times \mathcal{S} \times \mathcal{A}_{\mathscr{M}} \to \Delta_{\mathcal{S}}$ by $\tilde{p}(s' \mid h,s,a) = (1 - \frac{1}{THSA})p(s' \mid h,s,a) + \frac{1}{THS^2A}$ and notice that $\tilde{p} \in \mathcal{P}$ since $|p(s' \mid h,s,a) - \tilde{p}(s' \mid h,s,a)| \le \frac{1}{THSA}$. Next, let π_u be the uniformly random policy, and define $\tilde{\boldsymbol{q}} = (1 - \frac{1}{T})\boldsymbol{q} + \frac{1}{T}\boldsymbol{q}^{\pi_u,\tilde{p}}$. It holds that $\tilde{\boldsymbol{q}} \in \mathcal{W}$ because $\mathcal{W}$ is a

convex set. Moreover, notice that $\boldsymbol{q}^{\pi_u,\tilde{p}}(h,s,a,s') \geq \frac{1}{(THS^2A)^2A}$ which implies that $\tilde{\boldsymbol{q}}(h,s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$. Finally,

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

Now we prove the second claim. Define loss function $\tilde{\boldsymbol{\ell}}(h,s,a) = \operatorname{sign}\Big(\boldsymbol{q}^{\pi}(h,s,a) - \boldsymbol{v}(h,s,a)\Big)$ and note that $\|\boldsymbol{q}^{\pi}-\boldsymbol{v}\|_1 = V^{\pi,p,\tilde{\boldsymbol{\ell}}}(1,s_{\text{init}}) - V^{\pi,\hat{p},\tilde{\boldsymbol{\ell}}}(1,s_{\text{init}})$ for some $\hat{p}\in\mathcal{P}$. Combining the value difference lemma (see, e.g, Shani, Efroni, Rosenberg, and Mannor, 2020) with $\|p-\hat{p}\|_\infty \leq \frac{1}{THSA}$ proves that $\|\boldsymbol{q}^{\pi}-\boldsymbol{v}\|_1 \leq \frac{2H}{T}$. Now, let $\hat{p}^{h,s}$ be the transition function that corresponds to $\boldsymbol{u}(h,s)$. We have that, $\|\hat{p}^{h,s}-\hat{p}\|_\infty \leq \|\hat{p}^{h,s}-p\|_\infty + \|p-\hat{p}\|_\infty \leq \frac{2}{THSA}$. Thus, using the same argument as the above, $\|\boldsymbol{u}-\boldsymbol{v}\|_1 \leq \sum_{h,s}\|\boldsymbol{q}^{\pi,\hat{p}^{h,s}}-\boldsymbol{v}\|_1 \leq \frac{4H^2S}{T}$. □

Theorem 67 (RESTATED)**.** *Suppose that $T \geq H$. Algorithm 10 with*

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

guarantees

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

Proof. As in the proof of Theorem 63, the regularizer R_t as specified in (4.12) does not satisfy Assumption 57(c) because we can have $R_t(\boldsymbol{v}) \leq R_{t-1}(\boldsymbol{v})$, but, as argued in the proof of Theorem 63, we can overcome this issue in a relative straightforward manner via the regularizer $\tilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}'\in\mathcal{W}} R_t(\boldsymbol{v}')$, which has no impact on the iterates. We continue by decomposing the regret as

$$\mathcal{R}_T = \sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle = \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle}_{\textsc{Shift-Penalty}},$$

where, by the second property in Lemma 66, Error is bounded by $2H$; and by the first property $\tilde{\boldsymbol{u}} \in \mathcal{W}$ exists such that both Shift-Penalty is bounded by $2H$ and $\min_{h,s,a,s'}\tilde{\boldsymbol{u}}(h,s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$, which we will choose as our comparator to ensure that the regularization is always bounded. For Reg we use Lemma 58

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 80.

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

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

Thus, by definition of γ, Lemma 80 also implies that that

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

Next pick any $t \in [T], \tau \in \mathcal{M}_t$, then

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

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

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

Finally, pick any t, t',

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

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 58 with $\boldsymbol{b}_t = \mathbb{E}[\hat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t \mid \mathcal{F}_t]$, we get

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

Recall that $\min_{h,s,a,s'} \tilde{\boldsymbol{u}}(h,s,a,s') \geq \frac{1}{T^3H^2S^4A^2}$. Thus, using the third fact of Lemma 80 with $b = \frac{1}{T^3H^2S^4A^2}$, $K = HS^2A$ and $B = H$, we conclude

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

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

$$\begin{aligned}&\|\sum_{\tau\in\mathcal{M}_t}(\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} = \left(\sum_{\tau\in\mathcal{M}_t}\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau\right)\nabla^{-2}R_t(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))\left(\sum_{\tau\in\mathcal{M}_t}\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau\right) \\ &= \sum_{\tau\in\mathcal{M}_t}\|(\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} + \sum_{\tau\in\mathcal{M}_t}\sum_{\tau'\in\mathcal{M}_t\setminus\{\tau\}}(\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\nabla^{-2}R_t(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t))(\boldsymbol{\ell}_{\tau'} - \hat{\boldsymbol{\ell}}_{\tau'})\ .\end{aligned} \tag{4.24}$$

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

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

We bound the second term on the right hand side of (4.24) by first applying the law of total expectation,

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

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

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

where the third inequality is by Equation (4.37), the fourth inequality is since $\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau,h,s,a) \le \boldsymbol{q}_\tau^{\max}(h,s,a)$, and the fifth inequality is by Lemma 66. Likewise we can see that

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

These inequalities combined with Jensen's inequality gives

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

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 $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \boldsymbol{q}_t^{\max}$ and $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})$ are all determined

by $\mathcal{G}_t$. Therefore,

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

Now, as in the proof of Lemma 58, $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\gamma}})$. Thus, by Equation (4.37), $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}, h, s, a) \le 2\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a) \le 2\boldsymbol{q}_t^{\max}(h,s,a)$. Therefore,

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

where the last is by article 2 in Lemma 66.

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

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

□

4.D Adversarial MDPs with Unknown Transitions

Theorem 69 (RESTATED). *Algorithm 11 with* $\gamma = \frac{1}{128\sqrt{H}d_{\max}}$, $\eta = \frac{\sqrt{\log(SA)}}{\sqrt{SAT+D}}$, $\xi = \frac{1}{T}$ *and* $T \ge 4$ *guarantees,*

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

Proof. We introduce ς, the good event, where $p \in \mathcal{P}_{j'}$ for all $j' \le j_t$ and the compliment of the good event ς^c. By Lemma 68 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

$$\begin{aligned}\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \\ &= \mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle}_{\textsc{Shift-Penalty}}\right] . \end{aligned} \tag{4.25}$$

Under the good event and by Lemma 66 there exists an $\tilde{\boldsymbol{u}} \in \mathcal{W}_T$ such that Shift-Penalty is bounded by $2H$. That allows us to bound

$$\textsc{Shift-Penalty} = \mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle + \mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \leq 2H + 4\frac{HSA}{T} , \tag{4.26}$$

where we also used that $\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle \leq 1$ 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 84) by

$$\textsc{Error} \lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H^3S^3A\log^2(HSAT^3) + H^3S^2Ad_{max} . \tag{4.27}$$

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 58, our analysis is structured around epochs and to make sure that $\mathcal{W}_t = \mathcal{W}_\tau$ whenever $\tau \in \mathcal{M}_t$, as is required by Assumption 56, 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 outstanding observations, $\mathcal{M}_t \subseteq \Lambda$.

$\mathcal{W}_t \subseteq \mathcal{W}_{t-1}$ holds by the construction of $\mathcal{W}_t$ and $\mathcal{W}_T$ is non-empty under the good event, fulfilling the rest of Assumption 56. Our regularizer fulfils Assumptions 57 under the same caveats as in the previous section. We split Reg into the good and bad event

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

and bound the bad event first. By Lemma 68 the good event ς happens with a probability of at least $1 - \frac{4}{T^2}$ and we have that

$$\mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}})^\top\hat{\boldsymbol{\ell}}_t\right] \leq \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}\sum_{t=1}^{T}\|\hat{\boldsymbol{\ell}}_t\|_1\right] \leq \mathbb{E}\left[\mathbb{I}\{\varsigma^c\}4HT^2\right] \leq 4H , \tag{4.28}$$

where we used Hölders inequality on $(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}})^\top \hat{\boldsymbol{\ell}}_t$, upper bounded $\|\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}\|_\infty \le 1$ and finally upper bound all H non-zero elements of $\hat{\boldsymbol{\ell}}_t$ with $\hat{\boldsymbol{\ell}}_t(h,s,a) \le \frac{1}{\xi} = T$. Under the good event we already showed that Assumption 56 and Assuming 57 are satisfied, both of which are required for Lemma 58. Next we bound $\sqrt{\kappa}\|\hat{\boldsymbol{\ell}}_t\|_{R,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}$ for $\kappa = \gamma$. We have that

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

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

$$\|\boldsymbol{\ell}_\tau\|^2_{R,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \le \eta \sum_{h,s,a} \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t,h,s,a)\boldsymbol{\ell}_t(h,s,a)^2 \le \underbrace{\eta H}_{\alpha^2} . \tag{4.29}$$

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

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\ell}}_\tau\|^2_{R,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)} \,|\, \mathcal{F}_t, \varsigma] &\le \mathbb{E}\left[\eta \sum_{h,s,a} \boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau,h,s,a)\hat{\boldsymbol{\ell}}_\tau(h,s,a)^2 \,\middle|\, \mathcal{F}_t, \varsigma\right] \\ &\le \mathbb{E}\left[\eta \sum_{h,s,a} \frac{\mathbb{I}\{s_{\tau,h}=s, a_{\tau,h}=a\}}{\boldsymbol{q}^{\pi_\tau}(h,s,a)} \,\middle|\, \mathcal{F}_\tau, \varsigma\right] = \underbrace{\eta HSA}_{\beta^2} . \end{aligned} \tag{4.30}$$

As in the proof of Theorem 63, the regularizer R_t as specified in (4.12) does not satisfy Assumption 57(c) because we can have $R_t(\boldsymbol{v}) \le R_{t-1}(\boldsymbol{v})$, but, as argued in the proof of Theorem 63, we can overcome this issue in a relative straightforward manner via the regularizer $\tilde{R}_t(\boldsymbol{v}) = R_t(\boldsymbol{v}) - \min_{\boldsymbol{v}' \in \mathcal{W}} R_t(\boldsymbol{v}')$, which has no impact on the iterates. This already puts us in a position to apply Lemma 58 to find that

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}\{\varsigma\}\sum_{t=1}^T \langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] &\le \mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle \,\middle|\, \varsigma\right] \\ &\le \underbrace{\mathbb{E}\left[\sum_{t\in\Lambda} (\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}})^\top \boldsymbol{\ell}_t \,\middle|\, \varsigma\right]}_{\text{SKIPPED ROUNDS}} + \underbrace{R_T(\tilde{\boldsymbol{u}}) - \min_{v\in\mathcal{W}_1} R_1(\boldsymbol{v})}_{\text{PENALTY}} + \sum_{t\in\bar{\Lambda}} 8\beta_t^2 \\ &\quad - \underbrace{\sum_{t\in\bar{\Lambda}} \mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}) - \tilde{\boldsymbol{u}})^\top \boldsymbol{b}_t \,\middle|\, \varsigma\right]}_{\text{BIAS}} + \sum_{t\in\bar{\Lambda}} \left(8\alpha_t^2|\mathcal{M}_t| + \underbrace{8\alpha_t\, \mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t} (\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \,\middle|\, \varsigma\right]}_{\text{MISSING ESTIMATES}}\right)\end{aligned} \tag{4.31}$$

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:

$$\sum_{t\in\Lambda} (\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}})^\top \boldsymbol{\ell}_t \leq \sum_{t\in\Lambda} \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)^\top \boldsymbol{\ell}_t - \tilde{\boldsymbol{u}}^\top \boldsymbol{\ell}_t \leq d_{\max} H^2 SA \log(T) \ . \tag{4.32}$$

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

We now bound the BIAS term. We know that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}) \in \mathcal{D}_{R_t}(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \frac{1}{2\sqrt{\gamma}})$ when $t \in \bar{\Lambda}$, which is also shown in the proof of Lemma 58. By Lemma 74 we have that $\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}, h, s, a) \leq 2\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a)$. By definition $\boldsymbol{q}_t^{\max}(h,s,a) \geq \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a)$ and thus,

$$\begin{aligned}
&- \mathbb{E}\left[\sum_{t\in\bar{\Lambda}} \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}})^\top \boldsymbol{b}_t\right] \\
&= \mathbb{E}\left[\sum_{t\in\bar{\Lambda}}\sum_{h,s,a} \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}, h, s, a)\boldsymbol{\ell}_t(h,s,a)\left(1 - \frac{\boldsymbol{q}^{\pi_t}(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\right)\right] \\
&= \mathbb{E}\left[\sum_{t\in\bar{\Lambda}}\sum_{h,s,a} \frac{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}, h, s, a)\boldsymbol{\ell}_t(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\left(\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a) + \xi\right)\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{t\in\bar{\Lambda}}\sum_{h,s,a} \frac{\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t, h, s, a)\boldsymbol{\ell}_t(h,s,a)}{\boldsymbol{q}_t^{\max}(h,s,a) + \xi}\left(|\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a)| + \xi\right)\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{t\in\bar{\Lambda}}\left(\sum_{h,s,a} |\boldsymbol{q}_t^{\max}(h,s,a) - \boldsymbol{q}^{\pi_t}(h,s,a)| + \xi HSA\right)\right] \\
&\lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H^3S^3A\log^2(HSAT^3) + H^3S^2Ad_{max} + HSA \ ,
\end{aligned} \tag{4.33}$$

where the last inequality is due to $\xi = 1/T$ and Lemma 84, where we take an expectation over the event that the inequality in Lemma 84 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 $\mathbb{E}[\tilde{\boldsymbol{u}}^\top \boldsymbol{b}_t \mid \varsigma] \leq 0$, 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 59 out of the box because $\hat{\boldsymbol{\ell}}_t$ is a biased estimator, so we add and subtract the bias $\boldsymbol{b}_t$ and use the triangle inequality to find

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

Now we recognize that we are not using any information from rounds that we have not seen yet and thus $\hat{\boldsymbol{\ell}}_\tau$ and $\hat{\boldsymbol{\ell}}_{\tau'}$ are independent if $\tau, \tau' \in \mathcal{M}_t$. Furthermore, $\hat{\boldsymbol{\ell}}_t$ is

an unbiased estimator of $\boldsymbol{\ell}_t + \boldsymbol{b}_t$, which allows us to use the exact same arguments as in Corollary 59 for the first term to find

$$\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}(\boldsymbol{\ell}_\tau + \boldsymbol{b}_\tau - \hat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \middle| \varsigma\right] \leq \sqrt{\sum_{\tau\in\mathcal{M}_t}\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{\ell}}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right] - \mathbb{E}\left[\left\|\boldsymbol{\ell}_\tau + \boldsymbol{b}_\tau\right\|^2_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)}\right]\right)}$$
$$\leq \sqrt{4|\mathcal{M}_t|\beta_t^2}$$

For the second term we start with the triangle inequality and Lemma 74,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}\boldsymbol{b}_\tau\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \middle| \varsigma\right] &\leq \mathbb{E}\left[\sum_{\tau\in\mathcal{M}_t}\|\boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \middle| \varsigma\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{\tau\in\mathcal{M}_t}\|\boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)} \middle| \varsigma\right] \\
&\overset{(a)}{\leq} 2\,\mathbb{E}\left[\sum_{\tau\in\mathcal{M}_t}\left(\|\boldsymbol{\ell}_\tau + \boldsymbol{b}_\tau\|_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)} + \|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)}\right) \middle| \varsigma\right] \\
&\overset{(b)}{\leq} 2\,\mathbb{E}\left[\sum_{\tau\in\mathcal{M}_t}\left(2\|\boldsymbol{\ell}_\tau\|_{R_t,\boldsymbol{w}_\tau(\hat{\boldsymbol{L}}_\tau)}\right) \middle| \varsigma\right] \leq 4|\mathcal{M}_t|\alpha_t\ ,
\end{aligned}$$

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

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

$$\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{M}_t}(\boldsymbol{\ell}_\tau - \hat{\boldsymbol{\ell}}_\tau)\right\|_{R_t,\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t)} \middle| \varsigma\right] \leq \sqrt{4|\mathcal{M}_t|\beta_t^2} + 4|\mathcal{M}_t|\alpha_t\ . \tag{4.34}$$

The last thing to bound is the PENALTY term. Using the third fact of Lemma 80 with $b = \frac{1}{T^3H^2S^4A^2}$, $K = HS^2A$ and $B = H$, we conclude

$$\begin{aligned}
\text{PENALTY} &\leq \frac{H\left(1+\log\left(S^2A\right)\right)}{\eta} + \frac{HS^2A\log\left(T^3H^2S^4A^2\right)}{\gamma} \\
&\leq \frac{4H\log\left(SA\right)}{\eta} + \frac{4HS^2A\log\left(HSAT\right)}{\gamma}.
\end{aligned} \tag{4.35}$$

Putting things together for the REG term gives

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

where we used equations (4.28), (4.31), (4.32), (4.34), (4.35) in the first inequality and plugged in the values of α and β we found in equations (4.29) and (4.30) and also used that $\sqrt{ab} \le \frac{1}{2}(a+b)$ for $a, b > 0$. We plug in the learning rates to find

$$
\begin{aligned}
\textsc{Reg} &\lesssim \frac{H\log(SA)}{\eta} + \frac{HS^2A\log(HSAT)}{\gamma} + \eta HSAT \\
&\quad + \eta HD + d_{\max}H^2SA\log(T) \\
&\quad - \sum_{t\in\bar{\Lambda}}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + H \\
&\lesssim H\sqrt{SAT\log(SA)} + H\sqrt{D\log(SA)} \\
&\quad + d_{\max}\sqrt{H}S^2A\log(HSAT) + d_{\max}H^2SA\log(T) \\
&\quad - \sum_{t\in\bar{\Lambda}}\mathbb{E}\left[(\boldsymbol{w}_t(\hat{\boldsymbol{L}}_t^{\mathcal{M}}) - \boldsymbol{u})^\top \boldsymbol{b}_t \,|\, \varsigma\right] + H \ . \qquad (4.36)
\end{aligned}
$$

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

$$\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}^{\pi_t} - \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t), \boldsymbol{\ell}_t\rangle + \sum_{t=1}^{T}\langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle + \sum_{t=1}^{T}\langle \tilde{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Eqn (4.25))} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{w}_t(\hat{\boldsymbol{L}}_t) - \tilde{\boldsymbol{u}}, \boldsymbol{\ell}_t\rangle\right] + \sqrt{H^4S^2AT\log(HSAT^3)} \\
&\quad + H^3S^3A\log^2(HSAT^3) + d_{\max}H^3S^2A \qquad\qquad \text{(Eqns (4.26) and (4.27))} \\
&\lesssim \sqrt{H^4S^2AT\log(HSAT^3)} + H\sqrt{D\log(SA)} + d_{\max}H^3S^2A \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Eqns (4.36) and (4.33))} \\
&\quad + H^3S^3A\log^2(HSAT^3) + d_{\max}\sqrt{H}S^2A\log(HSAT) + d_{\max}H^2SA\log(T)\ ,
\end{aligned}$$

which concludes the proof. □

4.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, Zhou, Chen, Bambos, and Blanchet, 2019; Lancewicki, Rosenberg, and Mansour, 2022b).

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

Algorithm 13: Doubling procedure

Theorem 71. *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 13 with unknown d_{max} guarantees regret,*

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

where $M = \max_{t\in[T],\boldsymbol{a},\tilde{\boldsymbol{a}}\in\mathcal{A}}(\boldsymbol{a} - \tilde{\boldsymbol{a}})^\top \boldsymbol{\ell}_t$ is the maximal regret per round (e.g., in Section 4.6, $M \leq H$).

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 $\tilde{\mathcal{T}}_e = \{t \in \mathcal{T}_e : d_t \leq 2^e\}$ be the indices of epoch e with delay $\leq 2^e$. The regret in rounds $t \in \tilde{\mathcal{T}}_e$ 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 $\mathcal{T}_e \backslash \tilde{\mathcal{T}}_e$ is at most $M d_{max}$ since $|\mathcal{T}_e \backslash \tilde{\mathcal{T}}_e| \leq d_{max}$. Thus, the total regret in epoch e is at most,

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

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,

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

□

4.F Auxiliary Lemmas

Lemma 72. *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 73 (Be-The-Leader Lemma)**.** *Let $\tilde{\boldsymbol{w}}_t(\hat{\boldsymbol{L}}_t^\star) = \arg\min_{\boldsymbol{w} \in \mathcal{W}_t} \boldsymbol{w}^\top \hat{\boldsymbol{L}}_t^\star + R_t(\boldsymbol{w})$. 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*

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

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

$$\sum_{t\in\bar{\Lambda}\cap[T-1]} \hat{\boldsymbol{\ell}}_t^\top \tilde{\boldsymbol{w}}_t(\hat{\boldsymbol{L}}_t^\star) + R_1(\boldsymbol{w}_1(\hat{\boldsymbol{L}}_1)) \le \sum_{t\in\bar{\Lambda}\cap[T-1]} \hat{\boldsymbol{\ell}}_t^\top \boldsymbol{v} + R_{T-1}(\boldsymbol{v})$$

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

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

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

Lemma 74. *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 $i \in [n]$,*

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

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

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

On the other hand,

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

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

Lemma 75. *Let $a, b \in \mathbb{R}_+$ such that $a \ge b$, then*

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

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 76. $\log(x)^2 \leq \frac{1}{x}$ *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 rearrange to $-\sqrt{x}\log(x) \leq 1$. We maximize 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 77. *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 78 (Part of Lemma 14 from Gaillard, Stoltz, and Erven (2014))**.** *Let* $a_1, \ldots, a_{\mathcal{M}} \in \mathbb{R}_+$ *and call* $s_i = a_1 + \ldots + a_i$. *Let* $f : (0, \infty) \to [0, \infty]$ *be a non-increasing function. Then*

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

Proof.

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

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 79.

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}\,, \qquad \sum_{t=1}^{T} \frac{|\mathcal{M}_t|}{\sqrt{\sum_{t=1}^{T} |\mathcal{M}_\tau|}} \leq 2\sqrt{D}\,, \qquad \sum_{t=1}^{T} t^{-\frac{1}{4}} \leq \frac{4}{3} T^{\frac{3}{4}}\,.$$

Proof. By Lemma 78 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 $\mathcal{M}_t$ by $\tilde{m}_t$, where we used that $|\mathcal{M}_t| \leq |\tilde{m}_t|$ which holds with probability one, then by Lemma 78 with $a_i = |\tilde{m}_i|$ and $f(x) = \frac{1}{\sqrt{x}}$

$$\sum_{t=1}^{T} \frac{|\mathcal{M}_t|}{\sqrt{\sum_{t=1}^{T} |\mathcal{M}_\tau|}} \leq \sum_{t=1}^{T} \frac{|\tilde{m}_t|}{\sqrt{\sum_{t=1}^{T} |\tilde{m}_\tau|}} \leq \int_1^D \frac{1}{\sqrt{x}}\,\mathrm{d}x \leq 2\sqrt{D}\,.$$

One last time by Lemma 78 with $a_1, \ldots, a_T = 1$ and $f(x) = x^{-\frac{1}{4}}$

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

□

Lemma 80. *Let* $\mathcal{V}(b) \subseteq \{\boldsymbol{x} : 0 \leq b \leq \boldsymbol{x}(i) \leq 1\}$ *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* $\eta_t \geq \eta_{t+1}$*, 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* $\sqrt{\eta_{t+\delta}} \left(\frac{1}{\eta_{t+\delta}} - \frac{1}{\eta_t} \right) \leq \frac{\sqrt{\lambda}}{\sqrt{K}}$ *for a given* t *and* $\delta \in [d_{\max}]$*, then*

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

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

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

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 $j \neq i$. Now, we have that

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

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

$$\frac{1}{2}\boldsymbol{v}(i) \leq \boldsymbol{v}'(i) \leq \frac{3}{2}\boldsymbol{v}'(i)\,. \tag{4.37}$$

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 76.

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

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

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

Lemma 81. *Let* $\eta_t = \min\left\{a, \frac{1}{\sqrt{bt+c\sum_{t=1}^{T}|\mathcal{M}_\tau|}}\right\}$ *for some* $a, b, c \in \mathbb{R}_+$. *If* $a \leq \frac{d}{bd_{\max}+cd_{\max}^2}$ *for some* $d \in \mathbb{R}$, *then*

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

Proof. We start by showing that

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

where we used our assumption on η_t together with $\max\{x, y\} - \max\{x, z\} \leq y - z$ (Lemma 77) in the first inequality, $\sqrt{a} - \sqrt{b} \leq \sqrt{a-b}$ (Lemma 75) in the second inequality and the fact that $\delta \leq d_{\max}$ and $|\mathcal{M}_t| \leq d_{\max}$ 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 82. *Let* $\eta_t = \min\left\{a, \frac{1}{\sqrt{bt+c\sum_{t=1}^{T}|\mathcal{M}_\tau|}}\right\}$ *for some* $a, b, c \in \mathbb{R}_+$. *If* $a \leq \frac{d}{\sqrt{bd_{\max}+cd_{\max}^2}}$ *for some* $d \in \mathbb{R}$, *then*

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

Proof. We start by showing that

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

where we used our assumption on η_t together with $\max\{x,y\} - \max\{x,z\} \le y - z$ (Lemma 77) in the first inequality, $\sqrt{a} - \sqrt{b} \le \sqrt{a-b}$ (Lemma 75) in the second inequality and the fact that $\delta \le d_{\max}$ and $|\mathcal{M}_t| \le d_{\max}$ 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 83 (Lemma D.11 of Jin, Lancewicki, Luo, Mansour, and Rosenberg, 2022; see also Lemma 4 of Jin, Jin, Luo, Sra, and Yu, 2020)**.** *With probability* $1-\delta$*, for any collection of transition functions* $\{p^s_{t,h}\}_{s\in\mathcal{S}}$ *such that* $p^s_t \in \hat{P}_j$

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_{\mathscr{M}}}\left|q_t^{p_t^s,\pi_t}(h,s,a) - q_t^{\pi_t}(h,s,a)\right| \lesssim H\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S},a\in\mathcal{A}_{\mathscr{M}}}\epsilon_t(h,s,a)q^{\pi_t}(h,s,a)$$
$$+ HS\sum_{t=1}^{T}\sum_{1\le h<\tilde{h}\le H}\sum_{s\in\mathcal{S},a\in\mathcal{A}_{\mathscr{M}},s'\in\mathcal{S}}\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}_{\mathscr{M}}}\epsilon_t(s' \mid h,s,a)q^{\pi_t}(h,s,a)$$
$$\cdot\min\left\{2, \sum_{\tilde{s}'\in\mathcal{S}}\epsilon_t(\tilde{s}' \mid \tilde{h},\tilde{s},\tilde{a})\right\} q^{\pi_t}(\tilde{h},\tilde{s},\tilde{a} \mid s'; h+1) + H^3S^2Ad_{max} \tag{4.38}$$

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

Lemma 84 (Lemma D.12 of Jin, Lancewicki, Luo, Mansour, and Rosenberg, 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* $p_t^{h,s} \in \hat{P}_j$

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

Proof. Following the exact same steps as in the proof of Lemma E.5 in Lancewicki, Rosenberg, and Mansour, 2022a, with probability of at least $1-\delta$, the first term in (4.38) can be bounded by,

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s\in\mathcal{S},a\in\mathcal{A}_{\mathscr{M}}}\epsilon_t(h,s,a)q^{\pi_t}(h,s,a) \lesssim \sqrt{S\log(HSAT^3)}\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{\sqrt{N_{j(t)}(h,s,a)\vee 1}}$$
$$+ S\log(HSAT^3)\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{N_{j(t)}(h,s,a)\vee 1} + HS\log^2(HSAT^3) \tag{4.39}$$

Similarly, the second summation (4.38) is bounded by,

$$HS\log^2(HSAT^3)\sum_{t,h,s,a}\frac{\mathbb{I}\{s_{t,h}=s,a_{t,h}=a\}}{N_{j(t)}(h,s,a)\vee 1} \tag{4.40}$$

Finally, (4.39) and (4.40) 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, Jin, Luo, Sra, and Yu, 2020. □

4.G Further Results of the Experiments

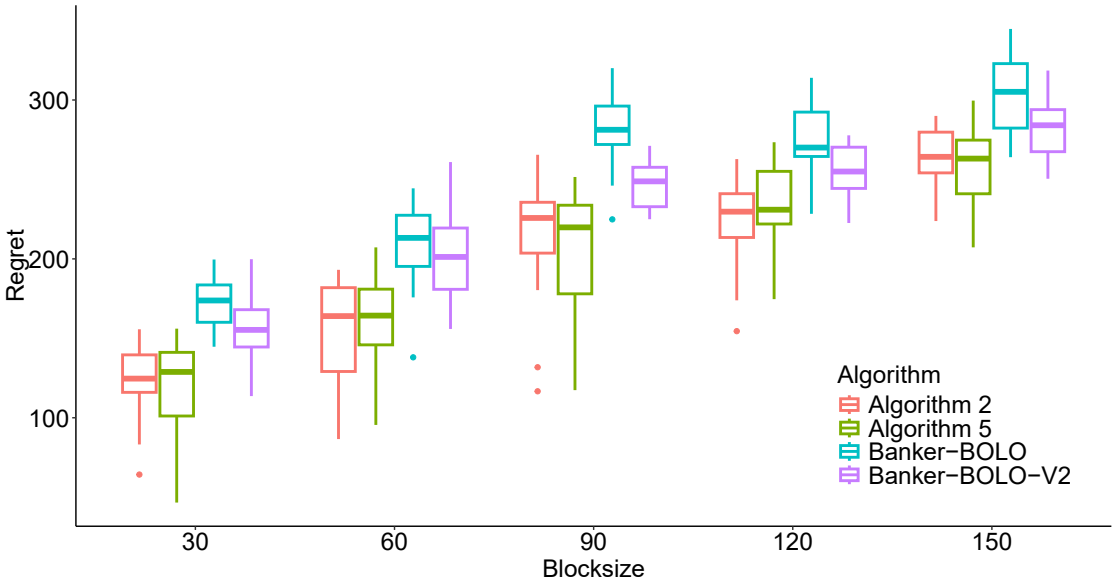


Figure 4.G.1: Boxplot of the regret over 20 repetitions over $T = 10000$ rounds with $K = 10$.

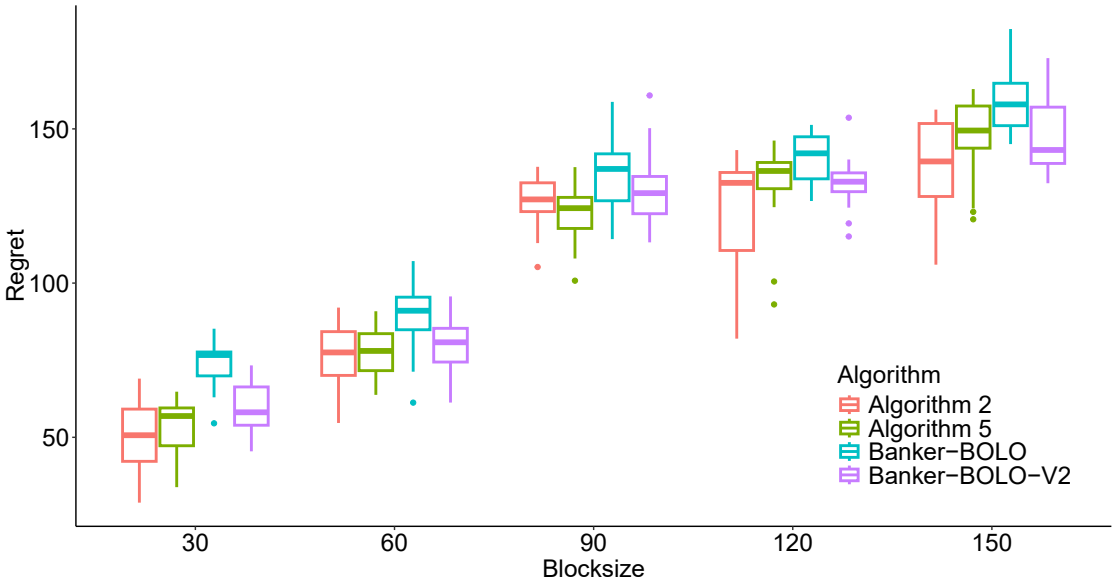


Figure 4.G.2: Boxplot of the regret over 20 repetitions over $T = 10000$ rounds with $K = 40$.

Bibliography

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

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

[3] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. "Interior-point methods for full-information and bandit online learning". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4164–4175.

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

[5] Shubhada Agrawal, Sandeep K. Juneja, and Wouter M. Koolen. "Regret Minimization in Heavy-Tailed Bandits". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 26–62. URL: https://proceedings.mlr.press/v134/agrawal21a.html.

[6] Reda Alami. "Bayesian Change-Point Detection for Bandit Feedback in Non-stationary Environments". In: *Proceedings of The 14th Asian Conference on Machine Learning*. Ed. by Emtiyaz Khan and Mehmet Gonen. Vol. 189. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 17–31. URL: https://proceedings.mlr.press/v189/alami23a.html.

[7] Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. "Hannan consistency in on-line learning in case of unbounded losses under partial monitoring". In: *Proceedings of the 17th International Conference on Algorithmic Learning Theory*. ALT'06. Barcelona, Spain: Springer-Verlag, 2006, pp. 229–243. ISBN: 3540466495. DOI: 10.1007/11894841_20. URL: https://doi.org/10.1007/11894841_20.

[8] Jean-Yves Audibert and Sébastien Bubeck. "Minimax policies for adversarial and stochastic bandits". In: *Proceedings of the 22th annual conference on learning theory*. Montreal, Canada, June 2009, pp. 217–226. URL: https://enpc.hal.science/hal-00834882.

[9] Jean-Yves Audibert and Sébastien Bubeck. "Regret Bounds and Minimax Policies under Partial Monitoring". In: *Journal of Machine Learning Research* 11.94 (2010), pp. 2785–2836. URL: http://jmlr.org/papers/v11/audibert10a.html.

[10] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. "Regret in online combinatorial optimization". In: *Mathematics of Operations Research* 39.1 (2014), pp. 31–45.

[11] Peter Auer. "Using Confidence Bounds for Exploitation-Exploration Tradeoffs". In: *Journal of Machine Learning Research* 3 (2002), pp. 397–422.

[12] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning Journal* 47.2-3 (2002), pp. 235–256.

[13] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. "The Nonstochastic Multiarmed Bandit Problem". In: *SIAM Journal on Computing* 32.1 (2002), pp. 48–77. DOI: 10.1137/S0097539701398375.

[14] Orly Avner, Shie Mannor, and Ohad Shamir. "Decoupling Exploration and Exploitation in Multi-Armed Bandits". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. ICML '12. Edinburgh, Scotland, GB: Omnipress, July 2012, pp. 409–416. ISBN: 978-1-4503-1285-1.

[15] 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*. 2004, pp. 45–53.

[16] Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. "Online Exp3 learning in adversarial bandits with delayed feedback". In: *Advances in Neural Information Processing Systems*. 2019, pp. 11349–11358.

[17] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. "Survey on Applications of Multi-Armed and Contextual Bandits". In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. 2020, pp. 1–8. DOI: 10.1109/CEC48606.2020.9185782.

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

[19] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. "Bandits With Heavy Tail". In: *Information Theory, IEEE Transactions on* 59 (Sept. 2012). DOI: 10.1109/TIT.2013.2277869.

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

[21] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. "Delay and cooperation in nonstochastic bandits". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 613–650.

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

[23] Nicolò Cesa-Bianchi and Gábor Lugosi. "Combinatorial bandits". In: *Journal of Computer and System Sciences* 78.5 (2012). JCSS Special Issue: Cloud Computing 2011, pp. 1404–1422. ISSN: 0022-0000. DOI: https://doi.org/10.1016/j.jcss.2012.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0022000012000219.

[24] Nicolò Cesa-Bianchi and Gábor Lugosi. "Combinatorial bandits". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1404–1422.

[25] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[26] Wei Chen, Yajun Wang, and Yang Yuan. "Combinatorial Multi-Armed Bandit: General Framework and Applications". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 151–159. URL: https://proceedings.mlr.press/v28/chen13a.html.

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

[28] Alon Cohen, Tamir Hazan, and Tomer Koren. "Online Learning with Feedback Graphs Without the Graphs". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 811–819. URL: https://proceedings.mlr.press/v48/cohena16.html.

[29] Alon Cohen, Tamir Hazan, and Tomer Koren. "Tight bounds for bandit combinatorial optimization". In: *Conference on Learning Theory*. 2017, pp. 629–642.

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

[31] Corinna Cortes, Giulia Desalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. "Online Learning with Sleeping Experts and Feedback Graphs". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 1370–1378. URL: https://proceedings.mlr.press/v97/cortes19a.html.

[32] Thomas M Cover. "Universal portfolios". In: *Mathematical finance* 1.1 (1991), pp. 1–29.

[33] 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*. Vol. 35. Curran Associates, Inc., 2022, pp. 11437–11449.

[34] Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. "Refined Regret for Adversarial MDPs with Linear Function Approximation". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 6726–6759. URL: https://proceedings.mlr.press/v202/dai23b.html.

[35] Varsha Dani and Thomas P Hayes. "Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary". In: *SODA*. Vol. 6. 2006, pp. 937–943.

[36] Varsha Dani, Sham M Kakade, and Thomas Hayes. "The price of bandit information for online optimization". In: *Advances in Neural Information Processing Systems* 20 (2007).

[37] Rémy Degenne, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet. "Bridging the gap between regret minimization and best arm identification, with application to A/B tests". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 1988–1996. URL: https://proceedings.mlr.press/v89/degenne19a.html.

[38] David Delande, Patricia Stolf, Raphaël Feraud, Jean-Marc Pierson, and Andre Bottaro. "Horizontal Scaling in Cloud Using Contextual Bandits". In: Sept. 2021.

[39] Travis Dick, András György, and Csaba Szepesvari. "Online learning in Markov decision processes with changing cost sequences". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 512–520.

[40] 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*. 2011, pp. 169–178.

[41] Stephen G. Eick. "The two-armed bandit with delayed responses". In: *The Annals of Statistics* (1988).

[42] Albert Einstein. "Die Grundlage der allgemeinen Relativitätstheorie". In: *Annalen der Physik* 354.7 (Jan. 1916), pp. 769–822. DOI: 10.1002/andp.19163540702.

[43] Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. "Learning on the Edge: Online Learning with Stochastic Feedback Graphs". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 34776–34788. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/e0e956681b04ac126679e8c7ddPaper-Conference.pdf.

[44] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. "Online Markov Decision Processes". In: *Mathematics of Operations Research* 34.3 (2009), pp. 726–736.

[45] 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*. 2021, pp. 3363–3373.

[46] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

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

[48] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. "Combinatorial Network Optimization With Unknown Variables: Multi-Armed Bandits With Linear Rewards and Individual Observations". In: *IEEE/ACM Transactions on Networking* 20.5 (2012), pp. 1466–1478. DOI: 10.1109/TNET.2011.2181864.

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

[50] Gerald Goertzel, U.S. Atomic Energy Commission, and Oak Ridge National Laboratory. *Quota Sampling and Importance Functions in Stochastic Solution of Particle Problems*. AECD. U.S. Atomic Energy Commission, Technical Information Division, 1950. URL: https://books.google.nl/books?id=Su1EGYGagoAC.

[51] Geoffrey J. Gordon. "Regret bounds for prediction problems". In: *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. COLT '99. Santa Cruz, California, USA: Association for Computing Machinery, 1999, pp. 29–40. ISBN: 1581131674. DOI: 10.1145/307400.307410. URL: https://doi.org/10.1145/307400.307410.

[52] András György and Pooria Joulani. "Adapting to delays and data in adversarial multi-armed bandits". In: *International Conference on Machine Learning*. 2021, pp. 3988–3997.

[53] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. "The On-Line Shortest Path Problem Under Partial Monitoring". In: *Journal of Machine Learning Research* 8.79 (2007), pp. 2369–2403. URL: http://jmlr.org/papers/v8/gyoergy07a.html.

[54] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. "The On-Line Shortest Path Problem Under Partial Monitoring." In: *Journal of Machine Learning Research* 8.10 (2007).

[55] J. Hannan. "Approximation to Bayes risk in repeated play". In: *Contributions to the Theory of Games* III (1957).

[56] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[57] Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sebag. "Multi-armed Bandit, Dynamic Environments and Meta-Bandits". working paper or preprint. Nov. 2006. URL: https://hal.science/hal-00113668.

[58] Elad Hazan and Zohar Karnin. "Volumetric spanners: an efficient exploration basis for learning". In: *Journal of Machine Learning Research* (2016).

[59] Dirk Van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. "A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs". In: *Conference on Learning Theory*. 2023, pp. 1285–1321.

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

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

[62] 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*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 13814–13844.

[63] 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*. 2020, pp. 4872–4883.

[64] 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*. 2020, pp. 2028–2038.

[65] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning." In: *Journal of Machine Learning Research* 11.4 (2010).

[66] 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*. 2020, pp. 4860–4869.

[67] 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*. Vol. 35. Curran Associates, Inc., 2022, pp. 33469–33481.

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

[69] Pooria Joulani, András György, and Csaba Szepesvári. "Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms". In: *AAAI Conference on Artificial Intelligence*. 2016.

[70] Pooria Joulani, András György, and Csaba Szepesvári. "Online learning under delayed feedback". In: *International Conference on Machine Learning*. 2013, pp. 1453–1461.

[71] Herman Kahn and Theodore E Harris. "Estimation of particle transmission by random sampling". In: *National Bureau of Standards applied mathematics series* 12 (1951), pp. 27–30.

[72] Satyen Kale, Chansoo Lee, and David Pal. "Hardness of Online Sleeping Combinatorial Optimization Problems". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/184260348236f9554fe9375772Paper.pdf.

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

[74] Satyen Kale, Lev Reyzin, and Robert E Schapire. "Non-stochastic bandit slate problems". In: *Advances in Neural Information Processing Systems* 23 (2010).

[75] Varun Kanade and Thomas Steinke. "Learning Hurdles for Sleeping Experts". In: *ACM Trans. Comput. Theory* 6.3 (July 2014). ISSN: 1942-3454. DOI: 10.1145/2505983. URL: https://doi.org/10.1145/2505983.

[76] Konstantinos V Katsikopoulos and Sascha E Engelbrecht. "Markov decision processes with delays and asynchronous cost collection". In: *IEEE transactions on automatic control* 48.4 (2003), pp. 568–574.

[77] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. "Regret bounds for sleeping experts and bandits". In: *Machine Learning* 80.2 (Sept. 2010), pp. 245–272. ISSN: 1573-0565. DOI: 10.1007/s10994-010-5178-7. URL: https://doi.org/10.1007/s10994-010-5178-7.

[78] Teun Kloek and Herman van Dijk. "Bayesian estimates of equation system parameters, An application of integration by Monte Carlo". In: *Econometrica* (Jan. 1978). URL: http://hdl.handle.net/1765/11224.

[79] Donald Ervin Knuth. *The Art of Computer Programming: Volume 1: Fundamental Algorithms (3rd ed.)* Addison-Wesley, 1997.

[80] Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. "Efficient learning by implicit exploration in bandit problems with side observations". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/8169cf3dc05090c7774c8dc38317c43d-Paper.pdf.

[81] Wouter M Koolen, Manfred K Warmuth, and Jyrki Kivinen. "Hedging Structured Concepts." In: *Conference on Learning Theory*. 2010, pp. 93–105.

[82] Akshay Krishnamurthy, Alekh Agarwal, and Miro Dudik. "Contextual semibandits via supervised learning oracles". In: *Advances in Neural Information Processing Systems* 29 (2016).

[83] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. "Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, Sept. 2015, pp. 535–543. URL: https://proceedings.mlr.press/v38/kveton15.html.

[84] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. "Cooperative Online Learning in Stochastic and Adversarial MDPs". In: *International Conference on Machine Learning*. 2022, pp. 11918–11968.

[85] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. "Learning adversarial Markov decision processes with delayed feedback". In: *AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 7281–7289.

[86] 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*. PMLR. 2023, pp. 18482–18534.

[87] Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. "Stochastic Multi-Armed Bandits with Unrestricted Delay Distributions". In: *International Conference on Machine Learning*. 2021, pp. 5969–5978.

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

[89] 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.

[90] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.

[91] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. "A contextual-bandit approach to personalized news article recommendation". In: *International World Wide Web Conference*. 2010, pp. 661–670.

[92] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: Association for Computing Machinery, 2011, pp. 297–306. ISBN: 9781450304931. DOI: 10.1145/1935826.1935878. URL: https://doi.org/10.1145/1935826.1935878.

[93] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization". In: *Journal of Machine Learning Research* 18.185 (2018), pp. 1–52. URL: http://jmlr.org/papers/v18/16-558.html.

[94] Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. "Contextual combinatorial cascading bandits". In: *International Conference on Machine Learning*. 2016, pp. 1245–1253.

[95] N. Littlestone and M.K. Warmuth. "The Weighted Majority Algorithm". In: *Information and Computation* 108.2 (1994), pp. 212–261. ISSN: 0890-5401. DOI: https://doi.org/10.1006/inco.1994.1009. URL: https://www.sciencedirect.com/science/article/pii/S0890540184710091.

[96] 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)*. Hollywood, United States, May 2015, pp. 67–72. URL: https://hal.science/hal-04077707.

[97] 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.

[98] Shie Mannor and Ohad Shamir. "From Bandits to Experts: On the Value of Side-Observations". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/e1e32e235eee1f970470a3a6658dfdd5-Paper.pdf.

[99] 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*. Vol. 35. Curran Associates, Inc., 2022, pp. 11752–11762.

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

[101] Joseph Mellor and Jonathan Shapiro. "Thompson Sampling in Switching Environments with Bayesian Online Change Detection". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Carlos M. Carvalho and Pradeep Ravikumar. Vol. 31. Proceedings of Machine Learning Research. Scottsdale, Arizona, USA: PMLR, 29 Apr–01 May 2013, pp. 442–450. URL: https://proceedings.mlr.press/v31/mellor13a.html.

[102] Arkadi Nemirovski. "Interior point polynomial time methods in convex programming". In: *Lecture notes* 42.16 (2004), pp. 3215–3224.

[103] Arkadi S Nemirovski and Michael J Todd. "Interior-point methods for optimization". In: *Acta Numerica* 17 (2008), pp. 191–234.

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

[105] Gergely Neu. "Explore no more: Improved high-probability regret bounds for non-stochastic bandits". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/e5a4d6bf330f23a8707bb0d6001dfbe8-Paper.pdf.

[106] Gergely Neu and Gábor Bartók. "An Efficient Algorithm for Learning with Semi-Bandit Feedback". In: *International Conference on Algorithmic Learning Theory*. 2013, pp. 234–248.

[107] Gergely Neu and Gábor Bartók. "Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits". In: *Journal of Machine Learning Research* 17 (154 2016), pp. 1–21.

[108] Gergely Neu, András György, Csaba Szepesvári, and András Antos. "Online Markov Decision Processes Under Bandit Feedback". In: *IEEE Trans. Automat. Contr.* 59.3 (2014), pp. 676–691.

[109] Gergely Neu, András György, Csaba Szepesvári, and András Antos. "Online Markov Decision Processes under bandit feedback". In: *Advances in Neural Information Processing Systems*. 2010.

[110] Gergely Neu and Julia Olkhovskaya. "Efficient and robust algorithms for adversarial linear contextual bandits". In: *Conference on Learning Theory*. Sept. 2020, pp. 3049–3068. URL: https://proceedings.mlr.press/v125/neu20b.html.

[111] Gergely Neu and Michal Valko. "Online combinatorial optimization with stochastic decision sets and adversarial losses". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/06da2cfb2088f776d522b5cdafe677ab-Paper.pdf.

[112] He Ni, Hao Xu, Dan Ma, and Jun Fan. "Contextual combinatorial bandit on portfolio management". In: *Expert Systems with Applications* 221 (2023), p. 119677.

[113] Francesco Orabona. "A Modern Introduction to Online Learning". In: *CoRR* abs/1912.13213 (2019). arXiv: 1912.13213. URL: http://arxiv.org/abs/1912.13213.

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

[115] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. "Contextual combinatorial bandit and its application on diversified online recommendation". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. 2014, pp. 461–469.

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

[117] Alexander Rakhlin and Karthik Sridharan. "Online learning with predictable sequences". In: *Conference on Learning Theory*. 2013, pp. 993–1019.

[118] Aviv Rosenberg and Yishay Mansour. "Online convex optimization in adversarial Markov decision processes". In: *International Conference on Machine Learning*. 2019, pp. 5478–5486.

[119] Aviv Rosenberg and Yishay Mansour. "Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function". In: *Advances in Neural Information Processing Systems*. 2019, pp. 2209–2218.

[120] Aviv Rosenberg and Yishay Mansour. "Stochastic Shortest Path with Adversarially Changing Costs". In: *International Joint Conference on Artificial Intelligence*. Ed. by Zhi-Hua Zhou. 2021, pp. 2936–2942.

[121] Chloé Rouyer and Yevgeny Seldin. "Tsallis-INF for Decoupled Exploration and Exploitation in Multi-armed Bandits". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 3227–3249. URL: https://proceedings.mlr.press/v125/rouyer20a.html.

[122] Aadirupa Saha and Pierre Gaillard. "Dueling Bandits with Adversarial Sleeping". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 27761–27771. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/e97ee2054defb209c35fe4dc94Paper.pdf.

[123] Aadirupa Saha, Pierre Gaillard, and Michal Valko. "Improved Sleeping Bandits with Stochastic Action Sets and Adversarial Rewards". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 8357–8366. URL: https://proceedings.mlr.press/v119/saha20a.html.

[124] Shai Shalev-Shwartz. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

[125] Shai Shalev-Shwartz. "Online learning: theory, algorithms and applications". In: 2007. URL: https://api.semanticscholar.org/CorpusID:123001221.

[126] Shai Shalev-Shwartz and Yoram Singer. "A Primal-Dual Perspective of Online Learning Algorithms". In: *Machine Learning* 69,no. 2-3 (2007), pp. 115–142. URL: http://dx.doi.org/10.1007/s10994-007-5014-x.

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

[128] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. "Portfolio choices with orthogonal bandit learning". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, pp. 974–980. ISBN: 9781577357384.

[129] Nícollas Silva, Heitor Werneck, Thiago Silva, Adriano C.M. Pereira, and Leonardo Rocha. "Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions". In: *Expert Systems with Applications* 197 (2022), p. 116669. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116669. URL: https://www.sciencedirect.com/science/article/pii/S0957417422001543.

[130] Ambuj Tewari and Susan A. Murphy. "From Ads to Interventions: Contextual Bandits in Mobile Health". In: *Mobile Health - Sensors, Analytic Methods, and Applications*. 2017, pp. 495–517.

[131] William R Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3-4 (Dec. 1933), pp. 285–294. ISSN: 0006-3444. DOI: 10.1093/biomet/25.3-4.285. eprint: https://academic.oup.com/biomet/article-pdf/25/3-4/285/513725/25-3-4-285.pdf. URL: https://doi.org/10.1093/biomet/25.3-4.285.

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

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

[134] Dirk Van der Hoeven and Nicolò Cesa-Bianchi. "Nonstochastic bandits and experts with arm-dependent delays". In: *International Conference on Artificial Intelligence and Statistics*. 2022.

[135] Dirk Van der Hoeven, Tim Van Erven, and Wojciech Kotlowski. "The many faces of exponential weights in online learning". In: *Conference on Learning Theory*. 2018, pp. 2067–2092.

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

[137] Vladimir Vovk. "Aggregating strategies". English. In: *Proceedings of the Third Annual Workshop on Computational Learning Theory*. Ed. by M. Fulk and John Case. Morgan Kaufmann, 1990, pp. 371–383.

[138] Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. "Learning and planning in environments with delayed feedback". In: *Autonomous Agents and Multi-Agent Systems* 18.1 (2009), p. 83.

[139] Manfred K. Warmuth and Dima Kuzmin. "Randomized Online PCA Algorithms with Regret Bounds that are Logarithmic in the Dimension". In: *Journal of Machine Learning Research* 9.75 (2008), pp. 2287–2320. URL: http://jmlr.org/papers/v9/warmuth08a.html.

[140] Marcelo J Weinberger and Erik Ordentlich. "On delayed prediction of individual sequences". In: *IEEE Transactions on Information Theory* 48.7 (2002), pp. 1959–1976.

[141] Zheng Wen, Branislav Kveton, and Azin Ashkan. "Efficient Learning in Large-Scale Combinatorial Semi-Bandits". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1113–1122. URL: https://proceedings.mlr.press/v37/wen15.html.

[142] Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. "A Simple Approach for Non-stationary Linear Bandits". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 26–28 Aug 2020, pp. 746–755. URL: https://proceedings.mlr.press/v108/zhao20a.html.

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

[144] Lukas Zierahn, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Gergely Neu. "Nonstochastic Contextual Combinatorial Bandits". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, 25–27 Apr 2023, pp. 8771–8813. URL: https://proceedings.mlr.press/v206/zierahn23a.html.

[145] Lukas Zierahn, Dirk van der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. "A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs". In: *Journal of Machine Learning Research* 26.104 (2025), pp. 1–60. URL: http://jmlr.org/papers/v26/24-0496.html.

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

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

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

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

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.