UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD Thesis

Julian Zimmert

Adversarially robust stochastic multi-armed bandits

Advisors: Yevgeny Seldin, Christina Lioma

Handed in: February 29, 2020

Abstract

In the past decade, multi-armed bandits have attracted significant attention from the online learning and machine learning community due to many relevant applications in both theory and practice. A majority of the literature assumes that the environment is fundamentally stochastic. This assumption underlies popular algorithms such as Ucb1 or Thompson Sampling. These algorithms enjoy provably fast learning, as long as the model assumptions hold. In practice, there is little guarantee that the environment is actually stochastic. That is why there has been extensive research into robust algorithms that perform even in adversarial environments, a setting that includes complete absence of stochasticity. Although there exist algorithms, e.g. Exp3, that achieve robustness in the adversarial regime, they come at the cost of significantly slower learning.

Naturally the question arises if a combination of both properties is achievable. Are there algorithms that enjoy optimal worst-case guarantees in adversarial regimes but adapt automatically to easier, i.e. stochastic, environments? Although a number of algorithms have been suggested, all of them are suboptimal in theory and underperforming in practice.

We derive the first algorithm that achieves the optimal (within constants) pseudo-regret in both adversarial and stochastic multi-armed bandits without prior knowledge of the regime and time horizon. The algorithm is based on Online Mirror Descent (Omd) and Follow the Regularised Leader (Ftrl) with Tsallis entropy regulariser. In addition, the proposed algorithm enjoys improved regret guarantees in several intermediate regimes. We provide an empirical evaluation of the algorithm demonstrating that it significantly outperforms Ucb1 and Exp3 in stochastic environments. We also provide examples of adversarial environments, where Ucb1 and Thompson Sampling exhibit almost linear regret, whereas our algorithm suffers only logarithmic regret. To the best of our knowledge, this is the first example demonstrating vulnerability of Thompson Sampling in adversarial environments.

We extend our results to semi-bandits, a generalisation of multi-armed bandits with applications in online routing and recommender systems. This algorithm is based on a novel hybrid regulariser applied in the Ftrl framework and also obtains optimality in both the stochastic and adversarial regimes.

In another extension, we propose a new algorithm for adversarial multi-armed bandits with unrestricted delays. It achieves optimal dependency on the time horizon and cumulative delays, without knowledge of these quantities. Additionally, we propose a refined tuning of the algorithm, which achieves improved regret guarantees when the delays are highly unbalanced. All our bounds strictly improve the state of the art and the algorithm requires less information from the environment.

Finally, we provide additional insights into the class of algorithms studied in this thesis by showing a fundamental connection between Omd and the Bayesian regret analysis of Thompson Sampling. We derive the best known bound for adversarial bandits and further improve regret bounds in several bandit and online learning problems.

Resumé

Det seneste årti har multi-armed bandits oplevet stor interesse fra forskere inden for online learning og machine learning grundet relevante anvendelser i både teori og praksis. Størstedelen af litteraturen arbejder under antagelser om, at læringsscenariet fundamentalt set er stokastisk. Denne antagelse ligger til grund for populære algoritmer som Ucb1 og Thompson Sampling. Disse algoritmer lærer beviseligt hurtigt så længe antagelserne holder. I praksis er der ikke garanti for at læringsscenariet faktisk er stokastisk. Derfor har der været omfattende forskning i robuste algoritmer, der virker selv i modarbejdende scenarier, hvilket også dækker over den totale mangel på stokastisitet. Selvom der findes algoritmer, f.eks. Exp3, der opnår robusthed overfor modarbejdende scenarier, har det en omkostning i form af væsentligt langsommere læring.

Et naturligt spørgsmål er, om en kombination af disse egenskaber kan opnås. Findes der algoritmer, der både har optimale worst-case garantier i modarbejdende scenarier, men automatisk tilpasser sig nemmer, dvs. stokastiske, scenarier? Selvom en række algoritmer tidligere er blevet foreslået, er de alle teoretisk suboptimale og fungerer dårligt i praksis.

Vi udleder den første algoritme med optimal pseudo-regret (op til konstanter) overfor både modarbejdende og stokastiske multi-armed bandits uden forhåndskendskab til læringsscenarie eller tidshorisont. Algoritmen er baseret på Online Mirror Descent (Omd) og Follow the Regularised Leader (Ftrl) med Tsallis-entropi som regulariser. I tillæg har algoritmen også forbedrede garantier i flere mellemliggende scenarier. Vi udfører en empirisk evaluering af algoritmen, som demonstrerer, at den er væsentligt bedre end Ucb1 og Exp3 i stokastiske scenarier. Vi designer derudover også eksempler på modarbejdende scenarier, hvor Ucb1 og Thompson Sampling har tæt på lineær regret, hvorimod vores algoritme kun har logaritmisk regret. Vi har ikke kendskab til tidligere eksempler, der demonstrerer Thompson Samplings sårbarhed overfor modarbejdende scenarier.

Vi udvider vores resultater til semi-bandits, en generalisering af multi-armed bandit som anvendes i online routing og recommender-systemer. Denne algoritme er baseret på en ny hybrid regulariser sammen med algoritmeklassen Ftrl, og er optimal indenfor både stokastiske og modarbejdende scenarier.

I en anden udvidelse udvikler vi en ny algoritme til modarbejdende multi-armed bandits med forsinket feedback uden begrænsninger på forsinkelserne. Algoritmens afhængighed af tidshorisonten og de kumulative forsinkelser er optimal uden at kræve forhåndskendskab til disse. Derudover giver vi en fintuning af algoritmen som giver forbedret regret-garantier, når forsinkelserne er ubalancerede. Alle vores regret-begrænsninger forbedrer state-of-the-art, og algoritmen kræver mindre information om scenariet.

Endelig giver vi ny indsigt i klassen af algoritmer studeret i denne afhandling ved at påvise en fundamental forbindelse mellem Omd og den Bayesianske regret-analyse af Thompson Sampling. Vi udleder de hidtil bedste regret-begrænsninger for modarbejdende bandits og forbedrer derudover regret-begrænsninger for flere problemer indenfor bandits og online learning.

Acknowledgements

I would like to start by thanking the people with whom my journey into the field of bandits began: Tor Lattimore and Csaba Szepesvari. My research visit with Csaba at the University of Alberta in 2016 had been a major turning point in my academic career. Not only did I return with a glimpse into an intriguing world at the intersection of mathematics and computer science, I also made friends and collaborators. I am grateful for having been able to join their foundations team at DeepMind during an exhilarating internship in 2019.

I want to express my deep gratitude to my supervisors, Yevgeny Seldin and Christina Lioma, especially Yevgeny who had been a steady and important help throughout my studies. He was the one who strongly believed that the results we finally obtained were actually achievable, when many people in the field held the opposite opinion.

My last and most important thanks goes to the person who accompanied me throughout my years in Denmark. She gave up her cosy university position in Malaysia to join me in a cold and gloomy country on the other side of the world. 雷秋雁，谢谢你为我做的一切。

Contents

Chapter 1

Introduction

Multi-armed bandits are a fundamental paradigm in online learning which have seen an exponential growth in publications over the last two decades. The origins date back to Thompson [96], who studied optimal experimental design for medical studies. His point of departure was the observation, that blind controlled studies seem highly inefficient. In the later stages of such studies, there might be sufficient data available to focus the treatment on the most promising approaches, thereby improving the quality of treatment for the participants of the controlled study. If we truly care about providing the best treatment for the highest number of people, we should not divide the process into a *clinical trial* and a *market* phase, or in the bandit language, an *exploration* and *exploitation* phase.

Even though medical studies motivated the initial research in multi-armed bandits, to the best of our knowledge, it is a field where bandits have not actually been used. However, there are many applications in which bandits already play a fundamental role today. Major internet companies use bandit algorithms to optimise user interfaces, provide personalised news or content, deliver targeted ads and much more [38, 66, 73]. Bandit algorithms are also playing an important role in Monte-Carlo Tree Search, which is a crucial part of surpassing world experts in the game of Go [93].

Besides these applications, multi-armed bandits are perhaps the simplest model for decision making under uncertainty. Therefore, they can serve as an entry problem to gain elementary insights before tackling more complex problems. For example, many algorithms for reinforcement learning and partial monitoring have their roots in the bandit setting [19, 20]. For further reading on different bandit models and their importance in practise, we refer to Lattimore and Szepesvári [70].

This thesis focuses on bridging two distinct lines of work in the bandit literature, the *stochastic* and *adversarial* regimes. In multi-armed bandit *games*, an agent has to repeatedly choose an action (also called arm) from a finite set. The stochastic setting assumes that the outcomes of an agent's actions are drawn from i.i.d. distributions that only depend on the agent's choice. The seminal work of Lai and Robbins [67] provided an asymptotic lower bound on the performance of multi-armed bandit algorithms in the stochastic regime. We also have a multitude of optimal algorithms in various models, including Thompson's original algorithm from 1933 [61]. For a comprehensive overview of stochastic bandits, we refer to Bubeck and Cesa-Bianchi [26], Lattimore and Szepesvári [70], Slivkins et al. [94].

In real life applications, the stochastic model is often violated. Considering, for example, targeted ads, where the performance is measured in terms of clicks or interactions of the user. One can easily imagine that the weather strongly influences the user's temporary engagement, breaking the i.i.d. assumption.

To tackle this problem, the adversarial regime abandons all stochasticity assumption and replaces it with the assumption that outcomes are bounded in $[0, 1]$. At the beginning of the game, an adversary is allowed to allocate outcomes to counter our algorithm and we care about the worst-case guarantee against any adversary. Surprisingly, it is still possible to learn in that regime and algorithms are known with almost matching upper and lower bounds [14].

In recent years, the focus shifted towards the question of combining both regimes. Is it possible to design an “adversarially robust stochastic multi-armed bandit”? In other words, is there an algorithm that is optimal in the stochastic regime, which also obtains worst-case guarantees if the stochasticity assumption is violated?

One line of work initiated by Bubeck and Slivkins [31] and later improved by Auer and Chiang [18] starts with a stochastic algorithm and performs sophisticated tests whether the environment might violate the i.i.d. assumption. If and only if such a violation is detected, the learner switches irreversibly to an adversarial algorithm. An obvious disadvantage of this approach is the need to know the time horizon in advance to tune the confidence intervals. While there are doubling tricks to turn finite-time algorithms into anytime algorithms, no doubling trick can do so for both the adversarial and the stochastic setting simultaneously, without extra cost in the regret [23].

A different approach introduced by Seldin and Slivkins [91] and improved by Seldin and Lugosi [90] and Wei and Luo [103] starts with an adversarial algorithm and improves the performance guarantees in stochastic regimes. These algorithms are more useful in practice, since they are natively anytime, i.e. they work without knowing how long the game lasts. However, all previously proposed algorithms are suboptimal in at least one of the regimes. Furthermore the logarithmic bounds in the stochastic regime are asymptotical results. As we show in our experiments, their empirical performances over realistic time horizons (up to 10^7 timesteps) do not surpass the performance of the adversarial baseline.

It has remained an open problem since the work of Bubeck and Slivkins [31] whether simultaneous optimality in both worlds with no prior knowledge about the regime is possible at all. Auer and Chiang [18] have shown that high probability bounds on the adversarial regret are incompatible with optimal pseudo-regret in the stochastic regime. Abbasi-Yadkori et al. [1] have shown that in the pure exploration setting it is also impossible to obtain the optimal rates in both stochastic and adversarial regimes.

We show that for pseudo-regret, however, optimality in both regimes is indeed achievable.

1.1 Outline of the thesis

The thesis is structured in the following way.

Chapter 2 introduces Factored Bandits, a stochastic bandit problem in which each arm consists of taking independent actions over several sets. For example, an advertiser might choose a picture, a design, and a text for an online ad, each selected from a pool of arbitrarily combinable options. We assume that the identity of the optimal action in each set is independent of actions taken in other sets, however the reward distribution is dependent on the ensemble of actions. The main result of this chapter is an efficient meta-algorithm that runs almost communication-free sub-algorithms for each set and obtains up to constants optimal regret. From the inside of any sub-algorithm, the problem becomes an *intermediate* regime between stochastic and adversarial. Although the environment is stochastic, the sub-algorithm does not observe the full context, i.e. it does not observe the choices of the sub-algorithms dedicated to other sets.

Chapter 3 contains the main result of this thesis. We optimally solve the aforementioned problem of “adversarially robust stochastic multi-armed bandits”. This is achieved by developing a novel proof technique for OMD algorithms in stochastic regimes. Moreover, we define a more general *adversarial regime with a self-bounding constraint*, which includes the stochastic, stochastically constrained adversarial [103], and adversarially corrupted stochastic [76] regimes as special cases. We propose an algorithm that achieves logarithmic pseudo-regret guarantee in the adversarial regime with a self-bounding constraint simultaneously with the adversarial regret guarantee. The algorithm is based on OMD with regularisation by Tsallis entropy with power $\frac{1}{2}$. We name it TSALLIS-INF, where INF stands for Implicitly Normalised Forecaster [14]. The proposed algorithm is anytime: it requires neither the knowledge of the time horizon nor doubling schemes.

Chapter 4 extends the results to combinatorial semi-bandits, which is a natural generalisation of multi-armed bandits with applications in online routing and item recommendation. The learner has to pick a subset of actions, also called a combinatorial action and receives feedback for an action if and only if it was selected. Our algorithm is based on a novel hybrid regulariser applied in the follow the regularised leader (FTRL) framework and also obtains optimality in both the stochastic and adversarial regimes.

Chapter 5 extends the problem definition to arbitrarily delayed feedback. While the stochastic version of bandits with delay is well understood [57], the adversarial problem is significantly more challenging. We present the first simple anytime algorithm to obtain optimal performance in the adversarial setting.

Chapter 6 provides further insight into the family of algorithms used throughout this thesis. It shows that there is a remarkable relationship between OMD and the information theoretic analysis of the Bayesian regret of THOMPSON SAMPLING.

1.2 Main contributions

The main contributions of this thesis are as follows:

1. We provide an anytime algorithm for playing factored bandits in stochastic environments and analyse its regret. We also provide a lower bound matching up to constants.

2. We show that the algorithm can also be applied to utility-based dueling bandits, where the additive factor in the regret bound is reduced by a multiplicative factor of K compared to state-of-the-art (where K is the number of actions).

3. We propose the TSALLIS-INF algorithm, which is based on OMD with regularisation by Tsallis entropy with power $\alpha = \frac{1}{2}$. The algorithm achieves the optimal logarithmic pseudo-regret rate in the stochastic regime simultaneously with the optimal square-root adversarial regret guarantee with no prior knowledge of the regime. This resolves an open question of Bubeck and Slivkins [31].

4. When combined with reduced-variance loss estimators proposed by Zimmert and Lattimore [108], the leading constant of the stochastic regret bound for the TSALLIS-INF algorithm matches the asymptotic lower bound of Lai and Robbins [67] within a multiplicative factor of 2.

5. The leading constant of the adversarial regret bound for the same combination matches the minimax lower bound of Cesa-Bianchi and Lugosi [35, Theorem 6.1] within a multiplicative factor of less than 15, simultaneously with the stochastic bound. To the best of our knowledge, this is the best leading constant in an adversarial regret bound known today, matching the result of Zimmert and Lattimore [108].

6. We introduce an adversarial regime with a self-bounding constraint, which includes stochastic, stochastically constrained adversarial, and adversarially corrupted stochastic regimes as special cases. We show that TSALLIS-INF achieves logarthmic regret in the new regime simultaneously with the worst-case adversarial regret bound.

7. We improve the regret bound for adversarially corrupted stochastic regimes.

8. We use Tsallis-INF in a Sparring framework [8] to obtain an algorithm that achieves stochastic and adversarial optimality in utility-based dueling bandits.

9. We provide an empirical comparison of Tsallis-INF with standard algorithms from the literature. In one of the comparisons we design a stochastically constrained adversarial environment, where Thompson Sampling suffers almost linear regret. To the best of our knowledge, this is the first evidence that Thompson Sampling is not suitable for adversarial environments.

10. We propose a simple and general semi-bandit algorithm based on the Ftrl framework with a novel hybrid regulariser.

11. For any combinatorial action set, we prove that our algorithm achieves $\mathcal{O}(C_{sto} \log T)$ regret for stochastic environments and $\mathcal{O}(C_{adv}\sqrt{T})$ regret for adversarial environments, where C_{sto} and C_{adv} are problem-dependent factors (that do not depend on T) and are worst-case optimal.

12. We conduct experiments with synthetic data to show that our algorithm indeed adapts well to the nature of the environment. Additionally, we present a simple intermediate setting where our algorithm outperforms all baselines.

13. We provide an anytime Ftrl algorithm based on a novel hybrid regulariser. The regulariser combines $\frac{1}{2}$-Tsallis entropy and negative entropy, each with its own learning rate. The algorithm requires no advance knowledge of the delays and achieves a regret bound of $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$, which matches the lower bound within constants.

14. We provide a novel "skipping" technique, which allows to "ignore" rounds with excessively large delays with no advance knowledge of the delays. We put "skipping" and "ignore" in quotation marks, because the observations are still used by the algorithm and the "skipped" rounds are only excluded from the update of the learning rate. We prove an $\mathcal{O}(\sqrt{kn} + \min_S |S| + \sqrt{D_{\bar{S}} \log(k)})$ regret bound for the refined algorithm. The bound is slightly tighter than the refined regret bound of Thune et al. [98], but most importantly it requires no advance knowledge of the delays.

15. We prove a formal connection between the information-theoretic analysis and OMD. Specifically, we show how tools for analysing OMD can be applied to a modified version of Thompson sampling that uses the same sampling strategy as OMD, but replaces the mirror descent update with a Bayesian update.

16. We provide an efficient algorithm for adversarial k-armed bandits with regret $\sqrt{2kn} + O(k)$, matching the information-theoretic upper bound except for small lower-order terms.

17. Finally, we improve the regret guarantees for two online learning problems. First, for bandits with graph feedback we improve the minimax regret in the "easy" setting by a $\log(n)$ factor, matching the lower bound up to a factor of $\log^{3/2}(k)$. Second, for online linear optimisation over the ℓ_p-balls we improve existing bounds by arbitrarily large constant factors.

Chapter 2

Factored bandits

The work presented in this chapter is based on a paper that has been accepted as [110].

[110] Zimmert, J. and Seldin, Y. (2018). Factored bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2835–2844

Abstract

We introduce the factored bandits model, which is a framework for learning with limited (bandit) feedback, where actions can be decomposed into a Cartesian product of atomic actions. Factored bandits incorporate rank-1 bandits as a special case, but significantly relax the assumptions on the form of the reward function. We provide an anytime algorithm for stochastic factored bandits and up to constants matching upper and lower regret bounds for the problem. Furthermore, we show how a slight modification enables the proposed algorithm to be applied to utility-based dueling bandits. We obtain an improvement in the additive terms of the regret bound compared to state-of-the-art algorithms (the additive terms are dominating up to time horizons that are exponential in the number of arms).

2.1 Introduction

We introduce *factored bandits*, which is a bandit learning model, where actions can be decomposed into a Cartesian product of atomic actions. As an example, consider an advertising task, where the actions can be decomposed into (1) selection of an advertisement from a pool of advertisements and (2) selection of a location on a web page out of a set of locations, where it can be presented. The probability of a click is then a function of the quality of the two actions, the attractiveness of the advertisement and the visibility of the location it was placed at. In order to maximise the reward the learner has to maximise the quality of actions along each dimension of the problem. Factored bandits generalise the above example to an arbitrary number of atomic actions and arbitrary reward functions satisfying some mild assumptions.

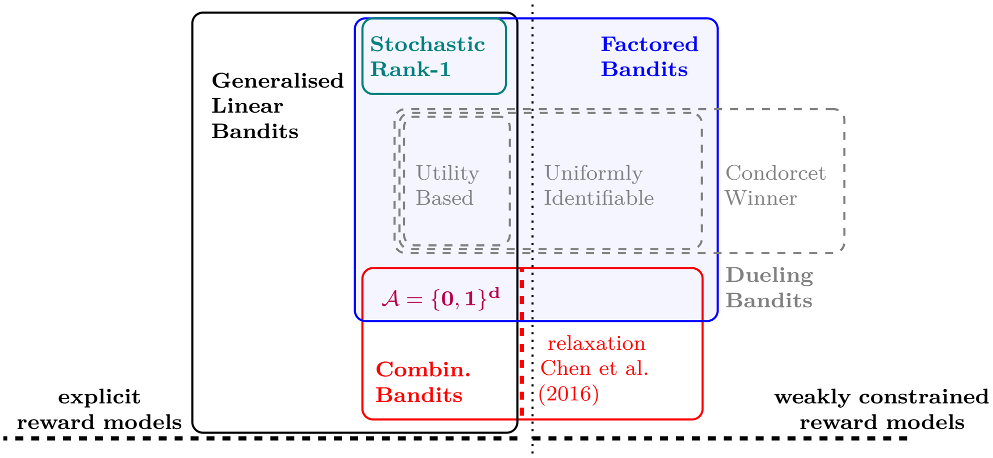


Figure 2.1: Relations between factored bandits and other bandit models.

In a nutshell, at every round of a factored bandit game the player selects L atomic actions, $a_1, \dots, a_L$, each from a corresponding finite set $\mathcal{A}_\ell$ of size $|\mathcal{A}_\ell|$ of possible actions. The player then observes a reward, which is an arbitrary function of $a_1, \dots, a_L$ satisfying some mild assumptions. For example, it can be a sum of the quality of atomic actions, a product of the qualities, or something else that does not necessarily need to have an analytical expression. The learner does not have to know the form of the reward function.

Our way of dealing with combinatorial complexity of the problem is through introduction of *unique identifiability* assumption, by which the best action along each dimension is uniquely identifiable. A bit more precisely, when looking at a given dimension we call the collection of actions along all other dimensions a *reference set*. The unique identifiability assumption states that in expectation the best action along a dimension outperforms any other action along the same dimension by a certain margin when both are played with the same reference set, irrespective of the composition of the reference set. This assumption is satisfied, for example, by the reward structure in linear and generalised linear bandits, but it is much weaker than the linearity assumption.

In Fig. 2.1, we sketch the relations between factored bandits and other bandit models. We distinguish between bandits with explicit reward models, such as linear and generalised linear bandits, and bandits with weakly constrained reward models, including factored bandits and some relaxations of combinatorial bandits. A special case of factored bandits are rank-1 bandits [60]. In rank-1 bandits the player selects two actions and the reward is the product of their qualities. Factored bandits generalise this to an arbitrary number of actions and significantly relax the assumption on the form of the reward function.

The relation with other bandit models is a bit more involved. There is an overlap between factored bandits and (generalised) linear bandits [2, 47], but neither is a special case of the other. When actions are represented by unit vectors, then for (generalised) linear reward functions the models coincide. However, the (generalised) linear bandits allow a continuum of actions, whereas factored bandits relax the (generalised) linearity assumption on the reward structure to uniform identifiability.

There is a partial overlap between factored bandits and combinatorial bandits [36]. The action set in combinatorial bandits is a subset of $\{0,1\}^d$. If the action set is unrestricted, i.e. $\mathcal{A} = \{0,1\}^d$, then combinatorial bandits can be seen as factored bandits with just two actions along each of the d dimensions. However, typically in combinatorial bandits the action set is a strict subset of $\{0,1\}^d$ and one of the parameters of interest is the permitted number of non-zero elements. This setting is not covered by factored bandits. While in the classical combinatorial bandits setting the reward structure is linear, there exist relaxations of the model, e.g. Chen et al. [40].

Dueling bandits are not directly related to factored bandits and, therefore, we depict them with faded dashed blocks in Fig. 2.1. While the action set in dueling bandits can be decomposed into a product of the basic action set with itself (one for the first and one for the second action in the duel), the observations in dueling bandits are the identities of the winners rather than rewards. Nevertheless, we show that the proposed algorithm for factored bandits can be applied to utility-based dueling bandits.

The main contributions of the paper can be summarised as follows:

1. We introduce factored bandits and the uniform identifiability assumption.

2. Factored bandits with uniformly identifiable actions are a generalisation of rank-1 bandits.

3. We provide an anytime algorithm for playing factored bandits under uniform identifiability assumption in stochastic environments and analyse its regret. We also provide a lower bound matching up to constants.

4. Unlike the majority of bandit models, our approach does not require explicit specification or knowledge of the form of the reward function (as long as the uniform identifiability assumption is satisfied). For example, it can be a weighted sum of the qualities of atomic actions (as in linear bandits), a product thereof, or any other function not necessarily known to the algorithm.

5. We show that the algorithm can also be applied to utility-based dueling bandits, where the additive factor in the regret bound is reduced by a multiplicative factor of K compared to state-of-the-art (where K is the number of actions). It should be emphasised that in state-of-the-art regret bounds for utility-based dueling bandits the additive factor is dominating for time horizons below $\Omega(\exp(K))$, whereas in the new result it is only dominant for time horizons up to $\mathcal{O}(K)$.

6. Our work provides a unified treatment of two distinct bandit models: rank-1 bandits and utility-based dueling bandits.

This chapter is organised in the following way. In Section 2.2 we introduce the factored bandit model and uniform identifiability assumption. In Section 2.3 we provide algorithms for factored bandits and dueling bandits. In Section 2.4 we analyse the regret of our algorithm and provide matching upper and lower regret bounds. In Section 2.5 we compare our work empirically and theoretically with prior work. We finish with a discussion in Section 2.6.

2.2 Problem Setting

2.2.1 Factored bandits

We define the game in the following way. We assume that the set of actions $\mathcal{A}$ can be represented as a Cartesian product of atomic actions, $\mathcal{A} = \bigotimes_{\ell=1}^{L} \mathcal{A}^\ell$. We call the elements of $\mathcal{A}^\ell$ *atomic arms*. For rounds $t = 1, 2, ...$ the player chooses an action $\mathbf{A}_t \in \mathcal{A}$ and observes a reward r_t drawn according to an unknown probability distribution $p_{\mathbf{A}_t}$ (i.e., the game is "stochastic"). We assume that the mean rewards $\mu(\mathbf{a}) = \mathbb{E}[r_t|\mathbf{A}_t = \mathbf{a}]$ are bounded in $[-1, 1]$ and that the noise $\eta_t = r_t - \mu(\mathbf{A}_t)$ is conditionally 1-sub-Gaussian. Formally, this means that

$$\forall \lambda \in \mathbb{R} \qquad \mathbb{E}\left[e^{\lambda \eta_t} | \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\right),$$

where $\mathcal{F}_t := \{\mathbf{A}_1, r_1, \mathbf{A}_2, r_2, ..., \mathbf{A}_t, r_t\}$ is the filtration defined by the history of the game up to and including round t. We denote $\mathbf{a}^* = (a_1^*, a_2^*, ..., a_L^*) = \arg\max_{\mathbf{a}\in\mathcal{A}} \mu(\mathbf{a})$.

Definition 2.1 (uniform identifiability). *An atomic set $\mathcal{A}^k$ has a uniformly identifiable best arm a_k^* if and only if*

$$\forall a \in \mathcal{A}^k \setminus \{a_k^*\} : \Delta_k(a) := \min_{\mathbf{b}\in\bigotimes_{\ell\neq k}\mathcal{A}^\ell} \mu(a_k^*, \mathbf{b}) - \mu(a, \mathbf{b}) > 0. \tag{2.1}$$

We assume that all atomic sets have uniformly identifiable best arms. The goal is to minimise the pseudo-regret, which is defined as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu(\mathbf{a}^*) - \mu(\mathbf{A}_t)\right].$$

Due to generality of the uniform identifiability assumption we cannot upper bound the instantaneous regret $\mu(\mathbf{a}^*) - \mu(\mathbf{A}_t)$ in terms of the gaps $\Delta_\ell(a_\ell)$. However, a sequential application of Eq. (2.1) provides a lower bound

$$\begin{aligned}\mu(\mathbf{a}^*) - \mu(\mathbf{a}) &= \mu(\mathbf{a}^*) - \mu(a_1, a_2^*, ..., a_L^*) + \mu(a_1, a_2^*, ..., a_L^*) - \mu(\mathbf{a}) \\ &\geq \Delta_1(a_1) + \mu(a_1, a_2^*, ..., a_L^*) - \mu(\mathbf{a}) \geq ... \geq \sum_{\ell=1}^{L} \Delta_\ell(a_\ell).\end{aligned} \tag{2.2}$$

For the upper bound let κ be a problem dependent constant, such that $\mu(\mathbf{a}^*) - \mu(\mathbf{a}) \leq \kappa \sum_{\ell=1}^{L} \Delta_\ell(a_\ell)$ holds for all $\mathbf{a}$. Since the mean rewards are in $[-1, 1]$, the condition is always satisfied by $\kappa = \min_{\mathbf{a},\ell} 2\Delta_\ell^{-1}(a_\ell)$ and by Eq. (2.2) κ is always larger than 1. The constant κ appears in the regret bounds. In the extreme case when $\kappa = \min_{\mathbf{a},\ell} 2\Delta_\ell^{-1}(a_\ell)$ the regret guarantees are fairly weak. However, in many specific cases mentioned in the previous section, κ is typically small or even 1. We emphasise that algorithms proposed in the paper do not require the knowledge of κ. Thus, the dependence of the regret bounds on κ is not a limitation and the algorithms automatically adapt to more favorable environments.

2.2.2 Dueling bandits

The set of actions in dueling bandits is factored into $\mathcal{A} \times \mathcal{A}$. However, strictly speaking the problem is not a factored bandit problem, because the observations in dueling bandits are not the rewards.[1] When playing two arms, a and b, we observe the identity of the winning arm, but the regret is typically defined via average relative quality of a and b with respect to a "best" arm in $\mathcal{A}$.

The literature distinguishes between different dueling bandit settings. We focus on *utility-based dueling bandits* [106] and show that they satisfy the uniform identifiability assumption.

In utility-based dueling bandits, it is assumed that each arm has a utility $u(a)$ and that the winning probabilities are defined by $\mathbb{P}[a \text{ wins against } b] = \nu(u(a) - u(b))$ for a monotonously increasing link function ν. Let $w(a, b)$ be 1 if a wins against b and 0 if b wins against a. Let $a^* := \arg\max_{a \in \mathcal{A}} u(a)$ denote the best arm. Then for any arm $b \in \mathcal{A}$ and any $a \in \mathcal{A} \setminus a^*$, it holds that $\mathbb{E}[w(a^*, b)] - \mathbb{E}[w(a, b)] = \nu(u(a^*) - u(b)) - \nu(u(a) - u(b)) > 0$, which satisfies the uniform identifiability assumption. For the rest of the paper we consider the linear link function $\nu(x) = \frac{1+x}{2}$. The regret is then defined by

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \frac{u(a^*) - u(A_t)}{2} + \frac{u(a^*) - u(B_t)}{2}\right]. \tag{2.3}$$

2.3 Algorithms

Although in theory an asymptotically optimal algorithm for any structured bandit problem was presented in [42], for factored bandits this algorithm does not only require solving an intractable semi-infinite linear program at every round, but it also suffers from additive constants which are exponential in the number of atomic actions L. An alternative naive approach could be an adaptation of sparring [105], where each factor runs an independent K-armed bandit algorithm and does not observe the atomic arm choices of other factors. The downside of sparring algorithms, both theoretically and practically, is that each algorithm operates under limited information and the rewards become non i.i.d. from the perspective of each individual factor.

Our Temporary Elimination Algorithm (TEA, Algorithm 1) avoids these downsides. It runs independent instances of the Temporary Elimination Module (TEM, Algorithm 3) in parallel, one per each factor of the problem. Each TEM operates on a single atomic set. The TEA is responsible for the synchronisation of TEM instances. Two main ingredients ensure information efficiency. First, we use relative comparisons between arms instead of comparing absolute mean rewards. This cancels out the effect of non-stationary means. The second idea is to use local randomisation in order to obtain unbiased estimates of the relative performance without having to actually play each atomic arm with the same reference, which would have led to prohibitive time complexity.

[1] In principle, it is possible to formulate a more general problem that would incorporate both factored bandits and dueling bandits. But such a definition becomes too general and hard to work with. For the sake of clarity we have avoided this path.

Algorithm 1: Factored Bandit TEA

```
$\forall \ell : \text{TEM}^\ell \leftarrow \text{new TEM}(\mathcal{A}^\ell)$
$t \leftarrow 1$
for $s = 1, 2, \dots$ do
    $M_s \leftarrow \arg\max_\ell |\text{TEM}^\ell.\text{getActiveSet}(f(t)^{-1})|$
    $T_s \leftarrow (t, t+1, \dots, t + M_s - 1)$
    for $\ell \in \{1, \dots, L\}$ in parallel do
        $\text{TEM}^\ell.\text{scheduleNext}(T_s)$
    for $t \in T_s$ do
        $r_t \leftarrow \mathit{play}((\text{TEM}^\ell.A_t)_{\ell=1,\dots,L})$
    for $\ell \in \{1, \dots, L\}$ in parallel do
        $\text{TEM}^\ell.\text{feedback}((r_{t'})_{t' \in T_s})$
    $t \leftarrow t + |T_s|$
```

Algorithm 2: Dueling Bandit TEA

```
1  $\text{TEM} \leftarrow \text{new TEM}(\mathcal{A})$
2  $t \leftarrow 1$
3  for $s = 1, 2, \dots$ do
4      $\mathcal{A}_s \leftarrow \text{TEM}.\text{getActiveSet}(f(t)^{-1})$
5      $T_s \leftarrow (t, t+1, \dots, t + |\mathcal{A}_s| - 1)$
6      $\text{TEM}.\text{scheduleNext}(T_s)$
7      for $b \in \mathcal{A}_s$ do
8          $r_t \leftarrow \mathit{play}(\text{TEM}.A_t, b)$
9          $t \leftarrow t + 1$
10     $\text{TEM}.\text{feedback}((r_{t'})_{t' \in T_s})$
```

The TEM instances run in parallel in externally synchronised phases. Each module selects active arms in *getActiveSet*(δ), such that the optimal arm is included with high probability. The length of a phase is chosen such that each module can play each potentially optimal arm at least once in every phase. All modules schedule all arms for the phase in *scheduleNext*. This is done by choosing arms in a round robin fashion (random choices if not all arms can be played equally often) and ordering them randomly. All scheduled plays are executed and the modules update their statistics through the call of *feedback* routine. The modules use slowly increasing lower confidence bounds for the gaps in order to temporarily eliminate arms that are with high probability suboptimal. In all algorithms, we use $f(t) := (t+1)\log^2(t+1)$.

Dueling bandits For dueling bandits we only use a single instance of TEM. In each phase the algorithm generates two random permutations of the active set and plays the corresponding actions from the two lists against each other. (The first permutation is generated in Line 5 and the second in Line 6 of Algorithm 2.)

2.3.1 TEM

The TEM tracks empirical differences between rewards of all arms a_i and a_j in D_{ij}. Based on these differences, it computes lower confidence bounds for all gaps. The set $\mathcal{K}^*$ contains those arms where all LCB gaps are zero. Additionally the algorithm keeps track of arms that were never removed from $\mathcal{B}$. During a phase, each arm from $\mathcal{K}^*$ is played at least once, but only arms in $\mathcal{B}$ can be played more than once. This is necessary to keep the additive constants at $M\log(K)$ instead of MK.

Algorithm 3: Temporary Elimination Module (TEM) Implementation

```
global : $N_{i,j}, D_{i,j}, \mathcal{K}^*, \mathcal{B}$
Function initialize($\mathcal{K}$)
    $\forall a_i, a_j \in \mathcal{K} : N_{i,j}, D_{i,j} \leftarrow 0, 0$
    $\mathcal{B} \leftarrow \mathcal{K}$

Function getActiveSet($\delta$)
    if $\exists N_{i,j} = 0$ then
        $\mathcal{K}^* \leftarrow \mathcal{K}$
    else
        for $a_i \in \mathcal{K}$ do
            $\hat{\Delta}^{LCB}(a_i) \leftarrow \max_{a_j \neq a_i} \frac{D_{j,i}}{N_{j,i}} - \sqrt{\frac{12 \log(2Kf(N_{j,i})\delta^{-1})}{N_{j,i}}}$
        $\mathcal{K}^* \leftarrow \{a_i \in \mathcal{K} | \hat{\Delta}^{LCB}(a_i) \leq 0\}$
        if $|\mathcal{K}^*| = 0$ then
            $\mathcal{K}^* \leftarrow \mathcal{K}$
        $\mathcal{B} \leftarrow \mathcal{B} \cap \mathcal{K}^*$
        if $|\mathcal{B}| = 0$ then
            $\mathcal{B} \leftarrow \mathcal{K}^*$
    return $\mathcal{K}^*$

Function scheduleNext($\mathcal{T}$)
    for $a \in \mathcal{K}^*$ do
        $\tilde{t} \leftarrow$ random unassigned index in $\mathcal{T}$
        $A_{\tilde{t}} \leftarrow a$
    while not all $A_{t_s}, \dots, A_{t_s+|\mathcal{T}|-1}$ assigned do
        for $a \in \mathcal{B}$ do
            $\tilde{t} \leftarrow$ random unassigned index in $\mathcal{T}$
            $A_{\tilde{t}} \leftarrow a$

Function feedback($\{R_t\}_{t_s,\dots,t_s+M_s-1}$)
    $\forall a_i : N_s^i, R_s^i \leftarrow 0, 0$
    for $t = t_s, \dots, t_s + M_s - 1$ do
        $R_s^{A_t} \leftarrow R_s^{A_t} + R_t$
        $N_s^{A_t} \leftarrow N_s^{A_t} + 1$
    for $a_i, a_j \in \mathcal{K}^*$ do
        $D_{i,j} \leftarrow D_{i,j} + \min\{N_i^s, N_j^s\}(\frac{R_s^i}{N_s^i} - \frac{R_s^j}{N_s^j})$
        $N_{i,j} \leftarrow N_{i,j} + \min\{N_i^s, N_j^s\}$
```

2.4 Analysis

We start this section with the main theorem, which bounds the number of times the TEM pulls sub-optimal arms. Then we prove upper bounds on the regret for our main algorithms. Finally, we prove a lower bound for factored bandits that shows that our regret bound is tight up to constants.

2.4.1 Upper bound for the number of sub-optimal pulls by TEM

Theorem 2.1. *For any TEM submodule* TEM^ℓ *with an arm set of size* $K = |\mathcal{A}^\ell|$, *running in the TEA algorithm with* $M := \max_\ell |\mathcal{A}^\ell|$ *and any suboptimal atomic arm* $a \neq a^*$, *let* $N_t(a)$ *denote the number of times TEM has played the arm* a *up to time* t. *Then there exist constants* $C(a) \leq M$ *for* $a \neq a^*$, *such that*

$$\mathbb{E}[N_t(a)] \leq \frac{120}{\Delta(a)^2}\left(\log(2Kt\log^2(t)) + 4\log\left(\frac{48\log(2Kt\log^2(t))}{\Delta(a)^2}\right)\right) + C(a),$$

where $\sum_{a \neq a^*} C(a) \leq M\log(K) + \frac{5}{2}K$ *in the case of factored bandits and* $C(a) \leq \frac{5}{2}$ *for dueling bandits.*

Proof sketch. [The complete proof is provided in the Appendix.]

Step 1 We show that the confidence intervals are constructed in such a way that the probability of all confidence intervals holding at all epochs up from s' is at least $1 - \max_{s \geq s'} f(t_s)^{-1}$. This requires a novel concentration inequality Lemma 2.3) for a sum of conditionally σ_s-sub-gaussian random variables, where σ_s can be dependent on the history. This technique might be useful for other problems as well.

Step 2 We split the number of pulls into pulls that happen in rounds where the confidence intervals hold and those where they fail: $N_t(a) = N_t^{conf}(a) + N_t^{\overline{conf}}(a)$.

We can bound the expectation of $N_t^{\overline{conf}}(a)$ based on the failure probabilities given by $\mathbb{P}[\text{conf failure at round s}] \leq \frac{1}{f(t_s)}$.

Step 3 We define s' as the last round in which the confidence intervals held and a was not eliminated. We can split $N_t^{conf}(a) = N_{t_{s'}}^{conf}(a) + C(a)$ and use the confidence intervals to upper bound $N_{t_{s'}}^{conf}(a)$. The upper bound on $\sum_a C(a)$ requires special handling of arms that were eliminated once and carefully separating the cases where confidence intervals never fail and those where they might fail. □

2.4.2 Regret Upper bound for Dueling Bandit TEA

A regret bound for the Factored Bandit TEA algorithm, Algorithm 1, is provided in the following theorem.

Theorem 2.2. *The pseudo-regret of Algorithm 1 at any time T is bounded by*

$$\begin{aligned}\mathfrak{R}_T \leq \kappa \left(\sum_{\ell=1}^{L} \sum_{a_\ell \neq a_\ell^*} \frac{120}{\Delta_\ell(a_\ell)} \left(\log(2|\mathcal{A}^\ell| t \log^2(t)) + 4 \log \left(\frac{48 \log(2|\mathcal{A}^\ell| t \log^2(t))}{\Delta_\ell(a_\ell)} \right) \right) \right) \\ + \max_\ell |\mathcal{A}^\ell| \sum_\ell \log(|\mathcal{A}^\ell|) + \sum_\ell \frac{5}{2} |\mathcal{A}^\ell|.\end{aligned}$$

Proof. The design of TEA allows application of Theorem 2.1 to each instance of TEM. Using $\mu(\mathbf{a}_*) - \mu(\mathbf{a}) \leq \kappa \sum_{\ell=1}^{L} \Delta_\ell(a_\ell)$, we have that

$$\mathfrak{R}_T = \mathbb{E}[\sum_{t=1}^{T} \mu(\mathbf{a}^*) - \mu(\mathbf{a}_t)]] \leq \kappa \sum_{l=1}^{L} \sum_{a_\ell \neq a_\ell^*} \mathbb{E}[N_T(a_\ell)] \Delta_\ell(a_\ell).$$

Applying Theorem 2.1 to the expected number of pulls and bounding the sums $\sum_a C(a)\Delta(a) \leq \sum_a C(a)$ completes the proof. □

2.4.3 Dueling bandits

A regret bound for the Dueling Bandit TEA algorithm (DBTEA), Algorithm 2, is provided in the following theorem.

Theorem 2.3. *The pseudo-regret of Algorithm 2 for any utility-based dueling bandit problem at any time T (defined in Eq. (2.3)) satisfies $\mathfrak{R}_T \leq \mathcal{O}\left(\sum_{a \neq a^*} \frac{\log(T)}{\Delta(a)}\right) + \mathcal{O}(K)$.*

Proof. At every round, each arm in the active set is played once in position A and once in position B in *play*(A, B). Denote by $N_t^A(a)$ the number of plays of an arm a in the first position, $N_t^B(a)$ the number of plays in the second position, and $N_t(a)$ the total number of plays of the arm. We have

$$\mathfrak{R}_T = \sum_{a \neq a_*} \mathbb{E}[N_t(a)]\Delta(a) = \sum_{a \neq a_*} \mathbb{E}[N_t^A(a) + N_t^B(a)]\Delta(a) = \sum_{a \neq a_*} 2\mathbb{E}[N_t^A(a)]\Delta(a).$$

The proof is completed by applying Theorem 2.1 to bound $\mathbb{E}[N_t^A(a)]$. □

2.4.4 Lower bound

We show that without additional assumptions the regret bound cannot be improved. The lower bound is based on the following construction. The mean reward of every arm is given by $\mu(\mathbf{a}) = \mu(\mathbf{a}^*) - \sum_\ell \Delta_\ell(a_\ell)$. The noise is Gaussian with variance 1. In this problem, the regret can be decomposed into a sum over atomic arms of the regret induced by pulling these arms, $\mathrm{Reg}_T = \sum_\ell \sum_{a_\ell \in \mathcal{A}^\ell} \mathbb{E}[N_T(a_\ell)]\Delta_\ell(a_\ell)$. Assume that we only want to minimise the regret induced by a single atomic set $\mathcal{A}^\ell$. Further, assume that $\Delta_k(a)$ for all $k \neq \ell$ are given. Then the problem is reduced to a regular K-armed bandit problem. The asymptotic lower bound for K-armed bandit under 1-Gaussian noise goes back to [67]: For any consistent strategy θ, the asymptotic regret is lower bounded by $\liminf_{T\to\infty} \frac{\mathrm{Reg}_T^\theta}{\log(T)} \geq \sum_{a \neq a_*} \frac{2}{\Delta(a)}$. Due to regret decomposition, we can apply this bound to every atomic set separately. Therefore, the asymptotic regret in the factored bandit problem is

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T^\theta}{\log(T)} \geq \sum_{\ell=1}^{L} \sum_{a^\ell \neq a^\ell_*} \frac{2}{\Delta^\ell(a^\ell)}.$$

This shows that our general upper bound is asymptotically tight up to leading constants and κ.

κ-gap We note that there is a problem-dependent gap of κ between our upper and lower bounds. Currently we believe that this gap stems from the difference between information and computational complexity of the problem. Our algorithm operates on each factor of the problem independently of other factors and is based on the "optimism in the face of uncertainty" principle. It is possible to construct examples in which the optimal strategy requires playing surely sub-optimal arms for the sake of information gain. For example, this kind of constructions were used by Lattimore and Szepesvári [69] to show suboptimality of optimism-based algorithms. Therefore, we believe that removing κ from the upper bound is possible, but requires a fundamentally different algorithm design. What is not clear is whether it is possible to remove κ without significant sacrifice of the computational complexity.

2.5 Comparison to Prior Work

2.5.1 Stochastic rank-1 bandits

Stochastic rank-1 bandits introduced by Katariya et al. [60] are a special case of factored bandits. The authors published a refined algorithm for Bernoulli rank-1 bandits using KL confidence sets in Katariya et al. [59]. We compare our theoretical results with the first paper because it matches our problem assumptions. In our experiments, we provide a comparison to both the original algorithm and the KL version.

In the stochastic rank-1 problem there are only 2 atomic sets of size K_1 and K_2. The matrix of expected rewards for each pair of arms is of rank 1. It means that for each $u \in \mathcal{A}^1$ and $v \in \mathcal{A}^2$, there exist $\overline{u}, \overline{v} \in [0,1]$ such that $\mathbb{E}[r(u,v)] = \overline{u} \cdot \overline{v}$. The proposed Stochastic rank-1 Elimination algorithm introduced by Katariya et al. is a typical elimination style algorithm. It requires knowledge of the time horizon and uses phases that increase exponentially in length. In each phase, all arms are played uniformly. At the end of a phase, all arms that are sub-optimal with high probability are eliminated.

Theoretical comparison It is hard to make a fair comparison of the theoretical bounds because TEA operates under much weaker assumptions. Both algorithms have a regret bound of $\mathcal{O}\left(\left(\sum_{u\in\mathcal{A}^1\setminus u^*} \frac{1}{\Delta_1(u)} + \sum_{v\in\mathcal{A}^2\setminus v^*} \frac{1}{\Delta_2(v)}\right)\log(t)\right)$. The problem independent multiplicative factors hidden under $\mathcal{O}$ are smaller for TEA, even without considering that rank-1 Elimination requires a doubling trick for anytime applications. However, the problem dependent factors are in favor of rank-1 Elimination, where the gaps correspond to the mean difference under uniform sampling $(\overline{u}^* - \overline{u}) \sum_{v \in \mathcal{A}^2} \overline{v}/K_2$. In factored bandits, the gaps are defined as $(\overline{u}^* - \overline{u}) \min_{v \in \mathcal{A}^2} \overline{v}$, which is naturally smaller. The difference stems from different problem assumptions. Stronger assumptions of rank-1 bandits make elimination easier as the number of eliminated suboptimal arms increases. The TEA analysis holds in cases where it becomes harder to identify suboptimal arms after removal of bad arms. This may happen when highly suboptimal atomic actions in one factor provide more discriminative information on atomic actions in other factors than close to optimal atomic actions in the same factor (this follows the spirit of illustration of suboptimality of optimistic algorithms in [69]). We leave it to future work to improve the upper bound of TEA under stronger model assumptions.

In terms of memory and computational complexity, TEA is inferior to regular elimination style algorithms, because we need to keep track of relative performances of the arms. That means both computational and memory complexities are $\mathcal{O}(\sum_\ell |\mathcal{A}^\ell|^2)$ per round in the worst case, as opposed to rank-1 Elimination that only requires $\mathcal{O}\left(|\mathcal{A}^1| + |\mathcal{A}^2|\right)$.

Empirical comparison The number of arms is set to 16 in both sets. We always fix $\overline{u^*} - \overline{u} = \overline{v^*} - \overline{v} = 0.2$. We vary the absolute value of $\overline{u^* v^*}$. As expected, rank1ElimKL has an advantage when the Bernoulli random variables are strongly biased towards one side. When the bias is close to $\frac{1}{2}$, we clearly see the better constants of TEA. In the evaluation we clearly outperform rank-1 Elimination over different parameter settings and even beat the KL optimised version if the means are not too close to zero or one. This supports that our algorithm does not only provide a more practical anytime version of elimination, but also improves on constant factors in the regret. We believe that our algorithm design can be used to improve other elimination style algorithms as well.

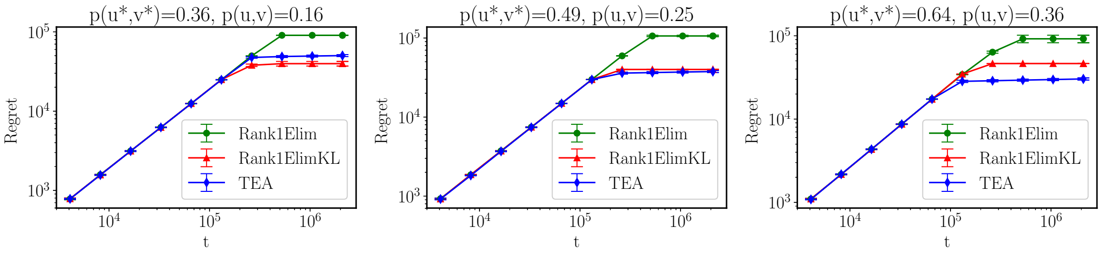


Figure 2.2: Comparison of Rank1Elim, Rank1ElimKL, and TEA for $K = L = 16$. The results are averaged over 20 repetitions of the experiment.

2.5.2 Dueling Bandits: Related Work

To the best of our knowledge, the proposed Dueling Bandit TEA is the first algorithm that satisfies the following three criteria simultaneously for utility-based dueling bandits:

- It requires no prior knowledge of the time horizon (nor uses the doubling trick or restarts).
- Its pseudo-regret is bounded by $\mathcal{O}(\sum_{a \neq a^*} \frac{\log(t)}{\Delta(a)})$.
- There are no additive constants that dominate the regret for time horizons $T > \mathcal{O}(K)$.

We want to stress the importance of the last point. For all state-of-the-art algorithms known to us, when the number of actions K is moderately large, the additive term is dominating for any realistic time horizon T. In particular, Ailon et al. [8] introduces three algorithms for the utility-based dueling bandit problem. The regret of Doubler scales with $\mathcal{O}(\log^2(t))$. The regret of MultiSBM has an additive term of order $\sum_{a\neq a^*}\frac{K}{\Delta(a)}$ that is dominating for $T < \Omega(\exp(K))$. The last algorithm, Sparring, has no theoretical analysis.

Algorithms based on the weaker Condorcet winner assumption apply to utility-based setting, but they all suffer from equally large or even larger additive terms. The RUCB algorithm introduced by Zoghi et al. [114] has an additive term in the bound that is defined as $2D\Delta_{max}\log(2D)$, for $\Delta_{max} = \max_{a\neq a^*}\Delta(a)$ and $D > \frac{1}{2}\sum_{a_i\neq a^*}\sum_{a_j\neq a_i}\frac{4\alpha}{\min\{\Delta(a_i)^2,\Delta(a_j)^2\}}$. By unwrapping these definitions, we see that the RUCB regret bound has an additive term of order $2D\Delta_{max} \geq \sum_{a\neq a^*}\frac{K}{\Delta(a)}$. This is again the dominating term for time horizons $T \leq \Omega(\exp(K))$. The same applies to the RMED algorithm introduced by Komiyama et al. [62], which has an additive term of $\mathcal{O}(K^2)$. (The dependencies on the gaps are hidden behind the $\mathcal{O}$-notation.) The D-TS algorithm by Wu and Liu [104] based on Thompson Sampling shows one of the best empirical performances, but its regret bound includes an additive constant of order $\mathcal{O}(K^3)$.

Other algorithms known to us, Interleaved Filter [105], Beat the Mean [107], and SAVAGE [100], all require knowledge of the time horizon T in advance.

Empirical comparison We have used the framework provided by Komiyama et al. [62]. We use the same utility for all sub-optimal arms. In Fig. 2.3, the winning probability of the optimal arm over suboptimal arms is always set to 0.7, we run the experiment for different number of arms K. TEA outperforms all algorithms besides RMED variants, as long as the number of arms are sufficiently big. To show that there also exists a regime where the improved constants gain an advantage over RMED, we conducted a second experiment in 2.4 (in the Appendix), where we set the winning probability to 0.95[2] and significantly increase the number of arms. The evaluation shows that the additive terms are indeed non-negligible and that Dueling Bandit TEA outperforms all baseline algorithms when the number of arms is sufficiently large.

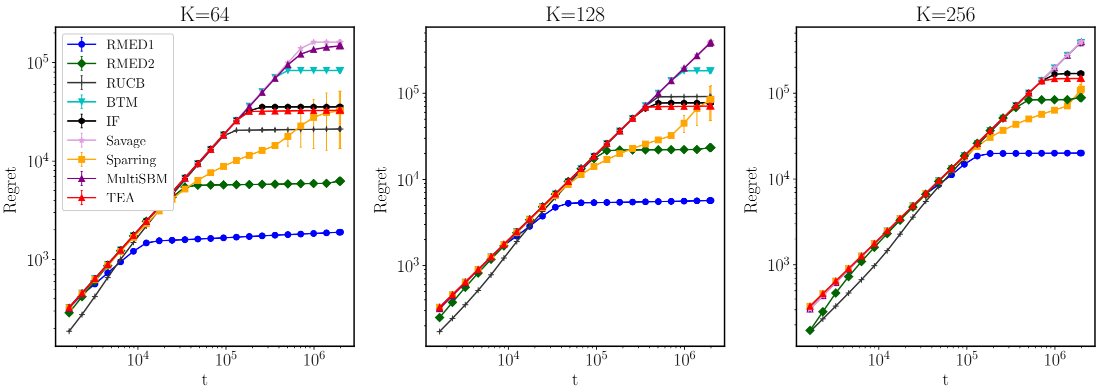


Figure 2.3: Comparison of Dueling Bandits algorithms with identical gaps of 0.4. The results are averaged over 20 repetitions of the experiment.

2.6 Discussion

We have presented the factored bandits model and uniform identifiability assumption, which requires no knowledge of the reward model. We presented an algorithm for playing stochastic factored bandits with uniformly identifiable actions and provided matching upper and lower

[2] Smaller gaps show the same behavior but require more arms and more timesteps.

bounds for the problem up to constant factors. Our algorithm and proofs might serve as a template to turn other elimination style algorithms into improved anytime algorithms.

Factored bandits with uniformly identifiable actions generalise rank-1 bandits. We have also provided a unified framework for the analysis of factored bandits and utility-based dueling bandits. Furthermore, we improve the additive constants in the regret bound compared to state-of-the-art algorithms for utility-based dueling bandits.

There are multiple potential directions for future research. One example mentioned in the text is the possibility of improving the regret bound when additional restrictions on the form of the reward function are introduced or improvements of the lower bound when algorithms are restricted in computational or memory complexity. Another example is the adversarial version of the problem.

Appendix

Auxiliary Lemmas

Lemma 2.1. *Given positive real numbers $\sigma_1, \sigma_2, \dots, \sigma_n$.. If $(X_i)_{i=1,\dots,n}$ is a sequence of random variables such that X_i conditioned on $X_{i-1}, X_{i-2}, \dots$ is σ_i-sub-Gaussian. Then $Z = \sum_{i=1}^n X_i$ is $\sqrt{\sum_{i=1}^n \sigma_i^2}$-sub-Gaussian.*

We believe this is a standard result, however we only found references for independent sub-Gaussian random variables.

Proof of Lemma 2.1. For $t = 1, \dots, n$ define $M_{s,t} = \exp(s\sum_{i=1}^t X_i - \frac{1}{2}\sum_{i=1}^t s^2\sigma_i^2)$. We claim $M_{s,t}$ is a super-martingale. Given that X_i are conditionally sub-Gaussian, we have $\mathbb{E}[\exp(sX_{t+1})|X_t, X_{t-1}, \dots] \leq \exp(\frac{s^2\sigma_{t+1}^2}{2})$. So

$$\begin{aligned}\mathbb{E}[M_{s,t+1}|M_{s,t}] &= \mathbb{E}[\exp(sX_{t+1} - \frac{1}{2}s^2\sigma_{t+1}^2)M_{s,t}|M_{s,t}] \\ &= \mathbb{E}[\exp(sX_{t+1} - \frac{1}{2}s^2\sigma_{t+1}^2)|M_{s,t}]M_{s,t} \leq M_{s,t}.\end{aligned}$$

Additionally by definition of sub-Gaussian $\mathbb{E}[M_{s,1}] \leq 1$. Therefore $\mathbb{E}[M_{s,n}] \leq 1$. Finally we get that $\mathbb{E}[\exp(sZ)] = \mathbb{E}[M_{s,n} \cdot \exp(\sum_{i=1}^n \frac{s^2\sigma_i^2}{2})] \leq \exp(\sum_{i=1}^n \frac{s^2\sigma_i^2}{2})$. So Z is $\sqrt{\sum_{i=1}^n \sigma_i^2}$-sub-Gaussian. □

Lemma 2.2. *Let $y \geq 1, z \geq 10$, then for any $x > zy + 4z\log(zy)$:*

$$\frac{z(\log(f(x)) + y)}{x} < 1.$$

Proof. We can reparameterise $x = zy + \alpha z \log(zy)$ for $\alpha > 4$. Then

$$\begin{aligned}&\frac{zy + z\log(f\left(zy + \alpha z\log(zy)\right))}{zy + \alpha z\log(zy)} < 1 \\ \Leftrightarrow\ &\frac{\log(f\left(zy + \alpha z\log(zy)\right))}{\alpha\log(zy)} < 1 \\ \Leftrightarrow\ &f\left(zy + \alpha z\log(zy)\right) < (zy)^\alpha \\ \Leftarrow\ &f\left(zy + \alpha zy\log(zy)\right) < (zy)^\alpha.\end{aligned}$$

Using $\log(x) \leq \sqrt{x} - \frac{1}{2}$ and $\alpha > 4$, we have that

$$f\left(zy + \alpha zy\log(zy)\right) < f\left(zy + \alpha zy(\sqrt{zy} - \frac{1}{2})\right) < f(\alpha(zy)^{\frac{3}{2}} - 1) = \alpha(zy)^{\frac{3}{2}}\log^2(\alpha(zy)^{\frac{3}{2}}).$$

It is therefore sufficient to prove that for all $\tilde{x} > 10$ and $\alpha > 4$:

$$\begin{aligned}&\alpha \log^2(\alpha\tilde{x}^{\frac{3}{2}}) < \tilde{x}^{\alpha-\frac{3}{2}}\\ &\Leftarrow \alpha(\sqrt{\alpha} + \tilde{x}^{\frac{3}{4}})^2 < \tilde{x}^{\alpha-\frac{3}{2}}\\ &\Leftrightarrow \sqrt{\alpha}(\sqrt{\alpha}\tilde{x}^{\frac{3}{4}-\frac{\alpha}{2}} + \tilde{x}^{\frac{3}{2}-\frac{\alpha}{2}}) < 1.\end{aligned}$$

The minimum on the left hand side is obtained for $\alpha = 4$ and $\tilde{x} = 10$ for with it holds true. □

Lemma 2.3. *Let $\sigma \in \mathbb{R}$ and $X_1, X_2, \dots$ be a sequence of sub-Gaussian random variables adapted to the filtration $\mathcal{F}_1, \mathcal{F}_2, \dots,$ i.e. $\mathbb{E}[e^{sX_t}|X_1, X_2, \dots, X_{t-1}] \leq e^{-\frac{\sigma_t^2 s^2}{2}}$. Assume for all $t : \sum_{i=1}^t \sigma_i^2 = n_t\sigma^2$, with $n_t \in \mathbb{N}$ almost surely. Then*

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \sum_{i=1}^t X_i \geq \sqrt{2\sigma^2 n_t \log\left(\frac{f(n_t)}{\delta}\right)}\right] \leq \delta,$$

where $f(n_t) = 2(1+n_t)\log^2(1+n_t)$.

Note that unlike in Lemma 2.1, we do not require σ_t to be independent of $X_1, \dots, X_{t-1}$.

Proof. The proof follows closely the arguments presented in the proofs of Lemma 8 in Abbasi-yadkori et al. [2] and Lemma 14 in Lattimore and Szepesvári [69]. For $\psi \in \mathbb{R}$ define

$$M_{t,\psi} = \exp\left(\sum_{s=1}^t \psi X_s - \frac{\psi^2\sigma_s^2}{2}\right).$$

If $t_0 \leq \tau \leq t$ is a stopping time with respect to $\mathcal{F}$, then as in the proof of Abbasi-yadkori et al. [2, Lemma 8] we have $\mathbb{E}[M_{\tau,\psi}] \leq 1$. By Markov's inequality, we have

$$\mathbb{P}[M_{\tau,\psi} \geq 1/\delta] \leq \delta \qquad \Leftrightarrow \qquad \mathbb{P}\left[\sum_{s=1}^{\tau} X_s \geq \frac{\log(\delta^{-1})}{\psi} + \frac{\psi n_\tau \sigma^2}{2}\right] \leq \delta.$$

An optimal choice of ψ would be $\psi = \sqrt{2\frac{\log(1/\delta)}{n_\tau\sigma^2}}$, however ψX_t would not be $\mathcal{F}_t$-measurable for $t \leq \tau$ and $M_{t,\psi}$ would not be well defined. Instead, for $k \geq 1$ we define

$$\psi_k := \sqrt{\frac{2\log(f(k)\delta^{-1})}{k\sigma^2}}.$$

With a union bound, we get that

$$\mathbb{P}\left[\exists k \geq 1 : \sum_{s=1}^{\tau} X_s \geq \frac{\log(f(k)\delta^{-1})}{\psi_k} + \frac{\psi n_\tau \sigma^2}{2}\right] \leq \sum_{k=1}^{\infty} \frac{\delta}{f(k)} \leq \delta.$$

Using now $k = n_\tau$, for which this also holds, we get that

$$\mathbb{P}\left[\sum_{s=1}^{\tau} X_s \geq \sqrt{2\sigma^2 n_\tau \log\left(\frac{f(n_\tau)}{\delta}\right)}\right] \leq \delta.$$

The proof is completed by choosing a stopping time τ:

$$\tau = \min\left(\infty \cup \left\{t \geq 1 : \sum_{s=1}^t X_s \geq \sqrt{2n_t\sigma^2 \log\left(\frac{f(n_t)}{\delta}\right)}\right\}\right).$$

□

Lemma 2.4. *Given $X_1, X_2, \ldots, X_n$ random variables with means $p_1, p_2, \ldots, p_n \in [-1, 1]$, such that all $X_i - p_i$ are 1-sub-Gaussian. (e.g. Bernoulli random variables) Given further two sample sizes $m, k \geq 1$, such that $m + k \leq n$. Then for $I_m : |I_m| = m$ and $I_k : |I_k| = k$ disjoint uniform samples of indices in $(1, 2, \ldots, n)$ without replacement, the random variable*

$$Z = \frac{1}{m} \sum_{i \in I_m} X_i - \frac{1}{k} \sum_{i \in I_k} X_i,$$

is $\sqrt{\frac{3(m+k)}{mk}}$-sub-Gaussian.

Proof. Without loss of generality, we set $m \leq k$. By definition, the random variables X_i can be decomposed into $X_i = p_i + \eta_i$, where η_i are conditionally independent 1-sub-Gaussian random variables. Decomposing Z gives:

$$Z = \frac{1}{m} \sum_{i \in I_m} p_i - \frac{1}{k} \sum_{i \in I_k} p_i + \frac{1}{m} \sum_{i \in I_m} \eta_i - \frac{1}{k} \sum_{i \in I_k} \eta_i.$$

We define $\overline{I} = \{1, ..., n\} \setminus (I_m \cup I_k)$, the indices of remaining X_i's and $\overline{p} = \frac{1}{n} \sum_{i=1}^{n} p_i$ the mean of means. In order to show that $\frac{1}{m} \sum_{i \in I_m} p_i - \frac{1}{k} \sum_{i \in I_k} p_i$ is sub-Gaussian, we first draw the elements in $(p_i)_{i \in \overline{I}} = (\overline{P}_i)_{i=1,...,n-m-k}$ and then the set $(p_i)_{i \in I_m} = (P_i^m)_{i=1,...,m}$. Drawing the first element $\overline{P}_1$ can be written as $\overline{P}_1 = \overline{p} + \zeta_1$, where ζ_1 is sub-Gaussian. With continuous drawings, it holds that

$$\begin{aligned}
&\mathbb{E}[\overline{P}_2 | \overline{P}_1] = \overline{p} - \frac{1}{n-1} \zeta_1 \\
&\overline{P}_2 = \overline{p} - \frac{1}{n-1} \zeta_1 + \zeta_2 \\
&\mathbb{E}[\overline{P}_3 | \overline{P}_1, \overline{P}_2] = \overline{p} - \frac{1}{n-1} \zeta_1 - \frac{1}{n-2} \zeta_2 \\
&\overline{P}_3 = \overline{p} - \frac{1}{n-1} \zeta_1 - \frac{1}{n-2} \zeta_2 + \zeta_3 \\
&... \\
&\mathbb{E}[\overline{P}_{n-m-k} | \overline{P}_1, ..., \overline{P}_{n-m-k-1}] = \overline{p} - \sum_{i=1}^{n-m-k-1} \frac{1}{n-i} \zeta_i \\
&\overline{P}_{n-m-k} = \overline{p} - \sum_{i=1}^{n-m-k-1} \frac{1}{n-i} \zeta_i + \zeta_{n-m-k} \\
&\sum_{i=1}^{n-m-k} \overline{P}_i = (n-m-k)\overline{p} + \sum_{i=1}^{n-m-k} \frac{m+k}{n-i} \zeta_i
\end{aligned}$$

The noise variables ζ_i are all conditionally independent and 1-sub-Gaussian.

We continue with P_i^m in the same fashion:

$$\begin{aligned}
&\mathbb{E}[P_1^m|\overline{P}] = \overline{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i}\zeta_i \\
&P_1^m = \overline{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i}\zeta_i + \zeta_{n-k-m+1} \\
&\mathbb{E}[P_2^m|\overline{P}, P_1^m] = \overline{p} - \sum_{i=1}^{n-m-k+1} \frac{1}{n-i}\zeta_i \\
&P_2^m = \overline{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i}\zeta_i + \zeta_{n-k-m+2} \\
&\ldots \\
&\mathbb{E}[P_m^m|\overline{P}, P_1^m, ..., P_{m-1}^m] = \overline{p} - \sum_{i=1}^{n-k-1} \frac{1}{n-i}\zeta_i \\
&P_m^m = \overline{p} - \sum_{i=1}^{n-k-1} \frac{1}{n-i}\zeta_i + \zeta_{n-k} \\
&\sum_{i=1}^{m} P_m^m = (n-k)\overline{p} + \sum_{i=1}^{n-k} \frac{k}{n-i}\zeta_i - \sum_{i=1}^{n-m-k} \overline{P}_i \\
&= m\overline{p} - \sum_{i=1}^{n-m-k} \frac{m}{n-i}\zeta_i + \sum_{i=n-m-k+1}^{n-k} \frac{k}{n-i}\zeta_i.
\end{aligned}$$

We can now use

$$\frac{1}{k}\sum_{i\in I_k} p_i = \frac{1}{k}\left(n\overline{p} - \sum_{i=1}^{n-m-k} \overline{P}_i - \sum_{i=1}^{m} P_i^m\right),$$

to substitute

$$\begin{aligned}
\frac{1}{m}\sum_{i\in I_m} p_i - \frac{1}{k}\sum_{i\in I_k} p_i &= \frac{1}{m}\sum_{i=1}^{m} P_i^m - \frac{1}{k}\left(n\overline{p} - \sum_{i=1}^{n-m-k} \overline{P}_i - \sum_{i=1}^{m} P_i^m\right) \\
&= \frac{m+k}{mk}\sum_{i=1}^{m} P_i^m + \frac{1}{k}\sum_{i=1}^{n-m-k} \overline{P}_i - \frac{n}{k}\overline{p} \\
&= \frac{m+k}{mk}\left(m\overline{p} - \sum_{i=1}^{n-m-k} \frac{m}{n-i}\zeta_i + \sum_{i=n-m-k+1}^{n-k} \frac{k}{n-i}\zeta_i\right) \\
&\quad + \frac{1}{k}\left((n-m-k)\overline{p} + \sum_{i=1}^{n-m-k} \frac{m+k}{n-i}\zeta_i\right) - \frac{n}{k}\overline{p} \\
&= \sum_{i=n-m-k+1}^{n-k} \frac{m+k}{m(n-i)}\zeta_i \\
&= \sum_{i=0}^{m-1} \frac{m+k}{m(k+i)}\zeta_{n-k-i}.
\end{aligned}$$

With these substitutions Z can be written as a weighted sum of conditionally independent sub-Gaussian random variables:

$$Z = \sum_{i=0}^{m-1} \frac{m+k}{m(k+i)}\zeta_{n-k-i} + \frac{1}{m}\sum_{i\in I_m} \eta_i - \frac{1}{k}\sum_{i\in I_k} \eta_i.$$

Therefore Z is according to Lemma 2.1 at least

$$\sqrt{\sum_{i=0}^{m-1}\left(\frac{m+k}{m(k+i)}\right)^2+\sum_{i=1}^{m}\frac{1}{m^2}+\sum_{i=1}^{k}\frac{1}{k^2}}\leq\sqrt{\frac{3(m+k)}{mk}}$$

-sub-Gaussian.

The last step uses the inequality

$$\begin{aligned}\sum_{i=0}^{m-1}\frac{1}{(k+i)^2}&=\int_0^m\frac{1}{(k+x)^2}\,dx+\sum_{i=0}^{m-1}\left(\frac{1}{(k+i)^2}-\int_{x=i}^{i+1}\frac{1}{(k+x)}\,dx\right)\\&=\frac{m}{(k+m)k}+\sum_{i=0}^{m-1}\frac{1}{(k+i)^2(k+i+1)}\\&\leq\frac{m}{(k+m)k}+\frac{1}{k+1}\sum_{i=0}^{m-1}\frac{1}{(k+i)^2}\\&\leq\frac{m(k+1)}{(k+m)k^2}\\&\leq\frac{2m}{(k+m)k}.\end{aligned}$$

□

Proof of Theorem 2.1

With the Lemmas from the previous section, we can proof our main theorem.

Proof of Theorem 2.1. We follow the steps from the sketch.

Step 1 We define the following shifted random variables.

$$\begin{aligned}\tilde{R}_t&:=R_t+\mu_t(a_*)-\mu_t(A_t)\\\tilde{R}_s^i&:=\sum_{t\in T_s}\mathbb{I}\{A_t=a_i\}\tilde{R}_t\\\Delta\tilde{D}_s^i&:=\frac{\tilde{R}_s^*}{N_s^*}-\frac{\tilde{R}_s^i}{N_s^i}\\\tilde{D}_s(a_i)&:=\sum_{k=1}^{s}\min\{N_s^i,N_s^*\}\Delta\tilde{D}_k^i\\\tilde{\Delta}_s(a_i)&:=\frac{\tilde{D}_s(a_i)}{N_{*,i}(s)}.\end{aligned}$$

The reward functions satisfy $\mu_t(a_*)-\mu_t(a_t)>\Delta(a_t)$ for all a_t. Therefore $R_t>\tilde{R}_t-\Delta(A_t)$. So we can bound $\frac{D_{*,i}}{N_{*,i}}>\Delta(a_i)+\tilde{\Delta}_s(a_i)$ and $\frac{D_{i,*}}{N_{i,*}}<-\Delta(a_i)-\tilde{\Delta}_s(a_i)$.

Define the events

$$\mathcal{E}_s:=\left\{\forall i:|\tilde{\Delta}_s(a_i)|\leq\sqrt{\frac{12\log(2Kf(N_{*,i})\delta_s^{-1})}{N_{*,i}}}\right\},\quad\mathcal{F}:=\bigcap_{s\geq 2}\mathcal{E}_s$$

and their complements $\overline{\mathcal{E}}_s,\overline{\mathcal{F}}$.

According to lemma 1, $\Delta\tilde{D}_s^i$ is $\sqrt{\frac{6}{\min\{N_s^*, N_s^i\}}}$-sub-Gaussian. So $\tilde{D}_s(a_i)$ is a sum of conditionally σ_i-sub-Gaussian random variables, such that $\sum_{i=1}^s \sigma_i^2 = 6N_{*,i}(s)$, Therefore we can apply Lemma 2.3. For both cases $\delta_s = \frac{1}{f(t_s)}$ and $\delta_s = \delta$, the probability never increases in time.

$$\mathbb{P}\left[\exists s' \geq s : \tilde{\Delta}_{s'}(a_i) \geq \sqrt{\frac{12\log(2Kf(\delta_{s'})N_{*,i})}{\delta_{s'}}}\right]$$
$$\leq \mathbb{P}\left[\exists s' \geq s : \tilde{D}_{s'}(a_i) \geq N_{*,i}\sqrt{\frac{12\log(2Kf(N_{*,i})\delta_s)}{N_{*,i}}}\right] \leq \frac{\delta_s}{2K}.$$

Using a union bound over $\pm\tilde{D}_s(a_i)$ for $a_i \in \mathcal{A}$, we get

$$\mathbb{P}[\overline{\mathcal{E}}_s] \leq \delta_s \text{ and } \mathbb{P}[\overline{\mathcal{F}}] \leq \delta_2.$$

step 2 We split the number of pulls in two categories: those that appear in rounds where the confidence intervals hold, and those that appear in rounds where they fail: $N_t^{\mathcal{E}}(a_i) = \sum_{s'=1}^{s(t)} \mathbb{I}\{\mathcal{E}_s\}N_s^i$, $N_t^{\overline{\mathcal{E}}}(a_i) = \sum_{s'=1}^{s(t)} \mathbb{I}\{\overline{\mathcal{E}}_s\}N_s^i$.

$$N_t(a_i) \leq N_t^{\mathcal{E}}(a_i) + N_t^{\overline{\mathcal{E}}}(a_i)$$
$$\mathbb{E}[N_t^{\mathcal{E}}(a_i)] = \mathbb{P}[\overline{\mathcal{F}}]\mathbb{E}[N_t^{\mathcal{E}}(a_i)|\mathcal{F}] + \mathbb{P}[\overline{\mathcal{F}}]\mathbb{E}[N_t^{\mathcal{E}}(a_i)|\overline{\mathcal{F}}].$$

In the high probability case, we are with probability $1-\delta$ in the event $\mathcal{F}$ and $N_t^{\overline{\mathcal{E}}}(a_i)$ is 0. In the setting of $\delta_s = f(t_s)^{-1}$, we can exclude the first round and start with $s = 2$ and $t_2 = M + 1$. This is because we do not use the confidence intervals in the first round.

$$\mathbb{E}[N_t^{\overline{\mathcal{E}}}(a_i)] \leq \sum_{s=2}^{\infty} \frac{t_{s+1} - t_s}{f(t_s)} \leq \sum_{s=1}^{\infty} \frac{M}{f(Ms)}$$
$$\leq \frac{M}{f(M)} + \sum_{s=2}^{\infty} \frac{M}{f(Ms)} \leq \frac{1}{2} + \sum_{s=1}^{\infty} \frac{1}{f(s)} \leq \frac{3}{2}$$

We use the fact that $\frac{1}{f(t_s)}$ is monotonically decreasing, so the expression gets minimized if all rounds are maximally long.

Step 3: bounding $\mathbb{E}[N_t^{\mathcal{E}}(a_i)|\mathcal{F}], \mathbb{E}[N_t^{\mathcal{E}}(a_i)|\overline{\mathcal{F}}]$

Let s' be the last round at which the arm a_i is not eliminated. We claim that $N_{i,*}$ at the beginning of round s' must be surely smaller or equal to $\frac{48}{\Delta(a_i)^2}\left(\log(2K\delta_{s'}^{-1}) + 4\log(\frac{48\log(2K\delta_{s'}^{-1})}{\Delta(a_i)^2})\right)$. Assume the opposite holds, then according to Lemma 2.2 with $z = \frac{48}{\Delta(a_i)^2}$ and $y = \log(2K\delta_{s'}^{-1})$:

$$\frac{\frac{48}{\Delta(a_i)^2}(\log(f(N_{i,*}(s'))) + \log(2K\delta_{s'}^{-1}))}{N_{i,*}(s')} < 1 \qquad \Leftrightarrow \qquad \sqrt{\frac{12\log(2Kf(N_{*,i})\delta_s^{-1})}{N_{*,i}}} < \frac{1}{2}\Delta(a_i).$$

So we have that

$$\hat{\Delta}_{s'}^{LCB}(a_i) \geq \Delta(a_i) - 2\sqrt{\frac{12\log(2Kf(N_{*,i})\delta_s^{-1})}{N_{*,i}}} > 0,$$

and a_i would have been excluded at the beginning of round s', which is a contradiction.

Let $C(a_i)$ denote the number of plays of a_i in round s'. Then for the different cases we have:

$$N_t^{\mathcal{E}}(a_i) - C(a_i) \leq \begin{cases} M \cdot N_{i,*}(s'), & \text{under the event } \overline{\mathcal{F}} \\ 2 \cdot N_{i,*}(s'), & \text{under the event } \mathcal{F} \\ N_{i,*}(s'), & \text{if } M_s = |\mathcal{A}_A| \end{cases}$$

$$\sum_{a \neq a_*} C(a) \leq \begin{cases} MK, & \text{under the event } \overline{\mathcal{F}} \\ M\log(K) + K, & \text{under the event } \mathcal{F} \\ K & \text{if } M_s = |\mathcal{A}_A| \end{cases}$$

The first case is trivial because each arm can only be played M times in a single round and $\min\{N_s^i, N_s^*\} \geq 1$ in rounds with $\mathcal{E}_s$. The second case follows from the fact that a_* is always in set $\mathcal{B}$ under the event $\mathcal{F}$. So $N_s^* \geq \max\{1, N_s^i - 1\}$ and $\min\{N_s^i, N_s^*\} \geq \frac{N_s^i}{2}$. The amount of pulls in a single round is naturally bounded by $\lceil \frac{M}{|\mathcal{B}|} \rceil \leq M$. Given that under the event $\mathcal{F}$, the set $\mathcal{B}$ never resets and the set $\mathcal{B}$ only decreases if an arm is eliminated, we can bound

$$\sum_{a_i \neq a_*} C(a_i) \leq \sum_{i=2}^{K} \lceil \frac{M}{i} \rceil \leq M\log(K) + K.$$

Finally the last case follows trivially because in the case of $M_s = |\mathcal{A}_A|$, we have $N_s^i = N_s^* = C(a_i) = 1$.

Step 4: combining everything

In the high probability case, we have with probability at least $1 - \delta$:

$$\begin{aligned} N_t(a_i) &\leq N_t^{\mathcal{E}}(a_i) + N_t^{\overline{\mathcal{E}}}(a_i) \\ &\leq 2N_{i,*}(s') + C(a_i) \\ &\leq \frac{96}{\Delta(a)^2}\left(\log(2K\delta^{-1}) + 4\log\left(\frac{48\log(2K\delta^{-1})}{\Delta(a)^2}\right)\right) + C(a_i) \end{aligned}$$

and also

$$\sum_{a \neq a_*} C(a) \leq M\log(K) + K.$$

If additionally $M_s = |\mathcal{A}_A|$, then the bound improves to

$$\begin{aligned} N_t(a_i) &\leq N_t^{\mathcal{E}}(a_i) + N_t^{\overline{\mathcal{E}}}(a_i) \\ &\leq N_{i,*}(s') + 1 \\ &\leq \frac{48}{\Delta(a)^2}\left(\log(2K\delta^{-1}) + 4\log\left(\frac{48\log(2K\delta^{-1})}{\Delta(a)^2}\right)\right) + 1. \end{aligned}$$

In the setting of $\delta_s = f(t_s)^{-1}$, we have

$$\begin{aligned} \mathbb{E}[N_t^{\mathcal{E}}(a_i) - C(a_i)] &\leq 2N_{i,*}(s') + \frac{1}{f(M)} M N_{i,*}(s') \\ &\leq \frac{120}{\Delta(a)^2}\left(\log(2Kt\log^2(t)) + 4\log\left(\frac{48\log(2Kt\log^2(t))}{\Delta(a)^2}\right)\right). \end{aligned}$$

So

$$\mathbb{E}[N_t(a_i)] \leq \mathbb{E}[C(a) + N_t^{\overline{\mathcal{E}}}(a_i)] + \frac{120}{\Delta(a)^2}\left(\log(2Kt\log^2(t)) + 4\log\left(\frac{48\log(2Kt\log^2(t))}{\Delta(a)^2}\right)\right).$$

where

$$\begin{aligned}\sum_{a\neq a_*} \mathbb{E}[C(a) + N_t^{\overline{\mathcal{E}}}(a_i)] &\leq M\log(K) + K + \frac{1}{f(M)}MK + \frac{3}{2}K \\ &\leq M\log(K) + \frac{5}{2}K.\end{aligned}$$

Finally if additionally $M_s = |\mathcal{A}_A|$, this bound improves to

$$\begin{aligned}\mathbb{E}[N_t(a_i)] &\leq \mathbb{E}[N_t^{\overline{\mathcal{E}}}(a_i)] + N_{*,i}(s') + 1 \\ &\leq \frac{5}{3} + \frac{48}{\Delta(a)^2}\left(\log(2Kt\log^2(t)) + 4\log\left(\frac{48\log(2Kt\log^2(t))}{\Delta(a)^2}\right)\right).\end{aligned}$$

□

Additional experiment

The winning probability is set to 0.95. All sub-optimal arms are identical

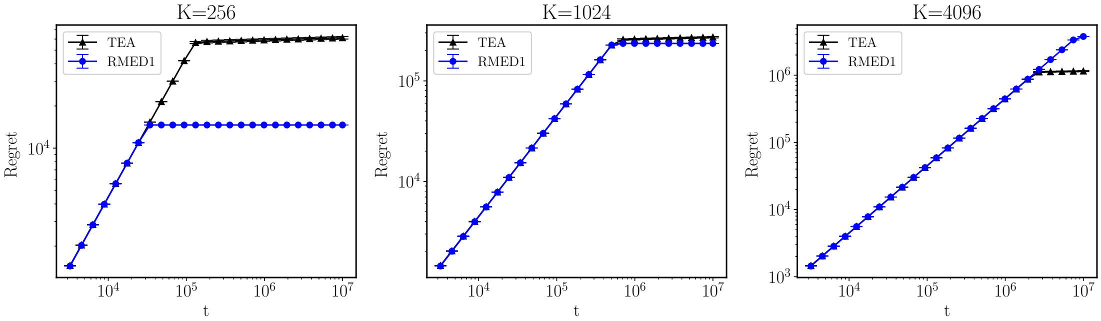


Figure 2.4: Comparison with identical gaps of 0.9. The results are averaged over 20 repetitions of the experiment.

Chapter 3

Multi-armed bandits

The work presented in this chapter is an extended version of a paper that has been accepted as [111]. The extended version is currently under submission as [113]. It has been accepted under minor revisions.

[111] Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475

[113] Zimmert, J. and Seldin, Y. (2020b). Tsalls-INF: An optimal algorithm for stochastic and adversarial bandits. *arXiv preprint arXiv:1807.07623*

Abstract

We derive an algorithm that achieves the optimal (within constants) pseudo-regret in both adversarial and stochastic multi-armed bandits without prior knowledge of the regime and time horizon.[1] The algorithm is based on online mirror descent (Omd) with Tsallis entropy regularisation with power $\alpha = 1/2$ and reduced-variance loss estimators. More generally, we define an adversarial regime with a self-bounding constraint, which includes stochastic regime, stochastically constrained adversarial regime Wei and Luo [103], and stochastic regime with adversarial corruptions [76] as special cases, and show that the algorithm achieves logarithmic regret guarantee in this regime and all of its special cases simultaneously with the adversarial regret guarantee. The algorithm also achieves adversarial and stochastic optimality in the utility-based dueling bandit setting. We provide empirical evaluation of the algorithm demonstrating that it significantly outperforms Ucb1 and Exp3 in stochastic environments. We also provide examples of adversarial environments, where Ucb1 and Thompson Sampling exhibit almost linear regret, whereas our algorithm suffers only logarithmic regret. To the best of our knowledge, this is the first example demonstrating vulnerability of Thompson Sampling in adversarial environments. Last, but not least, we present general stochastic and adversarial analyses of Omd algorithms with Tsallis entropy regularisation for $\alpha \in [0, 1]$ and explain the reason of why $\alpha = 1/2$ works best.

3.1 Introduction

Stochastic (i.i.d.) and adversarial multi-armed bandits are two fundamental sequential decision making problems in online learning [16, 17, 67, 83, 96]. When prior information about the nature of environment is available, it is possible to achieve $\mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i}\right)$ pseudo-regret in the stochastic case [16, 67] and $\mathcal{O}(\sqrt{KT})$ pseudo-regret in the adversarial case [13, 14], where T is the time horizon, K is the number of actions (a.k.a. arms), and Δ_i are suboptimality gaps. Both results match the lower bounds within constants, see [26] for a survey.[2] The challenge in recent years has been to achieve the optimal regret rates without prior knowledge about the nature of the problem.

One approach pursued by Bubeck and Slivkins [31] and later refined by Auer and Chiang [18] is to start playing under the assumption that the environment is i.i.d. and constantly monitor whether the assumption is satisfied. If a deviation from the i.i.d. assumption is detected, the algorithm performs an irreversible switch into an adversarial operation mode. This approach recovers the optimal bound in the stochastic case, but suffers from a multiplicative logarithmic factor in the regret in the adversarial case. Furthermore, the time horizon needs to be known in advance. The best known doubling schemes lead to extra multiplicative logarithmic factors in either the stochastic or the adversarial regime [23].

Another approach pioneered by Seldin and Slivkins [91] alters algorithms designed for adversarial bandits to achieve improved regret in the stochastic setting without losing the adversarial guarantees. They have introduced EXP3++, a modification of the EXP3 algorithm for adversarial bandits, which was later improved by Seldin and Lugosi [90] to achieve an anytime regret of $\mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\log(T)^2}{\Delta_i}\right)$ in the stochastic case while preserving optimality in the adversarial case. A related approach by Wei and Luo [103] uses log-barrier regularisation

[1] The paper expands and improves our earlier work [111].

[2] To be precise, the $\mathcal{O}(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i})$ stochastic regret rate is optimal when the means of the rewards are close to $\frac{1}{2}$, see Lai and Robbins [67], Cappé et al. [32], and Kaufmann et al. [61] for refined lower and upper bounds otherwise. However, the refined analysis applies to stochastic bandits, whereas we consider a more general setting, see Section 3.2 for details.

instead of entropic regularisation behind the EXP3. Their stochastic regret bound scales with $\log(T)$, although the constants are not spelled out explicitly and by empirical evaluation seem to be very large. Their adversarial regret guarantee scales with a square root of the cumulative loss of the best action in hindsight rather than a square root of the time horizon, but has an extra $\log T$ factor.

Seldin and Slivkins [91], Lykouris et al. [76], and Wei and Luo [103] also define a number of intermediate regimes between stochastic and adversarial bandits and provide improved regret guarantees for them.

The question of whether it is at all possible to achieve simultaneous optimality in both worlds with no prior knowledge about the regime has remained open since the work of Bubeck and Slivkins [31]. Auer and Chiang [18] have shown that no algorithm obtaining the optimal stochastic pseudo-regret bound can simultaneously achieve the optimal high-probability adversarial regret bound. Neither can an algorithm obtain the optimal stochastic pseudo-regret guarantee simultaneously with the optimal expected regret guarantee for adaptive adversaries.[3] In addition, Abbasi-Yadkori et al. [1] have shown that in the pure exploration setting it is also impossible to obtain the optimal rates in both stochastic and adversarial regimes.

We show that for pseudo-regret it is possible to achieve optimality in both regimes with a surprisingly simple algorithm. Moreover, we define a more general *adversarial regime with a self-bounding constraint*, which includes the stochastic, stochastically constrained adversarial [103], and adversarially corrupted stochastic [76] regimes as special cases. We propose an algorithm that achieves logarithmic pseudo-regret guarantee in the adversarial regime with a self-bounding constraint simultaneously with the adversarial regret guarantee. The algorithm is based on online mirror descent with regularisation by Tsallis entropy with power α. We name it α-TSALLIS-INF, or simply TSALLIS-INF for $\alpha = \frac{1}{2}$, where INF stands for Implicitly Normalised Forecaster [14]. The proposed algorithm is anytime: it requires neither the knowledge of the time horizon nor doubling schemes.

The main contributions of the paper are summarised in the following bullet points:

1. We propose the TSALLIS-INF algorithm, which is based on online mirror descent with regularisation by Tsallis entropy with power $\alpha = \frac{1}{2}$. The algorithm achieves the optimal logarithmic pseudo-regret rate in the stochastic regime simultaneously with the optimal square-root adversarial regret guarantee with no prior knowledge of the regime. This resolves an open question of Bubeck and Slivkins [31].

2. When combined with reduced-variance loss estimators proposed by Zimmert and Lattimore [108], the leading constant of the stochastic regret bound for the TSALLIS-INF algorithm matches the asymptotic lower bound of Lai and Robbins [67] within a multiplicative factor of 2.

3. The leading constant of the adversarial regret bound for the same combination matches the minimax lower bound of Cesa-Bianchi and Lugosi [35, Theorem 6.1] within a multiplicative factor of less than 15, simultaneously with the stochastic bound. To the best of our knowledge, this is the best leading constant in an adversarial regret bound known today, matching the result of Zimmert and Lattimore [108].

4. We introduce an adversarial regime with a self-bounding constraint, which includes stochastic, stochastically constrained adversarial, and adversarially corrupted stochastic regimes as special cases. We show that TSALLIS-INF achieves logarthmic regret in the new regime simultaneously with the worst-case adversarial regret bound.

[3] This does not contradict our result, because we bound the pseudo-regret, which is weaker than the expected regret.

	Regime	*Upper Bound* / *Lower Bound*
BROAD [103] *Corresponds to Tsallis entropy regularisation with* $\alpha = 0$. *Doubling is used for tuning the learning rate.*	Sto. Adv.	$\mathcal{O}(K)$ $\mathcal{O}\left(\sqrt{\log T}\right)$
$\alpha = \mathbf{\frac{1}{2}}$ (This paper) *Anytime. No need for gap estimation, doubling, or mixing.*	Sto. & Adv.	$\mathcal{O}(\mathbf{1})$
EXP3++ [90] *Corresponds to Tsallis entropy regularisation with* $\alpha = 1$. *Anytime. Mixed-in exploration is used for gap estimation.*	Sto. Adv.	$\mathcal{O}(\log T)$ $\mathcal{O}\left(\sqrt{\log K}\right)$

Table 3.1: Ratio of regret upper to lower bound for TSALLIS-INF and the closest prior work, BROAD and EXP3++.

5. We improve the regret bound for adversarially corrupted stochastic regimes.

6. We use TSALLIS-INF in a SPARRING framework [8] to obtain an algorithm that achieves stochastic and adversarial optimality in utility-based dueling bandits.

7. We provide a general analysis of OMD with Tsallis-Entropy regularisation with power $\alpha \in [0, 1]$ and explain the intuition of why $\alpha = \frac{1}{2}$ works best.

8. We provide an empirical comparison of TSALLIS-INF with standard algorithms from the literature. In one of the comparisons we design a stochastically constrained adversarial environment, where THOMPSON SAMPLING suffers almost linear regret. To the best of our knowledge, this is the first evidence that THOMPSON SAMPLING is not suitable for adversarial environments.

The paper is structured in the following way: In Section 3.2 we provide a formal definition of the problem setting, including the adversarial environment and the adversarial environment with a self-bounding constraint. Stochastic environments are a special case of the latter. In Section 3.3 we briefly review the framework of online mirror descent. We follow the techniques of Bubeck [25] to derive an anytime version of the family of algorithms based on regularisation by α-Tsallis Entropy [6, 99]. Section 3.4 contains the main theorems. We show that $\alpha = \frac{1}{2}$ provides an algorithm that is optimal in both adversarial regime and adversarial regime with a self-bounding constraint. The latter implies optimality in the stochastic regime. Interestingly, it is the same regularisation power $\alpha = \frac{1}{2}$ that has been used by Audibert and Bubeck [13, 14] in POLY-INF algorithm to achieve the optimal regret rate in the adversarial regime. We analyse the algorithm with standard importance-weighted loss estimators and with reduced-variance loss estimators proposed by [108]. The latter further reduces the constants and gets within a multiplicative factor of 15 from the minimax lower bound in the adversarial case and a multiplicative factor of 2 from the asymptotic lower bound in the stochastic case. Table 3.1 relates our results to the closest prior work on best-of-both-worlds algorithms. Wei and Luo [103] use logarithmic regularisation, which corresponds to Tsallis entropy with power $\alpha = 0$ and apply doubling for tuning the learning rate. Seldin and Lugosi [90] use entropic regularisation, which corresponds to Tsallis entropy with power $\alpha = 1$, and mix in additional exploration for estimation of the gaps. TSALLIS-INF with $\alpha = \frac{1}{2}$ requires neither doubling nor mixing nor estimation of the gaps. At the end of Section 3.4 we also provide a general analysis of the regret of α-TSALLIS-INF with $\alpha \in [0, 1]$ in adversarial environments and adversarial environments with a self-bounding constraint. We show that for $\alpha \neq \frac{1}{2}$ the optimal form of regularisation and learning rate for the adversarial regime and for the adversarial regime with a self-bounding constraint differ. Thus, for $\alpha \neq \frac{1}{2}$ the algorithm does not achieve simultaneous optimality in both regimes. Furthermore, for $\alpha \neq \frac{1}{2}$ the optimal regulariser for the adversarial regime with a self-bounding constraint requires oracle access to the unknown gaps. Prior work [90, 91, 103] used additional techniques, such as mixed-in exploration or doubling, to control the regret, but as we show in Table 3.1 the results were suboptimal. In Section 3.5 we show that the stochastic regime with adversarial corruptions of Lykouris et al. [76] is a special case of adversarial regime with a self-bounding constraint and that Tsallis-INF achieves the optimal regret rate there as well. In Section 3.6 we apply Tsallis-INF to dueling bandits. Section 3.7 contains proofs of our main theorems. In Section 3.8 we provide an empirical comparison of Tsallis-INF with baseline stochastic and adversarial bandit algorithms from the literature. We show that in stochastic environments with loss means close to $\frac{1}{2}$ Tsallis-INF with reduced-variance loss estimators significantly outperforms UCB1, EXP3, EXP3++, and Broad, and follows closely behind Thompson Sampling, whereas in certain adversarial environments it significantly outperforms UCB1 and Thompson Sampling, which suffer almost linear regret, and also significantly outperforms EXP3, EXP3++, and Broad. To the best of our knowledge, this is also the first evidence that Thompson Sampling is vulnerable in adversarial environments. We conclude with a summary in Section 3.9.

3.2 Problem Setting

At time $t = 1, 2, \dots,$ the agent chooses an arm $I_t \in \{1, \dots, K\}$ out of a set of K arms. The environment picks a loss vector $\ell_t \in [0,1]^K$ and the agent observes and suffers *only* the loss of the arm played, ℓ_{t,I_t}. The performance of an algorithm is measured in terms of pseudo-regret:

$$\overline{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_{t,I_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* \in \arg\min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$ is defined as a best arm in expectation in hindsight and the expectation is taken over internal randomisation of the algorithm and the environment.

In the (*adaptive*) *adversarial setting*, the adversary selects the losses arbitrarily, potentially based on the history of the agent's actions $(I_1, \dots, I_{t-1})$ and the adversary's own internal randomisation. For deterministic oblivious adversaries the definition of pseudo-regret coincides with the expected regret defined as $\mathbb{E}[Reg_T] := \mathbb{E}\left[\min_i \sum_{t=1}^{T} (\ell_{t,I_t} - \ell_{t,i})\right]$.

We further define an *adversarial regime with a* (Δ, C, T) *self-bounding constraint*, where the adversary selects losses such that at time T there exists a vector of gaps $\Delta \in [0,1]^K$ and a constant C for which the regret satisfies

$$\mathfrak{R} \geq \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{P}(I_t = i) - C. \tag{3.1}$$

The above condition should be satisfied at time T, but there is no requirement that it is satisfied for all $t < T$.

A simple instance of an adversarial regime with a self-bounding constraint is the *stochastic* regime. In the stochastic regime the losses $\ell_{t,i}$ are drawn from distributions with fixed means, $\mathbb{E}[\ell_{t,i}] = \mu_i$ independently of t, and the pseudo-regret can be written as

$$\mathfrak{R} = \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{P}(I_t = i), \tag{3.2}$$

where $\Delta_i = \mathbb{E}[\ell_{t,i}] - \min_i \mathbb{E}[\ell_{t,i}]$ is the suboptimality gap of action i. Thus, (3.1) is satisfied with Δ being the vector of suboptimality gaps and $C = 0$. In the stochastic regime the best arm $i^* = \arg\min_i \mu_i$ is the same for all the rounds, $i^*_T = i^*$ for all T (if there is more than one best arm we can pick one arbitrarily).

Another instance of an adversarial regime with a self-bounding constraint is the *stochastically constrained adversarial setting* [103]. In this setting the losses $\ell_{t,i}$ are drawn from distributions with fixed gaps, $\mathbb{E}[\ell_{t,i} - \ell_{t,j}] = \tilde{\Delta}_{i,j}$ independently of t, but the means, as well as other parameters of the distributions of all arms, are allowed to change with time and may depend on the agent's past actions $I_1, \dots, I_{t-1}$. Obviously, the stochastic regime is a special case of a stochastically constrained adversary. By using $i^* = \arg\min_i \tilde{\Delta}_{i,1}$ to denote an optimal arm (if there is more than one, we can pick one arbitrarily) we define a vector of suboptimality gaps Δ by taking $\Delta_i = \tilde{\Delta}_{i,i^*}$ and then the pseudo-regret satisfies the identity in (3.2) and the condition in equation (3.1) is satisfied with Δ and $C = 0$. In the stochastically constrained adversarial setting the best arm is also the same for all rounds, $i^*_T = i^*$ for all T.

In Section 3.5 we show that *stochastic bandits with adversarial corruptions* [76] are also a special case of an adversarial regime with a self-bounding constraint.

The motivation behind the definition of the adversarial regime with a self-bounding constraint will become clear when we explain the analysis. For simplified intuition, the reader can think about its special case, the stochastic regime, where the constraint (3.1) is satisfied by the identity in (3.2).

3.3 Online Mirror Descent

We recall a number of basic definitions and facts from convex analysis. The convex conjugate (a.k.a. Fenchel conjugate) of a function $f : \mathbb{R}^K \to \mathbb{R}$ is defined by

$$f^*(y) = \max_{x \in \mathbb{R}^K} \left\{ \langle x, y \rangle - f(x) \right\}.$$

We use

$$\mathcal{I}_{\mathcal{A}}(x) := \begin{cases} 0, & \text{if } x \in \mathcal{A} \\ \infty, & \text{otherwise} \end{cases}$$

to denote the characteristic function of a closed and convex set $\mathcal{A} \subset \mathbb{R}^K$. Hence, $(f + \mathcal{I}_{\mathcal{A}})^*(y) = \max_{x \in \mathcal{A}} \{\langle x, y\rangle - f(x)\}$. By standard results from convex analysis [84], for differentiable and convex f with invertible gradient $(\nabla f)^{-1}$ it holds that

$$\nabla (f + \mathcal{I}_{\mathcal{A}})^*(y) = \arg\max_{x \in \mathcal{A}} \left\{ \langle x, y \rangle - f(x) \right\} \in \mathcal{A}.$$

3.3.1 General Framework

The traditional online mirror descent (Omd) framework uses a fixed regulariser Ψ with certain regularity constraints [92]. The update rule is

$$w_1 = \min_{w \in \mathcal{A}} \Psi(w), \qquad w_{t+1} = \min_{w \in \mathcal{A}} a_t \langle w, \ell_t \rangle + D_\Psi(w, w_t),$$

where ℓ_t is the observed loss at time t, $\mathcal{A}$ is the convex body of the action set, and D_Ψ is the Bregman divergence $D_\Psi(x, y) = \Psi(x) - \Psi(y) - \langle x - y, \nabla\Psi(y)\rangle$. If the norm of the gradient of the regulariser $||\nabla\Psi(x)||$ is unbounded at the boundary of $\mathcal{A}$, then the update rule is equivalent to $w_{t+1} = \nabla(\Psi + \mathcal{I}_{\mathcal{A}})^*(-\sum_{s=1}^t a_s \ell_s)$, where $\sum_{s=1}^t a_s \ell_s$ is a weighted sum of past losses. This setting has been generalised to time-varying regularizers Ψ_t [81], where the updates are given by $w_{t+1} = \nabla(\Psi_t + \mathcal{I}_{\mathcal{A}})^*(-\sum_{s=1}^{t} \ell_s)$. Note that this formulation uses no weighting a_s of the losses. In the bandit setting we do not observe the complete loss vector ℓ_t. Instead, an unbiased estimator $\hat{\ell}_t : \mathbb{E}_{I_t \sim w_t}\left[\hat{\ell}_t\right] = \ell_t$ is used for updating the cumulative losses. The common way of constructing unbiased loss estimators is by using importance-weighted sampling:

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}_t(i)\ell_{t,i}}{w_{t,i}}\,,\ \text{where } \mathbb{1}_t(i) := \mathbb{1}(I_t = i) \text{ is the indicator function.} \tag{IW}$$

We use (IW) to denote these estimators. Zimmert and Lattimore [108] proposed reduced-variance importance-weighted loss estimators, which we call for brevity reduced-variance estimators or (RV)-estimators, and they are defined by

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}_t(i)(\ell_{t,i} - \mathbb{B}_t(i))}{w_{t,i}} + \mathbb{B}_t(i)\,,\ \text{where } \mathbb{B}_t(i) := \frac{1}{2}\mathbb{1}(w_{t,i} \geq \eta_t^2)\,. \tag{RV}$$

For any $\mathbb{B}_t(i) \in [0,1]$ the loss estimators remain unbiased, but their second moment $\mathbb{E}[\hat{\ell}_{t,i}^2]$ and variance are reduced. The value $\mathbb{B}_t(i) = \frac{1}{2}$ minimizes the worst-case variance of $\hat{\ell}_{t,i}$. However, the reduced-variance estimators can take negative values, $\hat{\ell}_{t,i} \geq -\frac{1}{2}\left(\frac{1}{w_{t,i}} - 1\right)$, while the analysis relies on non-negativity of the loss estimators. Zimmert and Lattimore [108] show that negative loss estimators can be dealt with, as long as they satisfy $\hat{\ell}_{t,i} \geq -\frac{1}{2}\eta_t^{-2}$. We achieve this by only reducing variance of the estimators with $w_{t,i} \geq \eta_t^2$.

At every step, we need to choose a probability distribution over arms w_t, so we add $\mathcal{I}_{\Delta^{K-1}}$ to the regularizers Ψ_t, thereby ensuring that $w_t \in \Delta^{K-1}$, where Δ^{K-1} is the probability simplex.

The algorithm is provided in Algorithm 4. Note that the framework is equivalent to what Abernethy et al. [5] call Gradient-Based Prediction (Gbp), where they replace $\nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*$ with suitable functions $\nabla\Phi_t : \mathbb{R}^K \to \Delta^{K-1}$. We adopt the notation $\Phi_t := (\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*$.

Algorithm 4: Online Mirror Descent for bandits

```
Input: $(\Psi_t)_{t=1,2,\ldots}$
Initialize: $\hat{L}_0 = \mathbf{0}_K$ (where $\mathbf{0}_K$ is a vector of $K$ zeros)
for $t = 1, \ldots$ do
    choose $w_t = \nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-\hat{L}_{t-1})$    % see Alg. 5 for an explicit calculation
    sample $I_t \sim w_t$
    observe $\ell_{t,I_t}$
    use (IW) or (RV) to construct $\hat{\ell}_t$
    update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
```

3.3.2 Omd with Tsallis Entropy Regularisation

We now consider a family of algorithms, which are regularised by the (negative) α-Tsallis entropy $H_\alpha(x) := \frac{1}{1-\alpha}\left(1 - \sum_i x_i^\alpha\right)$ [99]. We change the scaling and add linear terms, resulting in the following regulariser with learning rate η_t:

$$\Psi(w) := -\sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)\xi_i},$$

$$\Psi_t(w) := \frac{1}{\eta_t}\Psi(w).$$

Unless stated otherwise, we assume that $\xi_i = 1$ for all i, which leads to *symmetric regularisation.* In the stochastic analysis of α-Tsallis-INF with $\alpha \neq \frac{1}{2}$ we take $\xi_i = \Delta_i^{1-2\alpha}$, which leads to *asymmetric regularisation.* Since the gaps are unknown, the latter is mainly interesting from a theoretical point of view.

The resulting family of algorithms is a subset of Inf [14], which we call α-Tsallis-INF. α-Tsallis-INF with symmetric regularisation is related to the Poly-INF algorithm of Audibert and Bubeck [13, 14] and equivalent to the Gbp algorithm proposed by Abernethy et al. [6].

As has been observed earlier [6, 7], α-Tsallis-INF includes Exp3 based on the negative Shannon entropy $\sum_{i=1}^K w_i \log(w_i)$ [44] and algorithms based on the log-barrier potential $\sum_{i=1}^K -\log(w_i)$ [49] as special cases.[4] This can be seen by adding a constant term to the regularizer, so that $\Psi(w) = -\sum_i \frac{w_i^\alpha - \alpha w_i - (1-\alpha)}{\alpha(1-\alpha)\xi_i}$, and taking the respective limits $\alpha \to 0$ and $\alpha \to 1$. It gives:

$$\begin{aligned}
\lim_{\alpha\to 0} -\frac{w_i^\alpha - \alpha w_i - (1-\alpha)}{\alpha(1-\alpha)\xi_i} &= \lim_{\alpha\to 0} -\frac{\log(w_i) w_i^\alpha - w_i + 1}{(1-2\alpha)\xi_i} = -\xi_i^{-1}(\log(w_i) - w_i + 1),\\
\lim_{\alpha\to 1} -\frac{w_i^\alpha - \alpha w_i - (1-\alpha)}{\alpha(1-\alpha)\xi_i} &= \lim_{\alpha\to 1} -\frac{\log(w_i) w_i^\alpha - w_i + 1}{(1-2\alpha)\xi_i} = \xi_i^{-1}(\log(w_i) w_i - w_i + 1),
\end{aligned}$$

which are within linear and constant terms identical to the log-barrier potential and the negative Shannon entropy, respectively. Note that for symmetric regularisation neither the constant nor the linear terms influence the algorithm's choice of w, since it is normalised.

3.3.3 Implementation Details

The weights $w_{t,i}$ in Tsallis-INF are given implicitly through a solution of a constrained optimisation problem:

$$w_t = \operatorname*{arg\,max}_{w\in\Delta^{K-1}} \langle w, -\hat{L}_t \rangle + \frac{4}{\eta_t} \sum_i \sqrt{w_i}.$$

The solution takes the form

$$w_{t,i} = 4\left(\eta_t\left(\hat{L}_{t,i} - x\right)\right)^{-2},$$

where the normalisation factor x is defined implicitly through the constraint $\sum_i 4\left(\eta_t\left(\hat{L}_{t,i} - x\right)\right)^{-2} = 1$. The normalisation factor can be efficiently approximated by Newton's Method, reaching a sufficient precision in a very few iterations. Details of the computation are provided in Algorithm 5.

Algorithm 5: Newton's Method approximation of w_t in Tsallis-INF ($\alpha = \frac{1}{2}$)

```
Input: $x, \hat{L}_t, \eta_t$ %we use x from the previous iteration as a warmstart
repeat
    $\forall i : w_{t,i} \leftarrow 4(\eta_t(\hat{L}_{t,i} - x))^{-2}$
    $x \leftarrow x - (\sum_i w_{t,i} - 1)/(\eta_t \sum_i w_{t,i}^{\frac{3}{2}})$
until convergence
```

3.4 Main Results

In this section we present our main result, the Tsallis-INF algorithm with $\alpha = \frac{1}{2}$ that achieves the optimal regret bounds in both adversarial and stochastic bandits. We show that

[4] We use log to denote the natural logarithm throughout the paper.

it also achieves a logarithmic regret guarantee in the more general adversarial regime with a self-bounding constraint. In fact, the stochastic regret bound follows as a special case of the more general analysis. We then present a general analysis of α-Tsallis-INF with $\alpha \in [0,1]$ and explain the intuition of why $\alpha = \frac{1}{2}$ works best.

3.4.1 Analysis of Tsallis-INF with $\alpha = 1/2$

We show that Tsallis-INF with $\alpha = \frac{1}{2}$ and symmetric regulariser achieves the optimal $\sqrt{T}$ regret scaling in the adversarial regime and simultaneously $\log(T)$ regret scaling in the adversarial regime with a self-bounding constraint. The latter ensures the same regret scaling in stochastic and stochastically constrained adversarial environments as special cases. We analyse the algorithm with (IW) and (RV) loss estimators. Both estimators achieve the optimal regret scaling in both regimes, but the (RV) estimator yields better constants. The results for the two estimators are presented alongside each other using cases brackets and marked by (IW) and (RV), respectively.

Theorem 3.1. *The pseudo-regret of* Tsallis-INF *with $\alpha = \frac{1}{2}$, symmetric regularisation ($\xi_i = 1$), and learning rate*

$$\eta_t = \begin{cases} 2\sqrt{\frac{1}{t}}, & \textit{with } \text{(IW)} \textit{ estimators} \\ 4\sqrt{\frac{1}{t}}, & \textit{with } \text{(RV)} \textit{ estimators} \end{cases}$$

in any adversarial bandit problem satisfies:

$$\overline{Reg}_T \leq \begin{cases} 4\sqrt{KT} + 1, & \textit{with } \text{(IW)} \\ 2\sqrt{KT} + 10K\log(T) + 16, & \textit{with } \text{(RV)} \end{cases}.$$

If there exists a vector $\Delta \in [0,1]^K$ with a unique zero entry i^* (i.e., $\Delta_{i^*} = 0$ and $\Delta_i > 0$ for all $i \neq i^*$) and a constant C, such that the pseudo-regret at time T satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} w_{t,i}\Delta_i\right] - C \leq \overline{Reg}_T, \tag{3.3}$$

then the pseudo-regret further satisfies

$$\overline{Reg}_T \leq \begin{cases} \left(\sum_{i \neq i^*} \frac{4\log(T)+12}{\Delta_i}\right) + 4\log(T) + \frac{1}{\Delta_{\min}} + \sqrt{K} + 8 + C, & \textit{with } \text{(IW)} \\ \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + 20K\log(T) + \frac{1}{\Delta_{\min}} + \sqrt{K} + 32 + C, & \textit{with } \text{(RV)} \end{cases},$$

where $\Delta_{\min} := \min_{\Delta_i > 0} \Delta_i$. If C satisfies

$$\begin{array}{ll} C > \left(\sum_{i \neq i^*} \frac{4\log(T)+12}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}, & \textit{with } \text{(IW)} \\ C > \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}, & \textit{with } \text{(RV)} \end{array},$$

then the regret additionally satisfies

$$\overline{Reg}_t \leq \begin{cases} 2\sqrt{\left(\left(\sum_{i \neq i^*} \frac{4\log(T)+12}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}\right)C} + 4\log(T) + \sqrt{K} + 8, & \textit{with } \text{(IW)} \\ 2\sqrt{\left(\left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}\right)C} + 20K\log(T) + \sqrt{K} + 32, & \textit{with } \text{(RV)} \end{cases}.$$

The proof is postponed to Section 3.7.

Remark 3.1. *We call the condition in equation* (3.3) *a* self-bounding property *of the regret. As we have mentioned in Section 3.2, in the stochastically constrained adversarial environments and stochastic bandits as their special case* $\overline{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \sum_{i \neq i^*} w_{t,i}\Delta_i\right]$, *where* Δ *is the vector of suboptimality gaps, and under the assumption that the best arm is unique the condition in equation* (3.3) *is satisfied with* $C = 0$, *leading to a regret bound*

$$\overline{Reg}_T \leq \left(\sum_{i \neq i^*} \frac{\log(T) + 3}{\Delta_i}\right) + 20K\log(T) + \frac{1}{\Delta_{\min}} + \sqrt{K} + 32.$$

Remark 3.2. *The assumption that* Δ *has a unique zero entry and the corresponding assumption on uniqueness of the best arm in the stochastically constrained adversarial setting is a technical assumption we had to use in our proofs, but our experiments suggest that this is an artifact of the analysis. We conjecture that it can be removed, but explain the challenges in achieving the goal Section 3.7.*

The worst case lower bound for Stochastic MAB with Bernoulli losses is achieved when the expectations of the losses are close to $\frac{1}{2}$. Let Δ denote the vector of gaps and let $\mathbb{E}[\ell_{t,i}] = \frac{1}{2} + \Delta_i$, then for any consistent algorithm

$$\lim_{||\Delta|| \to 0} \left(\left(\sum_{i:\Delta_i > 0} \frac{1}{\Delta_i} \right)^{-1} \liminf_{t \to \infty} \frac{\mathbb{E}\left[\overline{Reg}_t\right]}{\log(t)} \right) \geq \frac{1}{2}.$$

The above lower bound follows from the well known divergence dependent lower bound of Lai and Robbins [67], see Appendix 3.9 for details. Therefore, the asymptotic regret upper bound of Tsallis-INF with RV-estimators is optimal within a multiplicative factor of 2, which is arguably a small price for a significant gain in robustness against adversaries. We leave it to future work to close the gap or prove that it is impossible to do so without compromising on the adversarial guarantees.

To the best of our knowledge, the leading constant 2 in the adversarial regret bound of Tsallis-INF with RV estimators provides the tightest adversarial regret guarantee known today. It matches the minimax adversarial lower bound in Cesa-Bianchi and Lugosi [35, Theorem 6.1] within a multiplicative factor of less than 15. Under the assumption of known time horizon, Zimmert and Lattimore [108] provide an adversarial regret bound with leading constant $\sqrt{2}$. The $\sqrt{2}$ multiplicative difference between their result and ours is the standard conversion rate between fixed-horizon and anytime regret bounds.

3.4.2 A General Alanysis of α-Tsallis-INF with $\alpha \in [0, 1]$

Now we provide a general analysis of α-Tsallis-INF with $\alpha \in [0, 1]$ and then explain the intuition of why $\alpha = \frac{1}{2}$ works best. Since $\alpha \neq \frac{1}{2}$ anyway leads to suboptimal regret rates and in order to keep things simple we restrict the general analysis to IW estimators. We note that in Theorem 3.1 the RV estimators helped improving the constants, but they did not change the rates and, therefore, we save the effort of optimising the constants in a priori suboptimal bounds. To keep things even simpler, we derive logarithmic bounds for stochastically constrained adversarial environments rather than the more general adversarial regime with a self-bounding constraint (technically speaking, we take $C = 0$).

Put attention that the adversarial analysis in Theorem 3.2 and stochastic analysis in Theorem 3.3 consider different versions of α-Tsallis-INF. The adversarial analysis uses *symmetric* regularisation, whereas stochastic analysis uses *asymmetric* regularisation. We get back to this point after we present the results.

Adversarial Regime

α-Tsallis-INF with symmetric regularisation has been previously analysed in the adversarial setting by Abernethy et al. [6] and Agarwal et al. [7]. Abernethy et al. provide a finite-time analysis for $\alpha \in (0,1]$, while Agarwal et al. analyse the case of $\alpha = 0$. The main contribution of the following theorem is that it provides a unified and anytime treatment for all $\alpha \in [0,1]$. The bound recovers the constants from Abernethy et al. without the need of tuning the learning rate by the time horizon T.

Theorem 3.2. *For any $\alpha \in [0,1]$, and any adversarial bandit problem the pseudo-regret of α-Tsallis-INF with symmetric regularizer, learning rate $\eta_t = \sqrt{\frac{K^{1-2\alpha}-K^{-\alpha}}{1-\alpha}\frac{1-t^{-\alpha}}{\alpha t}}$, and IW loss estimators at any time T satisfies*

$$\overline{Reg}_T \leq 2\sqrt{\min\left\{\frac{1}{\alpha-\alpha^2}, \frac{\log(K)}{\alpha}, \frac{\log(T)}{1-\alpha}\right\}KT} + 1.$$

(At the boundaries $\alpha = 0$ and $\alpha = 1$ the learning rates are defined by $\lim_{\alpha\to 0}\eta_t = \sqrt{\frac{(K-1)\log(t)}{t}}$ and $\lim_{\alpha\to 1}\eta_t = \sqrt{\frac{\log(K)(1-t^{-1})}{t}}$, respectively.)

The proof is postponed to Section 3.7.

Stochastically Constrained Adversarial Regime

Now we present an analysis of α-Tsallis-INF with $\alpha \in [0,1]$ and *asymmetric* regularisation in the stochastically constrained adversarial setting. We let $\bar{t} = \max\{e,t\}$. For learning rates $\eta_t = \frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}$ and asymmetric regulariser with $\xi_i = \Delta_i^{1-2\alpha}$ for $i \neq i^*$ and $\xi_{i^*} = \Delta_{\min}^{1-2\alpha}$, where $\Delta_{\min} = \min_{i\neq i^*}\Delta_i$, we prove the following theorem.

Theorem 3.3. *For any $\alpha \in [0,1]$ and any stochastically constrained adversarial regime with a unique best arm (i.e., $\Delta_i > 0$ for all i except a unique index i^* for which $\Delta_{i^*} = 0$), the pseudo-regret of α-Tsallis-INF with learning rate $\eta_t = \frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}$ and asymmetric regulariser with parameters $\xi_i = \Delta_i^{1-2\alpha}$ for $i \neq i^*$ and $\xi_{i^*} = \Delta_{\min}^{1-2\alpha}$ at any time T satisfies*

$$\overline{Reg}_T \leq \sum_{i\neq i^*}\left(\frac{(8\min\{\frac{1}{1-\alpha}, \log(T)\} + 64)\log(T)}{\Delta_i}\right) + \frac{16\log^4(\frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2}))}{\Delta_{\min}} + 4.$$

The proof is provided in Appendix 3.9.

Remark 3.3. *We emphasise that for $\alpha \neq \frac{1}{2}$ the result in Theorem 3.3 requires knowledge of the gaps Δ_i for tuning the regularisation parameters ξ_i. For $\alpha = \frac{1}{2}$ this knowledge is not required. Therefore, Theorem 3.3 is primarily interesting from the theoretical perspective of characterisation of behavior of α-Tsallis-INF in stochastically constrained adversarial environments, whereas $\alpha = \frac{1}{2}$ is the only practically interesting value with the refined analysis in Theorem 3.1.*

Remark 3.4. *For $\alpha \neq \frac{1}{2}$ the version α-Tsallis-INF in Theorem 3.3 uses* asymmetric *regularisation, whereas α-Tsallis-INF in Theorem 3.2 uses* symmetric *regularisation. The corresponding learning rates also differ. Therefore, for $\alpha \neq \frac{1}{2}$ neither of the two versions of α-Tsallis-INF achieves simultaneous optimality in the stochastic and adversarial setting. In fact, the time dependence of the adversarial regret guarantee for α-Tsallis-INF in Theorem 3.3 is in the order of $T^\alpha + T^{1-\alpha}$.*

Remark 3.5. *We note that while Tsallis entropy with $\alpha = 0$ corresponds to log-barrier potential used in* Broad*, and Tsallis entropy with $\alpha = 1$ corresponds to entropic regularisation used in* EXP3++*, the two algorithms (*Broad *and* EXP3++*) use* symmetric *regularisation, whereas α-*Tsallis-INF *in Theorem 3.3 uses* asymmetric *regularisation. Therefore, there is no direct relation between the result of Theorem 3.3 and these two algorithms. In particular,* Broad *and* EXP3++ *use other techniques to achieve slightly suboptimal, but simultaneous stochastic and adversarial regret guarantees (as described in Table 3.1), which is not the case for α-*Tsallis-INF *with asymmetric regularisation in Theorem 3.3.*

3.4.3 Intuition Behind the Success of Tsallis-INF with $\alpha = \frac{1}{2}$

It has been previously shown that regularisation by Tsallis entropy with power $\alpha = 1/2$ leads to the minimax optimal regret rate in the adversarial regime [14]. Here we provide some basic intuition on why the same value of α works well in the stochastic case. We also highlight the key breakthroughs that allow us to overcome challenges faced in prior work.

We start with a simple "back of the envelope" approximation of the form of the weights w_t played by Tsallis-INF. By definition of Algorithm 4, at round t we have

$$w_t = \operatorname*{arg\,max}_{w \in \Delta^{K-1}} \left\{ \left\langle w, -\hat{L}_{t-1} \right\rangle + \frac{1}{\eta_t} \sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)\xi_i} \right\}.$$

Taking a derivative of the Langrangian of the above expression with respect to w_i and equating it to zero, we obtain that

$$-\hat{L}_{t-1,i} + \frac{1}{\eta_t(1-\alpha)\xi_i}(w_{t,i}^{\alpha-1} - 1) - \nu = 0,$$

where ν is a Lagrange multiplier corresponding to the constraint that w is a probability distribution. We can express ν as

$$\nu = \frac{1}{\eta_t(1-\alpha)\xi_{i^*}}(w_{t,i^*}^{\alpha-1} - 1) - \hat{L}_{t-1,i^*}.$$

For $i \neq i^*$ this gives

$$\begin{aligned}
w_{t,i} &= \left(\eta_t(1-\alpha)\xi_i \left(\hat{L}_{t-1,i} + \nu\right) + 1\right)^{\frac{1}{\alpha-1}} \\
&= \left(\eta_t(1-\alpha)\xi_i \left(\hat{L}_{t-1,i} - \hat{L}_{t-1,i^*} + \frac{1}{\eta_t(1-\alpha)\xi_{i^*}}(w_{t,i^*}^{\alpha-1} - 1)\right) + 1\right)^{\frac{1}{\alpha-1}} \\
&= \left(\eta_t(1-\alpha)\xi_i \left(\hat{L}_{t-1,i} - \hat{L}_{t-1,i^*}\right) + \frac{\xi_i}{\xi_{i^*}}(w_{t,i^*}^{\alpha-1} - 1) + 1\right)^{\frac{1}{\alpha-1}} \\
&\approx \left(\eta_t(1-\alpha)\xi_i \left(\hat{L}_{t-1,i} - \hat{L}_{t-1,i^*}\right)\right)^{\frac{1}{\alpha-1}},
\end{aligned}$$

where the approximation holds because asymptotically the first term dominates the sum. A bit more explicitly, in order for the algorithm to deliver non-trivial regret guarantee, w_{t,i^*} should be close to 1. Thus, the last two terms in the brackets are roughly a constant. At the same time, as we discuss below, the whole expression in the brackets must grow roughly as $(\Delta_i^2 t)^{1-\alpha}$. Thus, the first term must dominate. In the stochastic regime $\mathbb{E}\left[\hat{L}_{t,i} - \hat{L}_{t,i^*}\right] = \Delta_i t$. If we use this in our back-of-the-envelope calculation, we obtain that for $i \neq i^*$ in the stochastic regime $\mathbb{E}[w_{t,i}] \approx \mathbb{E}\left[\left(\eta_t(1-\alpha)\xi_i(\hat{L}_{t-1,i} - \hat{L}_{t-1,i^*})\right)^{\frac{1}{\alpha-1}}\right] \propto (\eta_t \xi_i \Delta_i t)^{\frac{1}{\alpha-1}}$. (Strictly speaking, when we take the expectation inside the power we obtain an inequality, but we ignore this detail in the high-level discussion. We also ignore the $(1-\alpha)$ factor, which can be seen as a constant for $\alpha < 1$.)

In order to achieve a regret rate of $\Theta(\sum_{i\neq i^*} \frac{\log t}{\Delta_i})$ in the stochastic regime the suboptimal arms should be explored at a rate of $\Theta(\frac{1}{\Delta_i^2 t})$ per round (if $\mathbb{E}[w_{t,i}] = \Theta(\frac{1}{\Delta_i^2 t})$, then $\Delta_i \mathbb{E}\left[\sum_{s=1}^t w_{s,i}\right] = \Theta(\frac{\log t}{\Delta_i})$, as desired). Exploring more than that leads to excessive regret from the exploration alone. Exploring less is also prohibitive, because it leads to an overly high probability of misidentifying the best arm. By looking at the approximation of $\mathbb{E}[w_{t,i}]$ from the previous paragraph, we obtain that we should have $(\eta_t \xi_i \Delta_i t)^{\frac{1}{\alpha-1}} \propto \frac{1}{\Delta_i^2 t}$ or, equivalently, $\eta_t \xi_i \propto t^{-\alpha} \Delta_i^{1-2\alpha}$. The learning rate takes care of the time-dependent quantities, i.e., $\eta_t \propto t^{-\alpha}$, and ξ_i should take care of the arm-dependent quantities, i.e., we should have $\xi_i \propto \Delta_i^{1-2\alpha}$. Note that $\alpha = \frac{1}{2}$ leads to a symmetric regulariser Ψ (i.e., $\xi_i = 1$), whereas for $\alpha \neq \frac{1}{2}$ the regulariser must be tuned using unknown gaps Δ_i. The necessity to tune the regulariser based on unknown gaps has hindered progress in the work of Wei and Luo [103], who used the log-barrier regulariser corresponding to $\alpha = 0$.

Another crucial novelty behind the success of our analysis is basing it on the self-bounding property of the regret in equation (3.3). The new proof technique uses the same mechanism for controlling the regret in stochastic and adversarial regimes and we explain the intuition behind it in Section 3.7.1. The earlier approach by Seldin and Slivkins [91] and Seldin and Lugosi [90] has controlled the regret in stochastic and adversarial regimes through separate mechanisms. The stochastic analysis was based on using empirical estimates of the gaps and high-probability control of the weights $w_{t,i}$. However, gap estimation is challenging, because the variance of $\hat{L}_{t,i}$ is of the order of $\sum_{s=1}^t \frac{1}{w_{s,i}}$. If the arms are played according to the target probabilities of $w_{t,i} \approx \frac{1}{t\Delta_i^2}$, then the variance of $(\hat{L}_{t,i} - \hat{L}_{t,i^*})$ is of the order of $\Theta(\Delta_i^2 t^2)$. This is prohibitively large, because the square root of the variance is of the same order as the expected cumulative gap and standard tools, such as Bernstein's inequality, cannot guarantee concentration of $(\hat{L}_{t,i} - \hat{L}_{t,i^*})$ around $\Delta_i t$. Seldin and Slivkins [91] have coped with this by mixing in additional exploration, but this has led to a regret growth rate of the order of $(\log T)^3$ in the stochastic regime. Seldin and Lugosi [90] have mixed in less exploration and used unweighted losses for the gap estimates, which has decreased the regret growth rate down to $(\log T)^2$. It is currently unknown whether direct gap estimation can be further improved to support the desired $\log T$ stochastic regret rates. Additionally, existing oracle analysis in Seldin and Slivkins [91, Theorem 2] and Theorem 3.3 here only support $(\log T)^2$ regret rate for EXP3-based algorithms (corresponding to $\alpha = 1$) in the stochastic regime. It is currently unknown whether this rate can be improved either. To summarize, the main breakthrough compared to this line of work is moving from $\alpha = 1$ to $\alpha = \frac{1}{2}$ and shifting from an analysis based on gap estimation to an analysis based on self-boundedness of the regret. The proposed algorithm does not mix in any additional exploration.

3.5 Additional Intermediate Regimes Between Stochastic and Adversarial

In this section we show that stochastic bandits with adversarial corruptions proposed by Lykouris et al. [76] are also a special case of an adversarial environment with a self-bounding constraint. We further propose an extension of their regime by combining it with a stochastically constrained adversary. We show that the combination is also a special case of an adversarial environment with a $(\Delta, 2C, T)$ self-bounding constraint, where TSALLIS-INF achieves logarithmic regret. We finish the section with an open question on whether TSALLIS-INF can achieve logarithmic regret guarantees in the intermediate regimes defined by Seldin and Slivkins [91].

3.5.1 Stochastic Bandits with Adversarial Corruptions

Lykouris et al. [76] have proposed a regime in which an adversary is allowed to make corruptions to an otherwise stochastic environment. Let $\overline{\mathcal{L}}_T = (\bar{\ell}_1, \dots, \bar{\ell}_T)$ and $\mathcal{L}_T = (\ell_1, \dots, \ell_T)$ be two sequences of losses, then the amount of corruption is measured by $\sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty$.

Let $\overline{\mathcal{L}}_T$ be a sequence of losses generated by a stochastically constrained adversary with best arm i^* and gaps Δ_i, and let $\mathcal{L}_T$ be its adaptively corrupted version with corruption amount bounded by C. The regret of an algorithm executed on $\mathcal{L}_T$ satisfies

$$\begin{aligned}
\overline{Reg}_T =& \max_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i^*}\right] \\
=& \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_{t,I_t} - \bar{\ell}_{t,i^*}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \bar{\ell}_{t,I_t}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_{t,i^*} - \ell_{t,i^*}\right] \\
\geq& \sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] - 2C. \qquad (3.4)
\end{aligned}$$

Thus, a stochastically constrained adversary with adversarial corruptions is an adversarial regime with a $(\Delta, 2C, T)$ self-bounding constraint. This leads to a direct corollary of Theorem 3.1, which improves upon the pseudo-regret bounds of Lykouris et al. [76] and Gupta et al. [55], the latter providing an $\mathcal{O}\left(\sum_{i \neq i^*} \frac{\log(T)}{\Delta_i} + KC\right)$ guarantee. We note that Lykouris et al. [76] and Gupta et al. [55] do not assume uniqueness of the best arm and also provide high-probability regret guarantees, but they only consider the more restricted stochastic setting with adversarial corruptions.

Corollary 3.1. *The regret of* Tsallis-INF *in a stochastically constrained adversarial environment with a unique best arm i^*, adaptively corrupted with corruption amount bounded by C satisfies*

$$\overline{Reg}_T = \mathcal{O}\left(\sum_{i \neq i^*} \frac{\log(T)}{\Delta_i} + \sqrt{\sum_{i \neq i^*} \frac{\log(T)}{\Delta_i} C}\right).$$

Remark 3.6. *We emphasise that the assumption of best arm uniqueness is on the stochastically constrained adversary* before *corruption. After the adaptive corruption it is allowed to have multiple best arms and the identity of the best arm is allowed to change.*

Proof. By equation (3.4) the self-bounding condition of Theorem 3.1 is satisfied with Δ being the vector of gaps of the underlying stochastically constrained adversary and the constant being $2C$. Thus, with RV loss estimators for $2C \leq \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}$ Tsallis-INF achieves

$$\overline{Reg}_T \leq \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + 20K\log(T) + \frac{1}{\Delta_{\min}} + \sqrt{K} + 32 + 2C$$

and otherwise

$$\overline{Reg}_t \leq 2\sqrt{\left(\left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + \frac{1}{\Delta_{\min}}\right) 2C} + 20K\log(T) + \sqrt{K} + 32\,.$$

□

3.5.2 Open Problem: The Performance in Seldin and Slivkins' Environments

Seldin and Slivkins [91] define *moderately contaminated stochastic regime* and *an adversarial regime with a gap.* In the moderately contaminated stochastic regime the adversary is allowed to change up to $\frac{t\Delta_i}{4}$ arbitrarily selected observations for a suboptimal arm i and up to $\frac{t\Delta_{\min}}{4}$ observations for the optimal arm i^* (where $\Delta_{\min} = \min_{\Delta_i>0} \Delta_i$). The logic behind the definition is that in expectation the adversary can reduce the gap Δ_i by a factor of 2, but cannot eliminate it completely. The adversarial regime with a gap is an adversarial regime, where starting from a certain time τ (unknown to the algorithm) the cumulative loss of an optimal arm maintains a certain gap Δ_τ to all other arms until the end of the game. Seldin and Slivkins show that their EXP3++ algorithm achieves "logarithmic" regret in both regimes. Note that in the moderately contaminated stochastic regime the amount of contamination is allowed to grow linearly with time. While the regime could be seen as a special case of stochastic bandits with adversarial corruptions discussed earlier, the regret bound in Corollary 3.1 only supports "logarithmic" regret for "logarithmic" amount of corruption C. So far we have been unable to obtain "logarithmic" regret guarantees for TSALLIS-INF in the intermediate regimes of Seldin and Slivkins (the analysis proposed in Zimmert and Seldin [111] is incorrect). The challenge is that the gaps are defined through cumulative rather than instantaneous quantities. Deriving "logarithmic" regret guarantees for TSALLIS-INF in these regimes is an interesting open problem.

3.6 Dueling Bandits

In the sparring approach to stochastic utility-based dueling bandits [8] each side in the sparring can be modeled as a stochastically constrained adversarial environment. This makes it a perfect application domain for TSALLIS-INF. The problem is defined by K arms with utilities $u_i \in [0, 1]$. At each round, an agent has to select two arms, I_t and J_t, to "duel". The feedback is the winner W_t of the "duel", which is chosen according to $\mathbb{P}[W_t = I_t] = \frac{1+u_{I_t}-u_{J_t}}{2}$. The regret is defined by the distance to the optimal utility:

$$\overline{Reg}_T = \sum_{t=1}^{T} 2u_{i_T^*} - \mathbb{E}\left[\sum_{t=1}^{T}(u_{I_t} + u_{J_t})\right].$$

In the adversarial version of the problem, the utilities u_i are not constant, but time dependent, $u_{t,i}$, and selected by an adversary. The regret in this case is the difference to the optimal utility in hindsight:

$$\overline{Reg}_T = \max_i \mathbb{E}\left[\sum_{t=1}^{T} 2u_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T}(u_{t,I_t} + u_{t,J_t})\right].$$

Ailon et al. [8] have proposed the SPARRING algorithm, in which two black-box MAB algorithms spar with each other. The first algorithm selects I_t and receives the loss $\ell_{t,I_t} = \mathbb{1}(W_t \neq I_t)$. The second algorithm selects J_t and receives the loss $\ell_{t,J_t} = \mathbb{1}(W_t \neq J_t)$. They have shown that the regret is the sum of individual regret values for both MABs, thereby recovering $\mathcal{O}(\sqrt{KT})$ regret in the adversarial case, if MABs with $\mathcal{O}(\sqrt{KT})$ adversarial regret bound are used. In the stochastic case each black-box MAB is a system with a stochastically constrained adversary, because the relative winning probability of the arms stays fixed, but depending on the arm choice of the sparring partner the baseline shifts up and down. Since no algorithm has been known to achieve $\log(T)$ regret in stochastically constrained adversarial setting, Ailon et al. [8] provide no analysis of SPARRING in the stochastic case. Indeed, as we demonstrate in our experiments, standard algorithms for stochastic multiarmed bandits, such as UCB or THOMPSON SAMPLING, may exhibit almost linear regret in stochastically constrained adversarial setting and, therefore, are not suitable for sparring.

By applying Theorem 3.1, we directly obtain the following corollary.

Corollary 3.2. *In a utility-based dueling bandit problem* Sparring *with two independent versions of* Tsallis-INF *suffers a regret of*

$$\overline{Reg}_T \leq \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i}\right)$$

in the stochastic case with a unique best arm and

$$\overline{Reg}_T \leq \mathcal{O}\left(\sqrt{KT}\right)$$

in the adversarial case.

3.7 Proofs

In this section, we first revise the general proof framework of Omd and provide a compact summary of how to modify it to obtain stochastic guarantees. Afterward, we provide proofs of Theorems 3.2 and 3.1. A proof of Theorem 3.3 along with proofs of all the lemmas in this section are provided in the appendix.

3.7.1 High-Level Overview of Omd Modification for Stochastic Analysis

We follow the standard Omd analysis [70, Chapter 28] and introduce the potential function $\Phi_t(-L) = \max_{w\in\Delta^{K-1}}\{\langle w, -L\rangle - \Psi_t(w)\}$ to decompose the regret into *stability* and *penalty* terms.

$$\begin{aligned}\overline{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\ell_{t,I_t} - \ell_{t,i_T^*}\right)\right]\\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{stability} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i_T^*}\right]}_{penalty}. \end{aligned} \tag{3.5}$$

The Omd analysis bounds the *stability* and *penalty* terms separately. For Tsallis-entropy regularizers, Abernethy et al. [6] have proven the following bounds

$$\begin{aligned} stability &\leq \sum_{t=1}^{T}\eta_t\sum_{i=1}^{K} f(\mathbb{E}[w_{t,i}]),\\ penalty &\leq \sum_{t=1}^{T}({\eta_{t+1}}^{-1} - {\eta_t}^{-1})\sum_{i=1}^{K} g(\mathbb{E}[w_{t,i}]), \end{aligned}$$

where $f(x)$ and $g(x)$ are proportional to $x^{1-\alpha}$ and x^{α}, respectively. Adversarial bounds that scale with $\sqrt{T}$ are obtained by applying $\sum_{i=1}^{K} f(\mathbb{E}[w_{t,i}]) \leq \max_{w\in\Delta^{K-1}}\sum_{i=1}^{K} f(w)$, $\sum_{i=1}^{K} g(\mathbb{E}[w_{t,i}]) \leq \max_{w\in\Delta^{K-1}}\sum_{i=1}^{K} g(w)$, and choosing an appropriate learning rate. In particular, for $\alpha = \frac{1}{2}$ we have $f(x) = \theta(\sqrt{x})$ and $g(x) = \theta(\sqrt{x})$ and we use $\eta_t = \theta(1/\sqrt{t})$, for which $\eta_{t+1}^{-1} - {\eta_t}^{-1} = \theta(1/\sqrt{t})$. This gives

$$\overline{Reg}_T \leq \sum_{t=1}^{T} c\frac{1}{\sqrt{t}}\sum_{i=1}^{K}\sqrt{\mathbb{E}[w_{t,i}]} \leq \sum_{t=1}^{T} c\frac{1}{\sqrt{t}}\max_{z\in\Delta^{K-1}}\sum_{i=1}^{K}\sqrt{z_i} \leq \sum_{t=1}^{T} c\frac{1}{\sqrt{t}}\sqrt{K} \leq 2c\sqrt{KT},$$

where c is a small constant and we replace $\mathbb{E}[w_{t,i}]$ with z_i in the maximisation.

The main insight of the paper is that the same framework can be used to obtain "logarithmic" bounds in the stochastic case. The key novelty is that if we constrain the maximisation of $\mathbb{E}[w_{t,i}]$ by the self-bounding property of the regret (3.3), the space of solutions excludes the worst-case scenario, where the regret grows with the square root of the time horizon. For simplicity we first explain the approach with $C = 0$. By the self-bounding property (3.3) we then have $\mathfrak{R} \geq \sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] = \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{E}[w_{t,i}]$ (since $\Delta_{i^*} = 0$ by definition), which we can use to write

$$\mathfrak{R} \leq 2\mathfrak{R} - \sum_{t=1}^{T} \sum_{i} \Delta_i \mathbb{E}[w_{t,i}]. \tag{3.6}$$

The negative contributions $-\Delta_i \mathbb{E}[w_{t,i}]$ are used to achieve better control of the growth of $\mathbb{E}[w_{t,i}]$, but they are only helpful for i with $\Delta_i > 0$, i.e., only for $i \neq i^*$. In order to exploit them we derive refined bounds for the stability and penalty terms:

$$\begin{aligned} \textit{stability} &\leq \sum_{t=1}^{T} \eta_t \sum_{i \neq i^*} \tilde{f}(\mathbb{E}[w_{t,i}]), \\ \textit{penalty} &\leq \sum_{t=1}^{T} ({\eta_{t+1}}^{-1} - {\eta_t}^{-1}) \sum_{i \neq i^*} g(\mathbb{E}[w_{t,i}]), \end{aligned}$$

where the summation excludes the best arm i^*, which has no negative contribution in (3.6). The cost of excluding the best arm is an addition of a linear term to $f : \tilde{f}(x) = f(x) + c'x$, where c' is a small constant. In particular, for $\alpha = \frac{1}{2}$ and learning rate $\eta_t = \theta(1/\sqrt{t})$ we have

$$\begin{aligned} \mathfrak{R} &\leq 2\mathfrak{R} - \sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(2c \frac{1}{\sqrt{t}} \left(\sqrt{\mathbb{E}[w_{t,i}]} + c_1 \mathbb{E}[w_{t,i}] \right) - \Delta_i \mathbb{E}[w_{t,i}] \right) \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \max_z \left(2c \frac{1}{\sqrt{t}} \left(\sqrt{z} + c_1 z \right) - \Delta_i z \right) \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{c_2}{\Delta_i t} + \frac{c_3}{\Delta_i t (\Delta_i \sqrt{t} - 1)} \right) \\ &= O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i} \right), \end{aligned}$$

where c_1, c_2, c_3 are small constants and in the second line we use the refined stability and penalty bounds to bound $2\mathfrak{R}$. The negative contribution is exploited in the maximisation in the third line, which is now done coordinate-wise and the constraint that w_t is a probability distribution is dropped.

We assume uniqueness of the zero-entry in Δ, because currently we are only able to exclude one arm from the summation in refined bounds on stability and penalty. Had there been multiple arms with $\Delta_i = 0$ they would have no negative contributions to control $\mathbb{E}[w_{t,i}]$. The challenge in excluding more than one arm from the summation is explained in Lemma 3.1, where we derive the refined bound.

In the more general analysis with $C > 0$ we write $\mathfrak{R} \leq (1+\lambda)\mathfrak{R} - \lambda \left(\sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] - C \right)$ and use the parameter λ for optimizing the dependence on C. The parameter λ can also be seen as a Lagrange multiplier in a constrained optimisation problem of maximizing the regret bound (stability bound + penalty bound) under the self-bounding constraint that (stability bound + penalty bound) $\geq \sum_{t=1}^{T}\sum_{i\neq i^*}\Delta_i\mathbb{E}[w_{t,i}] - C$.

3.7.2 Key Lemmas

The proofs of Theorems 3.2, 3.1, and 3.3 are based on the following two lemmas that bound the *stability* and *penalty* terms. The proofs of the lemmas are provided in Appendix 3.9.

Lemma 3.1. *For a positive learning rate, the instantaneous* stability *of* α-Tsallis-INF *satisfies at any time* t

$$\mathbb{E}\left[\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \leq \begin{cases} \min\left\{\sum_{i=1}^{K}\frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha}, 1\right\}, & \textit{if 1.}\\ \frac{\eta_t^2}{2} + \sum_{i=1}^{K}\frac{\eta_t}{2}\mathbb{E}[w_{t,i}]^{\frac{1}{2}}(1-\mathbb{E}[w_{t,i}]), & \textit{if 2.}\\ \frac{5\eta_t^2}{8}K + \sum_{i=1}^{K}\frac{\eta_t}{8}\mathbb{E}[w_{t,i}]^{\frac{1}{2}}(1-\mathbb{E}[w_{t,i}]), & \textit{if 3.}\\ \sum_{i\neq j}\left(\frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha} + \frac{\eta_t(\xi_i+2\xi_j)}{2}\mathbb{E}\left[w_{t,i}\right]\right), & \textit{if 4.}, \end{cases}$$

where

1. $\hat{L}_t$ *is based on importance weighted estimators. The inequality holds for any* $\eta_t > 0$ *and* $\alpha \in [0,1]$.
2. $\hat{L}_t$ *is based on importance weighted estimators,* $1 \geq \eta_t > 0$, *and* $\alpha = \frac{1}{2}$.
3. $\hat{L}_t$ *is based on reduced-variance estimators,* $1 \geq \eta_t > 0$, *and* $\alpha = \frac{1}{2}$.
4. $\hat{L}_t$ *is based on importance weighted estimators and* $\eta_t\xi_i \leq \frac{1}{4}$ *for all* i. *The inequality holds for any* j *and* $\alpha \in [0,1]$.

The first part of the Lemma is due to Abernethy et al. [6]. The remaining parts are non-trivial refinements that are crucial for our analysis, as outlined in the previous section. The first inequality is used in the proof of Theorem 3.2, the second and third inequalities are used for the two results in Theorem 3.1, the last inequality is used in the proof of Theorem 3.3. In the proof of Theorem 3.1 we use $\mathbb{E}[w_{t,i}]^{\frac{1}{2}}(1-\mathbb{E}[w_{t,i}]) \leq \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$ for $i \neq i^*$, and for i^* we use $\mathbb{E}[w_{t,i^*}]^{\frac{1}{2}}(1-\mathbb{E}[w_{t,i^*}) \leq (1-\mathbb{E}[w_{t,i^*}]) = \sum_{i\neq i^*}\mathbb{E}[w_{t,i}]$. This eliminates $\mathbb{E}[w_{t,i^*}]$ from the regret bound and allows to exploit the self-bounding property. This approach only allows to eliminate one arm from the regret bound, which is the reason we rely on the assumption of uniqueness of the best arm.

Lemma 3.2. *For any* $\alpha \in [0,1]$ *and any unbiased loss estimators the* penalty *term of* α-Tsallis-INF *satisfies:*

1. *For the symmetric regulariser and non-increasing learning rate*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \leq \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{(1-\alpha)\alpha\eta_T} + 1.$$

2. *For an arbitrary regularizer, non-increasing learning rate, and any* $x \in [1,\infty]$

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right]$$

$$\leq \frac{1-T^{-\alpha x}}{\alpha}\sum_{i\neq i_T^*}\left(\frac{\mathbb{E}[w_{1,i}]^\alpha - \alpha\mathbb{E}[w_{1,i}]}{\eta_1\xi_i(1-\alpha)} + \sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\frac{\mathbb{E}[w_{t,i}]^\alpha - \alpha\mathbb{E}[w_{t,i}]}{\xi_i(1-\alpha)}\right) + T^{1-x}.$$

The first part of the Lemma is a straightforward improvement of the penalty bound in Abernethy et al. [6] with the techniques from Agarwal et al. [7]. The second part is again a crucial refinement. It is obtained by exploiting the negative contribution of $\Psi_T(\mathbf{e}_{i^*})$ in an intermediate step of the proof, which Abernethy et al. [6] trivially bounded by 0.

3.7.3 Proofs of Theorems 3.1 and 3.2

Now we are ready to present the proofs of the main theorems.

Proof of Theorem 3.1. We provide a proof of regret bounds for Tsallis-INF with reduced-variance (RV) estimators. The analysis of Tsallis-INF with IW estimators in the adversarial case is analogous to the proof of Theorem 3.2 and under the self-bounding constraint (3.3) it is analogous to the analysis of RV estimators with the bound in Part 3 of Lemma 3.1) replaced by the bound in Part 2. Therefore, the proofs of both results for the IW estimators are omitted.

To analyse the regret we start by bounding the *stability* term. We use Lemma 3.1, where for $t < 16$ we have $\eta_t > 1$ and the RV estimators are equivalent to IW estimators. Thus, we can apply the first part of the lemma to bound the instantaneous stability by 1. For $t \geq 16$, we use the third part of the lemma.

$$\begin{aligned}
\textit{stability} &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \\
&\leq 15 + \sum_{t=16}^{T}\left(\frac{5\eta_t^2}{8}K + \sum_{i=1}^{K}\frac{\eta_t}{8}\sqrt{\mathbb{E}[w_{t,i}]}(1-\mathbb{E}[w_{t,i}])\right) \\
&\leq 15 + 10K\log(T) + \sum_{t=16}^{T}\sum_{i=1}^{K}\frac{\sqrt{\mathbb{E}[w_{t,i}]}(1-\mathbb{E}[w_{t,i}])}{2\sqrt{t}}\,. \qquad (3.7)
\end{aligned}$$

Adversarial bound. We bound $\sum_{i=1}^{K}\sqrt{\mathbb{E}[w_{t,i}]}(1-\mathbb{E}[w_{t,i}]) \leq \sum_{i=1}^{K}\sqrt{\mathbb{E}[w_{t,i}]} \leq \sqrt{K\sum_{i=1}^{K}\mathbb{E}[w_{t,i}]} = \sqrt{K}$, which holds by Jensen's inequality. Then we have

$$\begin{aligned}
\textit{stability} &\leq 15 + 10K\log(T) + \sum_{t=16}^{T}\frac{\sqrt{K}}{2\sqrt{t}} \\
&\leq 15 + 10K\log(T) + \sqrt{KT}\,.
\end{aligned}$$

For the *penalty* term, we use the first part of Lemma 3.2 to obtain

$$\textit{penalty} \leq \sqrt{KT} + 1\,.$$

Combining *stability* and *penalty* completes the proof.

Bound under the self-bounding constraint (3.3). We continue bounding the *stability* up from equation (3.7). For $i \neq i^*$ we use $\sqrt{\mathbb{E}[w_{t,i}]}(1-\mathbb{E}[w_{t,i}]) \leq \sqrt{\mathbb{E}[w_{t,i}]}$. For i^* we use $\sqrt{\mathbb{E}[w_{t,i^*}]}(1-\mathbb{E}[w_{t,i^*}]) \leq (1-\mathbb{E}[w_{t,i^*}]) = \sum_{i\neq i^*}\mathbb{E}[w_{t,i}]$. For a constant $0 < \lambda \leq 1$ that will be specified at a later stage of the proof and $t \leq T_0 = \left\lceil (\frac{1}{2\lambda\Delta_{\min}})^2 \right\rceil$ we further bound the last expression as $\sum_{i\neq i^*}\mathbb{E}[w_{t,i}] \leq 1$. Altogether, this gives

$$\begin{aligned}
\sum_{t=16}^{T}\sum_{i=1}^{K}\frac{\sqrt{\mathbb{E}[w_{t,i}]}(1-\mathbb{E}[w_{t,i}])}{2\sqrt{t}} &\leq \sum_{t=16}^{T_0}\frac{1}{2\sqrt{t}} + \sum_{i\neq i^*}\left(\sum_{t=T_0+1}^{T}\frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}} + \sum_{t=16}^{T}\frac{\sqrt{\mathbb{E}[w_{t,i}]}}{2\sqrt{t}}\right) \\
&\leq \sqrt{T_0} + \sum_{i\neq i^*}\left(\sum_{t=1}^{T}\frac{\sqrt{\mathbb{E}[w_{t,i}]}}{2\sqrt{t}} + \sum_{t=T_0+1}^{T}\frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right)
\end{aligned}$$

and

$$\mathit{stability} \leq 15 + 10K\log(T) + \sqrt{T_0} + \sum_{i\neq i^*}\left(\sum_{t=1}^{T} \frac{\sqrt{\mathbb{E}[w_{t,i}]}}{2\sqrt{t}} + \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right).$$

In order to bound the *penalty* term, we use the second part of Lemma 3.2 with $x = \infty$. We drop the linear terms since they are all negative. Note that $2\sqrt{T} = 1 + \int_{t=1}^{T} \frac{1}{\sqrt{t}}\,dt \leq 1 + \sum_{t=1}^{T} \frac{1}{\sqrt{t}}$.

$$\begin{aligned}
\mathit{penalty} &= \mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i^*}\right] \\
&\leq 4\sum_{i\neq i^*}\left(\frac{\sqrt{\mathbb{E}[w_{1,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]}{\eta_1} + \sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\left(\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]\right)\right) \\
&= \sum_{i\neq i^*}\left(\left(\sqrt{\mathbb{E}[w_{1,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]\right) + \sum_{t=2}^{T}\left(\sqrt{t} - \sqrt{t-1}\right)\left(\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]\right)\right) \\
&= \sum_{t=1}^{T}\left(\sum_{i\neq i^*}\frac{\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right) + \sum_{i\neq i^*}\left(\frac{\sqrt{\mathbb{E}[w_{1,i}]} - \frac{1}{2}\mathbb{E}[w_{1,i}]}{2}\right. \\
&\qquad\qquad \left. + \sum_{t=2}^{T}\left(\sqrt{t} - \sqrt{t-1} - \frac{1}{2\sqrt{t}}\right)\left(\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]\right)\right) \\
&\leq \sum_{t=1}^{T}\left(\sum_{i\neq i^*}\frac{\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right) + \left(\frac{1}{2} + \sum_{t=2}^{T}\left(\sqrt{t} - \sqrt{t-1} - \frac{1}{2\sqrt{t}}\right)\right)\sqrt{K} \\
&= \sum_{t=1}^{T}\left(\sum_{i\neq i^*}\frac{\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right) + \left(\sqrt{T} - \sum_{t=1}^{T}\frac{1}{2\sqrt{t}}\right)\sqrt{K} \\
&\leq \sum_{t=1}^{T}\left(\sum_{i\neq i^*}\frac{\sqrt{\mathbb{E}[w_{t,i}]} - \frac{1}{2}\mathbb{E}[w_{t,i}]}{2\sqrt{t}}\right) + \frac{\sqrt{K}}{2}.
\end{aligned}$$

Combining *penalty* and *stability* gives the bound

$$\overline{\mathit{Reg}}_T \leq \sum_{i\neq i^*}\left(\sum_{t=1}^{T}\frac{\sqrt{\mathbb{E}[w_{t,i}]}}{\sqrt{t}} + \sum_{t=T_0+1}^{T}\frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}}\right) + \sqrt{T_0} + \underbrace{\frac{\sqrt{K}}{2} + 15 + 10K\log(T)}_{=:M}.$$

By using the self-bounding property (3.3) and $(1+\lambda) \leq 2$ we obtain

$$\begin{aligned}
\overline{Reg}_T &\leq \overline{Reg}_T + \lambda \left(\overline{Reg}_T - \sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] + C \right) \\
&\leq \sum_{i \neq i^*} \left(\sum_{t=1}^{T} \frac{2\sqrt{\mathbb{E}[w_{t,i}]}}{\sqrt{t}} + \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}} \right) + 2\sqrt{T_0} + 2M - \lambda \sum_{t=1}^{T} \sum_{i \neq i^*} \Delta_i \mathbb{E}[w_{t,i}] + \lambda C \\
&= \sum_{t=1}^{T_0} \sum_{i \neq i^*} \left(\frac{2\sqrt{\mathbb{E}[w_{t,i}]}}{\sqrt{t}} - \lambda \Delta_i \mathbb{E}[w_{t,i}] \right) + \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \left(\frac{2\sqrt{\mathbb{E}[w_{t,i}]} + \frac{1}{2}\mathbb{E}[w_{t,i}]}{\sqrt{t}} - \lambda \Delta_i \mathbb{E}[w_{t,i}] \right) \\
&\qquad + 2\sqrt{T_0} + 2M + \lambda C \\
&\leq \sum_{t=1}^{T_0} \sum_{i \neq i^*} \max_{z \geq 0} \left(\frac{2\sqrt{z}}{\sqrt{t}} - \lambda \Delta_i z \right) + \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \max_{z \geq 0} \left(\frac{2\sqrt{z} + \frac{1}{2}z}{\sqrt{t}} - \lambda \Delta_i z \right) \\
&\qquad + 2\sqrt{T_0} + 2M + \lambda C.
\end{aligned}$$

Simple optimisation shows that $\max_{z>0} 2\alpha\sqrt{z} - \beta z = \frac{\alpha^2}{\beta}$. Thus, we have

$$\max_{z \geq 0} \frac{2\sqrt{z}}{\sqrt{t}} - \lambda \Delta_i z = \frac{1}{\lambda \Delta_i t}.$$

and

$$\begin{aligned}
\max_{z \geq 0} \frac{2\left(\sqrt{z} + \frac{1}{4}z\right)}{\sqrt{t}} - \lambda \Delta_i z &= \frac{1}{(\lambda \Delta_i - \frac{1}{4\sqrt{t}})t} \\
&= \frac{1}{\lambda \Delta_i t} + \frac{1}{(\lambda \Delta_i - \frac{1}{4\sqrt{t}})t} - \frac{1}{\lambda \Delta_i t} \\
&= \frac{1}{\lambda \Delta_i t} + \frac{1}{4\lambda^2 \Delta_i^2 t^{\frac{3}{2}} - \lambda \Delta_i t}.
\end{aligned}$$

In order to bound the summation of the above terms we use the following bound from Lemma 3.4 in the appendix:

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{b\sqrt{T_0} - c}.$$

By definition of T_0 we have $\frac{1}{2\lambda\Delta_{\min}} \leq \sqrt{T_0} \leq \frac{1}{2\lambda\Delta_{\min}} + 1$ and

$$\frac{1}{2\lambda^2 \Delta_i^2 \sqrt{T_0} - \frac{1}{2}\lambda\Delta_i} = \frac{2}{\lambda\Delta_i} \cdot \frac{1}{4\lambda\Delta_i\sqrt{T_0} - 1} \leq \frac{2}{\lambda\Delta_i} \cdot \frac{1}{2\frac{\Delta_i}{\Delta_{\min}} - 1} \leq \frac{2}{\lambda\Delta_i}.$$

By plugging the calculations into the regret bound above we obtain:

$$\begin{aligned}
\overline{Reg}_T &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{1}{\lambda \Delta_i t} + \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \frac{1}{4\lambda^2 \Delta_i^2 t^{\frac{3}{2}} - \lambda \Delta_i t} + 2\sqrt{T_0} + 2M + \lambda C \\
&\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{1}{\lambda \Delta_i t} + \frac{1}{2\lambda^2 \Delta_i^2 \sqrt{T_0} - \frac{1}{2}\lambda\Delta_i} + 2\sqrt{T_0} + 2M + \lambda C \\
&\leq \sum_{i \neq i^*} \frac{\log(T) + 3}{\lambda \Delta_i} + \frac{1}{\lambda \Delta_{\min}} + 2(M+1) + \lambda C.
\end{aligned}$$

Finally choosing $\lambda = \min\left\{1, \sqrt{\left(\sum_{i\neq i^*} \frac{\log(T)+3}{\Delta_i} + \frac{1}{\Delta_{\min}}\right)/C}\right\}$ completes the proof.

□

Proof of Theorem 3.2. We start from equation (3.5). Since the regularisation is symmetric, we have $\xi_i = 1$ for all i. Using Lemma 3.1 we bound the *stability* term as

$$\begin{aligned} \textit{stability} &= \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \leq \sum_{t=1}^T \sum_{i=1}^K \frac{\eta_t}{2} \mathbb{E}\left[w_{t,i}\right]^{1-\alpha} \\ &\leq \left(\sum_{t=1}^T \frac{\eta_t}{2}\right) \max_{w\in\Delta^{K-1}} \sum_{i=1}^K w_i^{1-\alpha} = \left(\sum_{t=1}^T \sqrt{\frac{K^{1-2\alpha}-K^{-\alpha}}{1-\alpha}\frac{1-t^{-\alpha}}{\alpha t}}\right)\frac{K^\alpha}{2} \\ &\leq \left(\sum_{t=1}^T \sqrt{\frac{1-K^{\alpha-1}}{1-\alpha}\frac{1-T^{-\alpha}}{\alpha t}}\right)\frac{\sqrt{K}}{2} \leq \sqrt{\frac{1-K^{\alpha-1}}{1-\alpha}\frac{1-T^{-\alpha}}{\alpha}KT}. \end{aligned}$$

The *penalty* is bounded according to Lemma 3.2

$$\begin{aligned} \textit{penalty} &= \mathbb{E}\left[\sum_{t=1}^T -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i_T^*}\right] \\ &\leq \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{(1-\alpha)\alpha\eta_T} + 1 = \sqrt{\frac{1-K^{\alpha-1}}{1-\alpha}\frac{1-T^{-\alpha}}{\alpha}KT} + 1. \end{aligned}$$

The proof is completed by noting that the first factor is bounded by $\sqrt{\frac{1}{1-\alpha}}$ and monotonically increasing in α with the limit $\lim_{\alpha\to 1}\sqrt{\frac{1-K^{\alpha-1}}{1-\alpha}} = \sqrt{\log(K)}$ (details in Lemma 3.3 in the appendix). By the same argument, the second factor is bounded by $\sqrt{\frac{1}{\alpha}}$ and monotonically decreasing in α with the limit $\lim_{\alpha\to 0}\sqrt{\frac{1-T^{-\alpha}}{\alpha}} = \sqrt{\log(T)}$.

□

3.8 Experiments

We provide an empirical comparison of Tsallis-INF with the classical algorithms for stochastic bandits, Ucb1 [16, with parameter $\alpha = 1.5$] and Thompson Sampling [96][5], and the classical algorithm for adversarial bandits, Exp3, implemented for the losses [26]. We also compare with the state-of-the-art algorithms for stochastic and adversarial bandits, EXP3++ with parametrisation proposed in Seldin and Lugosi [90] and Broad [103]. The pseudo-regret is estimated by 100 repetitions of the corresponding experiments and two standard deviations of the empirical pseudo-regret, $\sum_{t=1}^T \Delta_{I_t}$, over the 100 repetitions are depicted by the shaded areas on the plots. We always show the first 10000 time steps on a linear plot and then the time steps from 10^4 to 10^7 on a separate log-log plot.

The first experiment, shown in Figures 3.1 and 3.2, is a standard stochastic MAB, where the mean rewards are $(1+\Delta)/2$ for the single optimal arm and $(1-\Delta)/2$ for all the suboptimal arms. The number of arms K and the gaps Δ are varied as described in the figures. Unsurprisingly, Thompson Sampling exhibits the lowest regret, but Tsallis-INF takes a confident second

[5] Another leading stochastic algorithm, KL-UCB [32], has performed comparably to Thompson Sampling in our experiments and, therefore, is not reported in the figures.

place, outperforming all other competitors. UCB1, EXP3, and EXP3++ fall roughly in the same league, while BROAD suffers from extremely large constant factors and is out of question for practical applications.

The second experiment, shown in Figures 3.3 and 3.4, simulates stochastically constrained adversaries. The mean loss of (optimal arm, all sub-optimal arms) switches between $(1-\Delta, 1)$ and $(0, \Delta)$, while staying unchanged for phases that are increasing exponentially in length. Both UCB1 and THOMPSON-SAMPLING suffer almost linear regret. To the best of our knowledge, this is the first empirical evidence clearly demonstrating that THOMPSON SAMPLING is unsuitable for the adversarial regime. All other algorithms are almost unaffected by the shifting of the means, with TSALLIS-INF being the only algorithm that achieves $\log(T)$ regret with practical constant factors.

3.8.1 Multiple Optimal Arms

Since our theoretical results for the stochastic setting do not include multiple optimal arms, we explore this setting empirically. We use a single suboptimal arm with mean loss of 9/16. All other arms are optimal with mean loss 7/16. We run the experiment with 1000 repetitions and increase the number of arms. Figure 3.5 clearly shows that the regret does not suffer if the optimal arm is not unique. Moreover, we observe that the regret decreases with the growth of the number of suboptimal arms. Therefore, we conjecture that the requirement of uniqueness is merely an artifact of the analysis.

3.9 Discussion

We have presented a general analysis of online mirror descent algorithms regularised by Tsallis entropy with $\alpha \in [0, 1]$. As the main contribution, we have shown that the special case of $\alpha = \frac{1}{2}$ achieves optimality in both adversarial and stochastic regimes, while being oblivious to the environment at hand. Thereby, we have closed logarithmic gaps to lower bounds, which were present in existing best-of-both-worlds algorithms. We introduced a novel proof technique based on the self-bounding property of the regret, circumventing the need of controlling the variance of loss estimates. We have provided an empirical evidence that our algorithm outperforms UCB1 in stochastic environments and is significantly more robust than UCB1 and THOMPSON SAMPLING in non-i.i.d. settings. We have introduced an adversarial regime with a self-bounding constraint, which includes stochastically constrained adversaries and adversarially corrupted stochastic bandits as special cases and improved regret bounds for the latter two regimes. We have also shown that TSALLIS-INF can be applied to achieve stochastic and adversarial optimality in utility-based dueling bandits.

A weak point of the current analysis is the assumption of uniqueness of the zero entry in a vector of suboptimality gaps in the adversarial regime with a self-bounding constraint. In the stochastic and stochastically constrained adversarial settings it corresponds to assumption of uniqueness of the best arm. Our experiments suggest that this is most likely an artifact of the analysis and we aim to address this shortcoming in future work.

Another open question is whether it is possible to remove the remaining factor 2 in the asymptotic gap-dependent stochastic bound. A complimentary open question is the possibility of derivation of tighter lower bounds for the adversarial regime with a self-bounding constraint.

One more open question is whether logarithmic regret is achievable by TSALLIS-INF in the intermediate regimes defined by Seldin and Slivkins [91]. We have discussed this question in more details in Section 3.5.2.

An additional direction for future research is application of TSALLIS-INF to other problems. The fact that the algorithm relies solely on importance weighted losses makes it a suitable

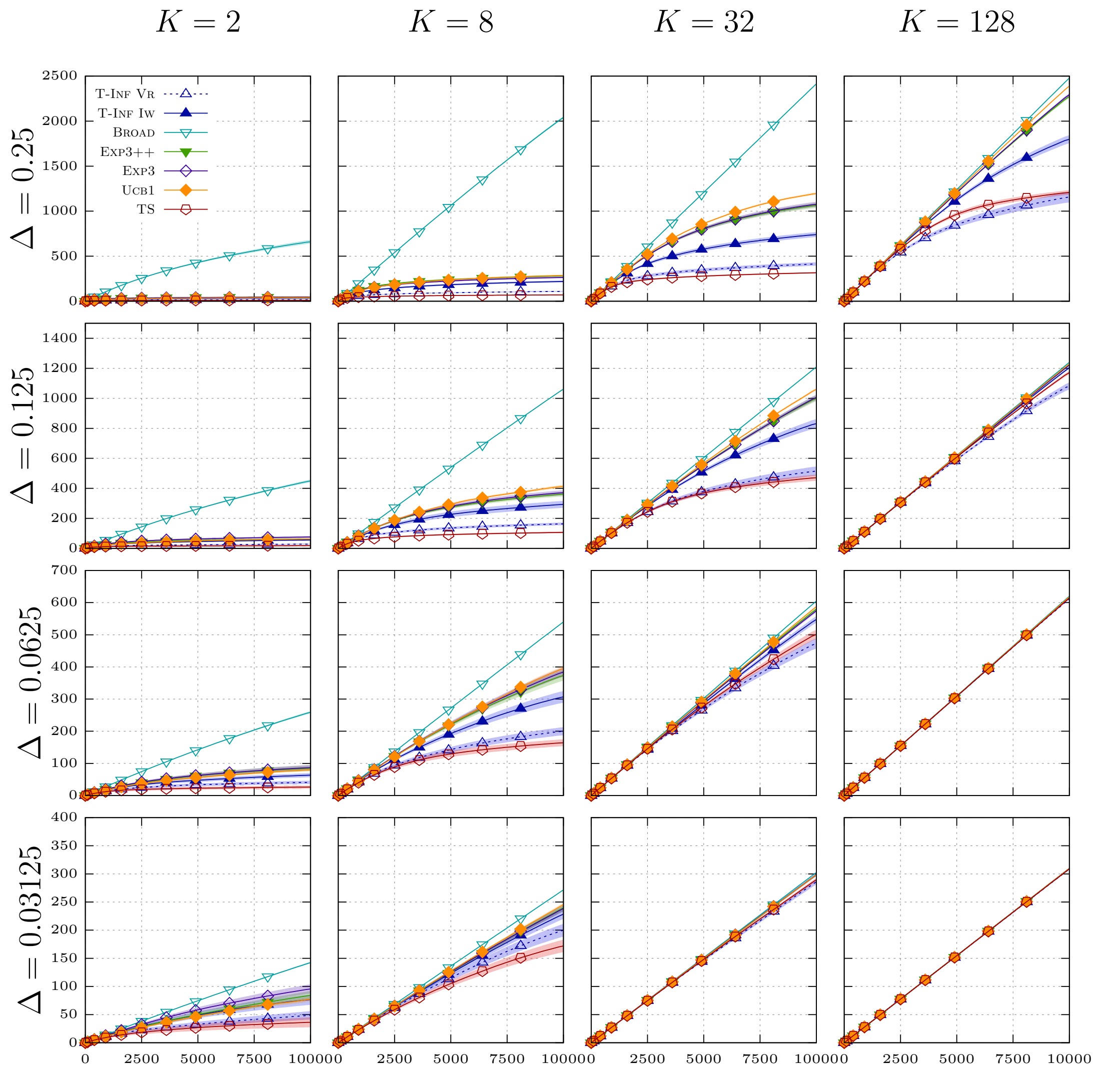


Figure 3.1: Comparison of Tsallis-INF with Thompson Sampling, Ucb1, Exp3, Exp3++, and Broad in a stochastic environment with fixed mean losses of $\frac{1-\Delta}{2}$ for the optimal arm and $\frac{1+\Delta}{2}$ for all sub-optimal arms. The experiment is repeated for different number of arms K and different gaps Δ. The figure shows the first 10000 time steps on a linear plot. The pseudo-regret is estimated by 100 repetitions and we depict 2 standard deviations of the empirical pseudo-regret by the shaded areas.

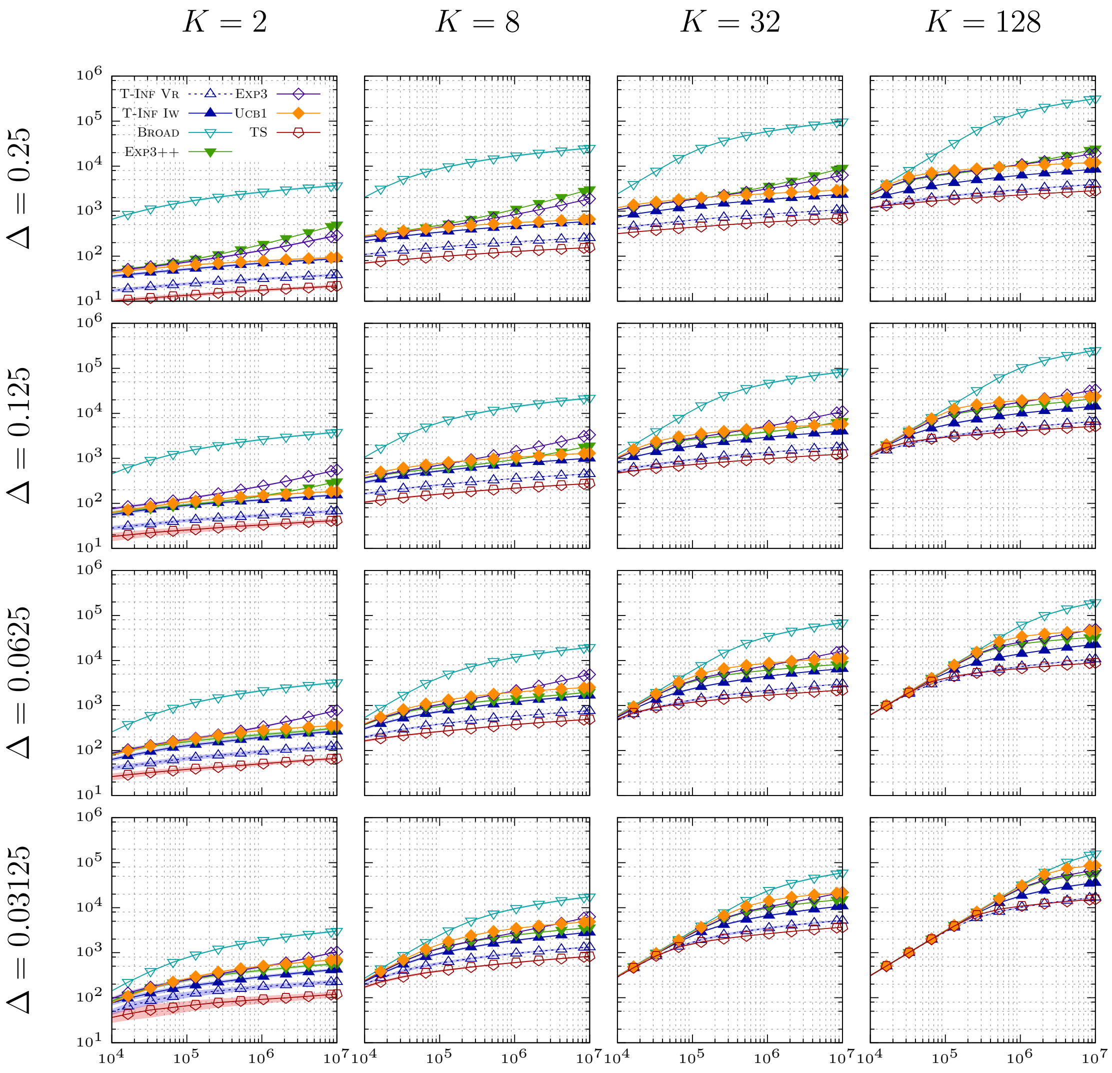


Figure 3.2: Comparison of Tsallis-INF with Thompson Sampling, Ucb1, Exp3, Exp3++, and Broad in a stochastic environment with fixed mean losses of $\frac{1-\Delta}{2}$ for the optimal arm and $\frac{1+\Delta}{2}$ for all sub-optimal arms. The experiment is repeated for different number of arms K and different gaps Δ. The figure shows the time steps from 10^4 to 10^7 on a log-log plot. The pseudo-regret is estimated by 100 repetitions and we depict 2 standard deviations of the empirical pseudo-regret by the shaded areas.

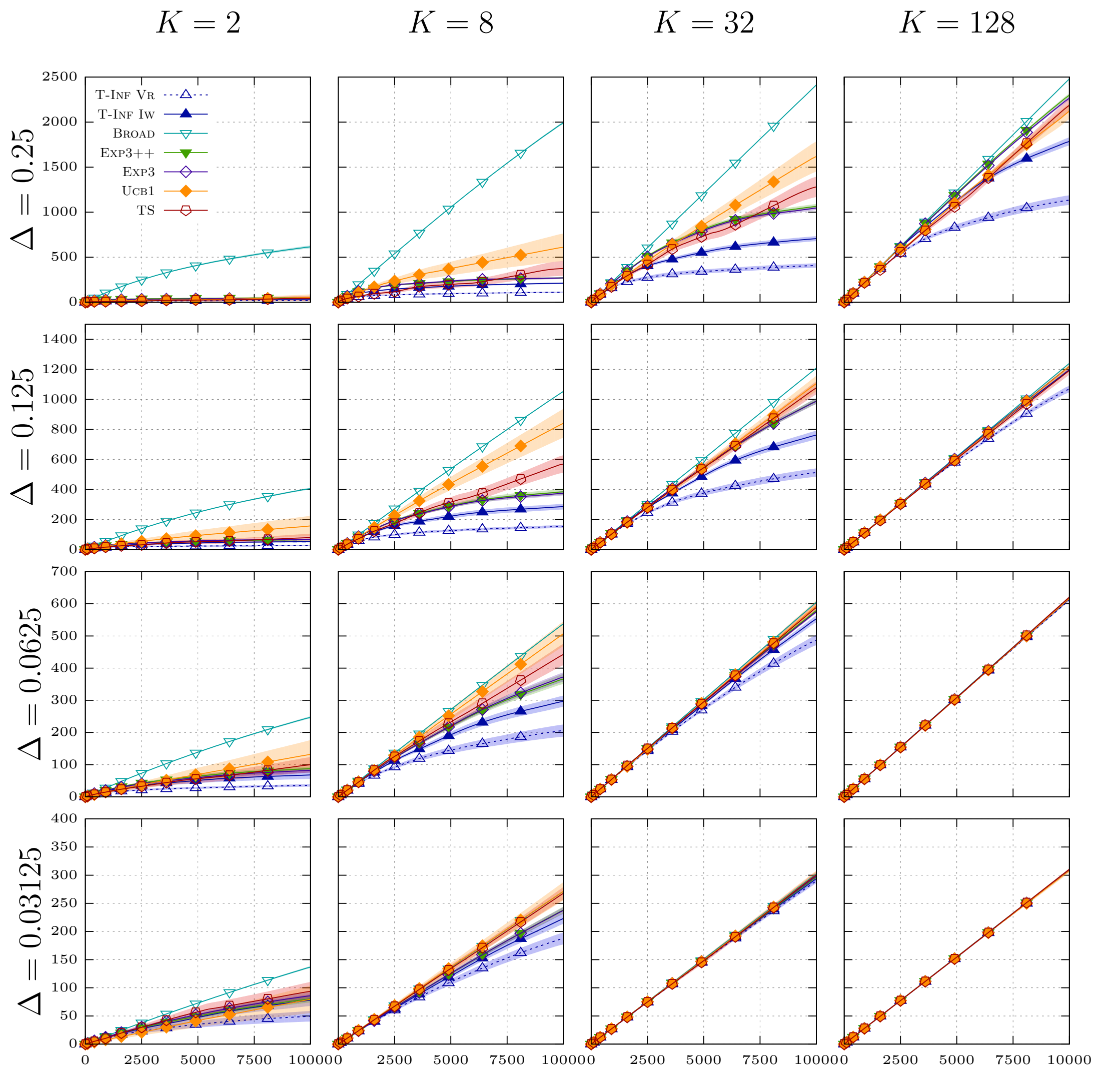


Figure 3.3: Comparison of Tsallis-INF with Thompson Sampling, Ucb1, Exp3, Exp3++, and Broad in a stochastically constrained adversarial environment. The environment (unknown to the agent) alternates between two stochastic settings. In the first setting the expected loss of the optimal arm is 0 and Δ for sub-optimal arms. In the second the expected losses are $1 - \Delta$ and 1, respectively. The time between alternations increases exponentially (with factor 1.6) after each switch. The experiment is repeated for different number of arms K and different gaps Δ. The figure shows the first 10000 time steps on a linear plot. The pseudo-regret is estimated by 100 repetitions and we depict 2 standard deviations of the empirical pseudo-regret by the shaded areas.

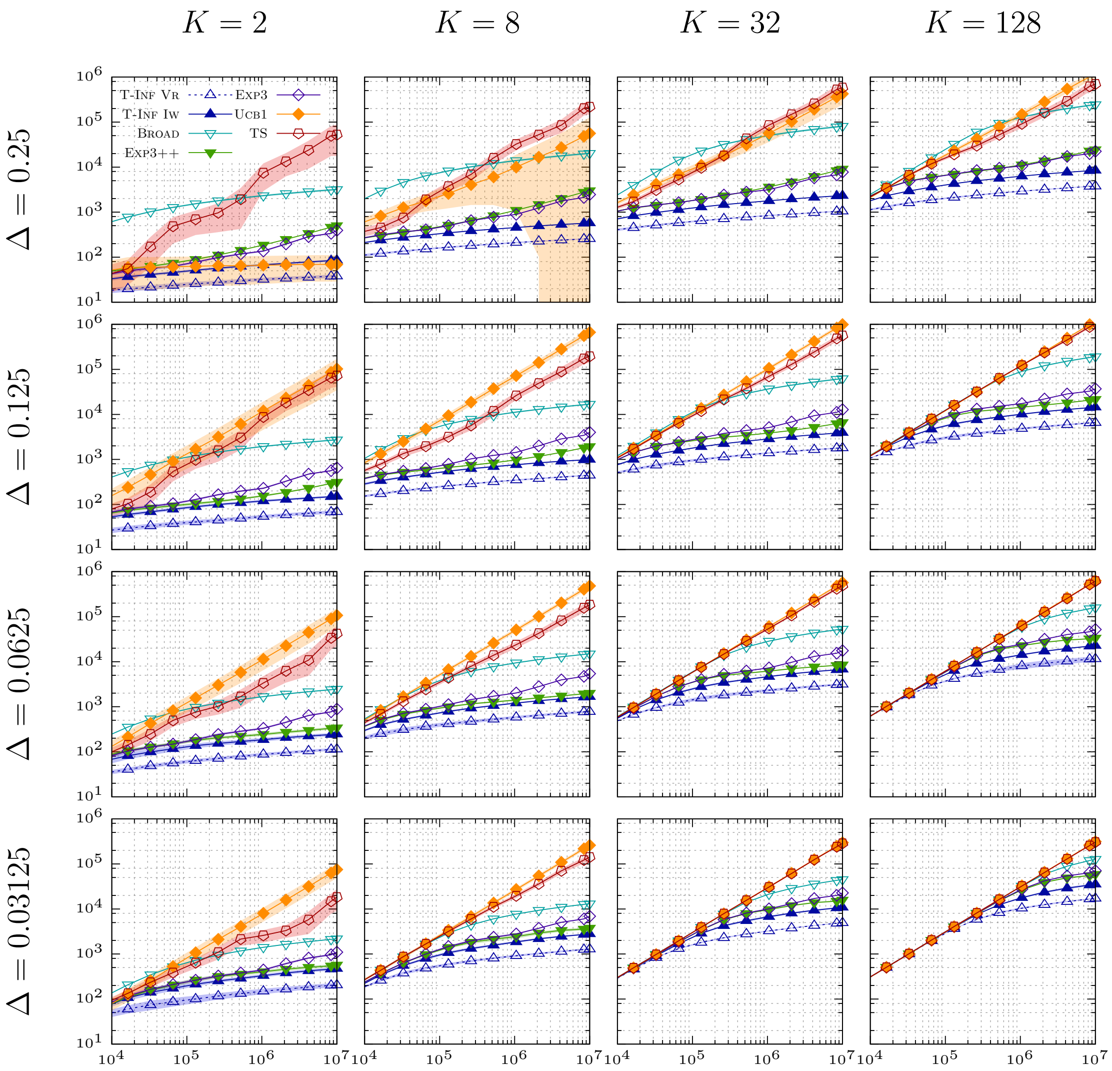


Figure 3.4: Comparison of Tsallis-INF with Thompson Sampling, Ucb1, Exp3, Exp3++, and Broad in a stochastically constrained adversarial environment. The environment (unknown to the agent) alternates between two stochastic settings. In the first setting the expected loss of the optimal arm is 0 and Δ for sub-optimal arms. In the second the expected losses are $1 - \Delta$ and 1 respectively. The time between alternations increases exponentially (with factor 1.6) after each switch. The experiment is repeated for different number of arms K and different gaps Δ. The figure shows the time steps from 10^4 to 10^7 on a log-log plot. The pseudo-regret is estimated by 100 repetitions and we depict 2 standard deviations of the empirical pseudo-regret by the shaded areas.

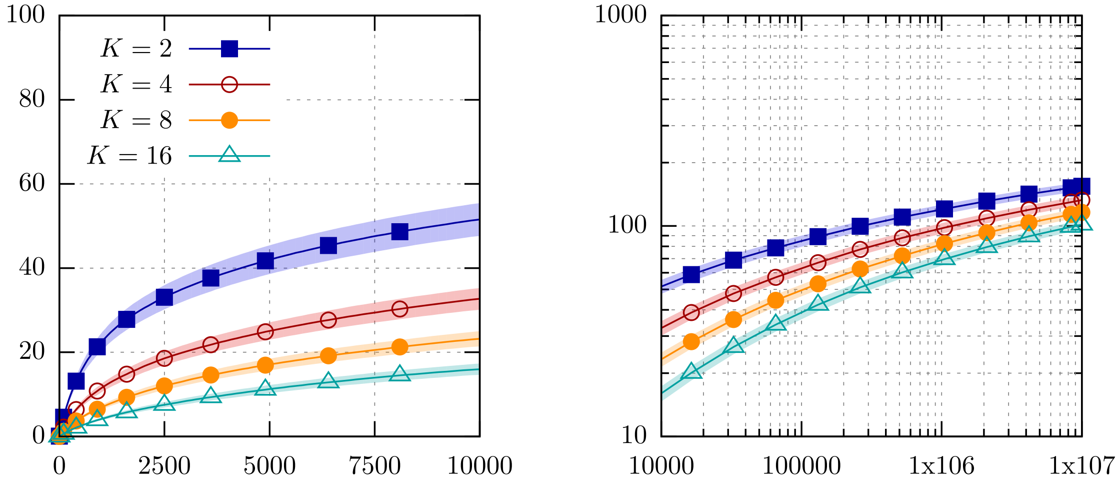


Figure 3.5: Increasing number of copies of the best arm

candidate for partial monitoring games. One step in this direction has already been taken by Zimmert et al. [109].

Acknowledgments We would like to thank Chloé Rouyer for pointing out several bugs in the previous version of the work [111] and Haipeng Luo for the idea on how to improve our regret bounds for stochastic bandits with adversarial corruptions in the large C case.

Appendix

Asymptotic Lower Bound

If the optimal arm has mean reward $\frac{1}{2}$ and suboptimal arms have the gaps Δ_i then the following lower bound for any consistent algorithm follows from Lai and Robbins [67, Theorem 2]

$$\lim_{t\to\infty}\frac{\overline{Reg}_t}{\log(t)} \geq \sum_{i:\Delta_i>0}\frac{\Delta_i}{\mathrm{kl}(\frac{1}{2}+\Delta_i,\frac{1}{2})}.$$

For any $\Delta_i \in [0, 0.5]$ the kl term can be upper bounded as

$$\mathrm{kl}(\frac{1}{2}+\Delta_i,\frac{1}{2}) \leq 2\Delta_i^2 + 3\Delta_i^3,$$

which can be verified by taking Taylor's expansion at $\Delta_i = 0$. Therefore,

$$\begin{aligned}\sum_{i:\Delta_i>0}\frac{\Delta_i}{\mathrm{kl}(\frac{1}{2}+\Delta_i,\frac{1}{2})} &\geq \sum_{i:\Delta_i>0}\frac{1}{2\Delta_i+3\Delta_i^2}\\ &= \frac{1}{2}\sum_{i:\Delta_i>0}\frac{1}{\Delta_i} - \sum_{i:\Delta_i>0}\frac{\frac{3}{2}\Delta_i}{2\Delta_i+3\Delta_i^2} \geq \frac{1}{2}\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i} - \frac{3}{2}K\right).\end{aligned}$$

Thus, for any consistent algorithm we obtain

$$\lim_{||\Delta||\to 0}\left(\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\right)^{-1}\liminf_{t\to\infty}\frac{\mathbb{E}\left[\overline{Reg}_t\right]}{\log(t)}\right) \geq \lim_{||\Delta||\to 0}\left(\frac{1}{2}-\frac{3}{4}\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\right)^{-1}K\right) = \frac{1}{2},$$

since $\lim_{||\Delta||\to 0}\left(\sum_{i:\Delta_i>0}\Delta_i^{-1}\right)^{-1}K = 0$.

Technical Lemmas

Lemma 3.3. *For any $y > 0, x > 0$, the function $\frac{1-y^{-x}}{x}$ is non-increasing in x and has the limit*

$$\lim_{x\to 0} \frac{1-y^{-x}}{x} = \log(y),$$

therefore, $\frac{1-y^{-x}}{x} \leq \min\{x^{-1}, \log(y)\}$.

Proof. Taking the derivative and using the inequality $z \leq e^z - 1$:

$$\frac{\partial}{\partial x}\left(\frac{1-y^{-x}}{x}\right) = \frac{\log(y)y^{-x}x - (1-y^{-x})}{x^2} \leq \frac{(e^{\log(y)x}-1)y^{-x} - (1-y^{-x})}{x^2} = 0.$$

The limit by L'Hôpital's rule is

$$\lim_{x\to 0} \frac{1-y^{-x}}{x} = \lim_{x\to 0} \frac{\log(y)y^{-x}}{1} = \log(y).$$

□

Lemma 3.4. *For any $b, c > 0$ and $T_0, T \in \mathbb{N}$ such that $T_0 < T$ and $b\sqrt{T_0} > c$, it holds that*

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{b\sqrt{T_0} - c}\,.$$

Proof. In the domain $(c^2/b^2, \infty)$, the function $f(t) = \frac{1}{bt^{\frac{3}{2}}-ct}$ is positive, monotonically decreasing and has the antiderivative

$$F(t) = \frac{1}{c}\left(2\log(b\sqrt{t}-c) - \log(t)\right)\,,$$

which can be verified by taking the respective derivatives. Therefore, we can bound

$$\sum_{t=T_0+1}^{T} f(t) \leq \int_{T_0}^{\infty} f(t)\,dt = \lim_{t\to\infty} F(t) - F(T_0) = \frac{1}{c}\left(2\log(b) - 2\log(b\sqrt{T_0}-c) + \log(T_0)\right)$$
$$= \frac{2}{c}\log\left(\frac{b\sqrt{T_0}}{b\sqrt{T_0}-c}\right) \leq \frac{2}{c}\left(\frac{b\sqrt{T_0}}{b\sqrt{T_0}-c} - 1\right) = \frac{2}{b\sqrt{T_0}-c}\,.$$

□

Lemma 3.5. *For any $\alpha \in [0,1]$ and $z \geq 1$, it holds that*

$$\frac{1-z^{-1+\alpha}}{1-\alpha} \leq \log(z)^{\alpha}\,.$$

Proof. The cases $\alpha \in \{0,1\}$ are trivial, since

$$\lim_{\alpha\to 0} \frac{1-z^{-1+\alpha}}{1-\alpha} = 1 - z^{-1} < \log(z)^0$$
$$\lim_{\alpha\to 1} \frac{1-z^{-1+\alpha}}{1-\alpha} = \log(z) = \log(z)^1\,.$$

We only need to verify the statement for $\alpha \in (0,1)$. Consider the function

$$f(z) = \log(z)^{\alpha} - \frac{1-z^{-1+\alpha}}{1-\alpha}\,.$$

The function is continuous for $z \geq 1$, takes the value 0 at $z = 1$ and goes to infinity for $z \to \infty$. If there is a point where the function is negative, there must also exist an extreme point. Setting the derivative to 0 shows that all extreme points z^* satisfy

$$z^* = \log(z^*)\alpha^{\frac{1}{\alpha-1}} .$$

The function values at the extreme points are therefore lower bounded by

$$f(z^*) \geq \min_{z\geq 1}\left(\log(z)^\alpha - \frac{1-(\log(z)\alpha^{\frac{1}{\alpha-1}})^{-1+\alpha}}{1-\alpha}\right) = \min_{\tilde{z}\geq 0}\left(\tilde{z}^\alpha - \frac{1-\tilde{z}^{-1+\alpha}\alpha}{1-\alpha}\right),$$

where we apply the substitution $\tilde{z} = \log(z)$. The RHS goes to infinity for $\tilde{z} \to 0$ and $\tilde{z} \to \infty$, which means the only extreme point (can be verified by taking the derivative) at $\tilde{z} = 1$ is the minimum. Since $1^\alpha - \frac{1-1^{-1+\alpha}\alpha}{1-\alpha} = 0$, the function f is always positive, which concludes the proof. □

Support Lemmas for Section 3.7

We use $v = (v_i)_{i=1,\dots,K}$ to denote a column vector $v \in \mathbb{R}^K$ with elements $v_1, \dots, v_K$ and $\operatorname{diag}(v)$ to denote a $K \times K$ matrix with $v_1, \dots, v_K$ on the diagonal and 0 elsewhere. For a positive semidefinite matrix M we use $||\cdot||_M := \sqrt{\langle \cdot, M\cdot\rangle}$ to denote the canonical norm with respect to M. We also use the following properties of the potential function.

$$\begin{aligned}
&\Psi_t(w) = -\sum_i \frac{{w_i}^\alpha - \alpha w_i}{\alpha(1-\alpha)\eta_t\xi_i},\\
&\nabla\Psi_t(w) = -\left(\frac{{w_i}^{\alpha-1}-1}{(1-\alpha)\eta_t\xi_i}\right)_{i=1,\dots,K},\\
&\nabla^2\Psi_t(w) = \operatorname{diag}\left(\left(\frac{{w_i}^{\alpha-2}}{\eta_t\xi_i}\right)_{i=1,\dots,K}\right),\\
&\text{For } Y \leq 0:\\
&\Psi_t^*(Y) = \max_w \langle w, Y\rangle + \frac{1}{\eta_t}\sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)\xi_i},\\
&\nabla\Psi_t^*(Y) = \arg\max_w \langle w, Y\rangle + \frac{1}{\eta_t}\sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)\xi_i} = \left((-\eta_t(1-\alpha)\xi_i Y_i + 1)^{\frac{1}{\alpha-1}}\right)_{i=1,\dots,K}. \qquad (3.8)
\end{aligned}$$

As we have shown in Section 3.4.3, there exists a Lagrange multiplier ν, such that the algorithm picks the probabilities

$$w_t = \nabla\Phi_t(-\hat{L}_{t-1}) = \nabla\Psi^*(-\hat{L}_{t-1} + \nu\mathbf{1}_K). \qquad (3.9)$$

Ψ_t is a Legendre function, which implies that its gradient is invertible and $\nabla\Psi_t^{-1} = \nabla\Psi_t^*$ [84]. Furthermore, by the Inverse Function theorem,

$$\nabla^2\Psi_t^*(\nabla\Psi_t(w)) = \nabla^2\Psi_t(w)^{-1} = \operatorname{diag}\left(\eta_t\xi_i{w_i}^{2-\alpha}\right)_{i=1,\dots,K}. \qquad (3.10)$$

The Bregman divergence associated with a Legendre function f is defined by

$$D_f(x,y) = f(x) - f(y) - \langle\nabla f(y), x-y\rangle .$$

Due to Taylor's theorem, it satisfies for some $z \in \operatorname{conv}(x,y)$

$$D_f(x,y) \leq \frac{1}{2}||x-y||^2_{\nabla^2 f(z)}. \qquad (3.11)$$

Controlling the *stability* Term

Equation (3.11) gives us a way of bounding the *stability* term. The following lemma allows us to control the eigenvalues of the Hessian $\nabla^2\Psi_t^*$.

Lemma 3.6. *Let $w \in \Delta^{K-1}$ and $\tilde{w} = \nabla\Psi_t^*(\nabla\Psi_t(w) - \ell)$. If $\eta_t\xi_i \leq \frac{1}{4}$ for all i, then for all $\ell \in \mathbb{R}^K$ with $\ell_i \geq -1$ for all i, it holds that $\tilde{w}_i^{2-\alpha} \leq 2w_i^{2-\alpha}$ for all i.*

Proof. Since $\nabla\Psi_t$ is the inverse of $\nabla\Psi_t^*$, we have

$$\begin{aligned}
&\nabla\Psi_t(w)_i - \nabla\Psi_t(\tilde{w})_i = \ell_i \geq -1,\\
&\frac{w_i^{\alpha-1} - 1}{(1-\alpha)\eta_t\xi_i} - \frac{\tilde{w}_i^{\alpha-1} - 1}{(1-\alpha)\eta_t\xi_i} \leq 1,\\
&\tilde{w}_i^{1-\alpha} \leq \frac{w_i^{1-\alpha}}{1 - \eta_t\xi_i(1-\alpha)w_i^{1-\alpha}} \leq \frac{w_i^{1-\alpha}}{1 - \eta_t\xi_i(1-\alpha)},\\
&\tilde{w}_i^{2-\alpha} \leq \frac{w_i^{2-\alpha}}{(1 - \eta_t\xi_i(1-\alpha))^{\frac{2-\alpha}{1-\alpha}}}.
\end{aligned}$$

It remains to bound $(1 - \eta_t\xi_i(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}}$. Note that this function is monotonically decreasing in α, which can be verified by confirming that the derivative is negative in $[0,1]$. Using the fact that $\eta_t\xi_i \leq \frac{1}{4}$, we have

$$(1 - \eta_t\xi_i(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}} \leq (1 - \eta_t\xi_i)^{-2} \leq \frac{4^2}{3^2} \leq 2.$$

□

For the reduced-variance estimators and $\alpha = 1/2$ we provide a tighter bound for the stability term by using the following two lemmas. For $\alpha = 1/2$ and $\xi_i = 1$ we have

$$\begin{aligned}
\Psi_t(w) &= -4\eta_t^{-1}\sum_{i=1}^K (w_i^{\frac{1}{2}} - \frac{1}{2}w_i)\\
\nabla\Psi_t(w) &= \left(-2\eta_t^{-1}(w_i^{-\frac{1}{2}} - 1)\right)_{i=1,\ldots,K}.
\end{aligned}$$

Lemma 3.7. *The convex conjugate of $\Psi_t(w) = -4\eta_t^{-1}\sum_{i=1}^K(w_i^{\frac{1}{2}} - \frac{1}{2}w_i)$ is*

$$\Psi_t^*(Y) = \begin{cases} \sum_{i=1}^K \frac{2\eta_t^{-1}}{1-\eta_t Y_i/2}, & \textit{if } Y_i < 2\eta_t^{-1} \textit{ for all } i\\ \infty, & \textit{otherwise.}\end{cases}$$

Proof.

$$\begin{aligned}
\Psi_t^*(Y) &= \sup_{w\in\mathbb{R}^K} \langle w, Y\rangle - \Psi_t(w)\\
&= \sum_{i=1}^K \sup_{w\in\mathbb{R}} w(Y_i - 2\eta_t^{-1}) + 4\eta_t^{-1}w^{\frac{1}{2}}.
\end{aligned}$$

For $Y_i \geq 2\eta^{-1}$, the term goes to infinity as $w \to \infty$. Otherwise, the maximum is obtained by $w = \frac{1}{(1-\frac{1}{2}\eta_t Y_i)^2}$, which concludes the proof. □

Using the explicit form of the convex conjugate, we can show a general bound on the stability.

Lemma 3.8. *Let $\alpha = 1/2$ and $\xi_i = 1$. Then for any x, such that $\min_i \eta_t(\hat{\ell}_{t,i} - x)w_{ti}^{\frac{1}{2}} \geq -1$, the instantaneous stability satisfies*

$$\langle w_t, \hat{\ell}_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}) \leq \sum_{i=1}^{K} \frac{\eta_t}{2} w_{t,i}{}^{\frac{3}{2}} (\hat{\ell}_{t,i} - x)^2 + \frac{\eta_t^2}{2} w_{t,i}{}^2 \left| x - \hat{\ell}_{t,i} \right|_+^3 ,$$

where $|z|_+ = \max\{z, 0\}$.

Proof. By equation (3.9) and since $\nabla \Psi_t^{-1} = \nabla \Psi_t^*$, there exists a Lagrange multiplier ν such that

$$-\hat{L}_{t-1} = \nabla \Psi_t(w_t) - \nu \mathbf{1}_K .$$

Furthermore, $\Phi_t(-L - \nu \mathbf{1}_K) = \Phi_t(-L) - \nu$, since the maximisation over w is restricted to the probability simplex. Using these two properties, we have for any $x \in \mathbb{R}$

$$\begin{aligned}
&\langle w_t, \hat{\ell}_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}) \\
&\qquad = \langle w_t, \hat{\ell}_t \rangle + \Phi_t(-\hat{\ell}_t + \nabla \Psi_t(w_t) - \nu \mathbf{1}_K) - \Phi_t(\nabla \Psi_t(w_t) - \nu \mathbf{1}_K) \\
&\qquad = \langle w_t, \hat{\ell}_t \rangle + \Phi_t(-\hat{\ell}_t + \nabla \Psi_t(w_t)) - \Phi_t(\nabla \Psi_t(w_t)) \\
&\qquad = \langle w_t, \hat{\ell}_t - x \mathbf{1}_K \rangle + \Phi_t(x \mathbf{1}_K - \hat{\ell}_t + \nabla \Psi_t(w_t)) - \Phi_t(\nabla \Psi_t(w_t)) \\
&\qquad \leq \langle w_t, \hat{\ell}_t - x \mathbf{1}_K \rangle + \Psi_t^*(x \mathbf{1}_K - \hat{\ell}_t + \nabla \Psi_t(w_t)) - \Psi_t^*(\nabla \Psi_t(w_t)) ,
\end{aligned}$$

where the last line uses $\Phi_t(\nabla \Psi_t(w)) = \Psi_t^*(\nabla \Psi_t(w))$ for any $w \in \Delta^{K-1}$, which holds because the argmax in both terms is w, and the inequality $\Phi_t(L) \leq \Psi_t^*(L)$, which holds because Φ_t is a constrained version of Ψ_t^*.

Using the explicit expression for the convex conjugate in Lemma 3.7 and assuming that x is in the range defined in the statement of Lemma 3.8, which ensures that the convex conjugate is bounded, we have

$$\begin{aligned}
&\langle w_t, \hat{\ell}_t - x \mathbf{1}_K \rangle + \Psi_t^*(x \mathbf{1}_K - \hat{\ell}_t + \nabla \Psi_t(w_t)) - \Psi_t^*(\nabla \Psi_t(w_t)) \\
&\qquad = \sum_{i=1}^{K} w_{t,i}(\hat{\ell}_{t,i} - x) + \frac{2}{\eta_t} \left(w_{t,i}{}^{-\frac{1}{2}} + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) \right)^{-1} - \frac{2}{\eta_t} w_{t,i}{}^{\frac{1}{2}} \\
&\qquad = \sum_{i=1}^{K} \frac{2}{\eta_t} w_{t,i}{}^{\frac{1}{2}} \left(\frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} + \left(1 + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} \right)^{-1} - 1 \right) \\
&\qquad = \sum_{i=1}^{K} \frac{\eta_t}{2} w_{t,i}{}^{\frac{3}{2}} (\hat{\ell}_{t,i} - x)^2 \left(1 + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} \right)^{-1} . \qquad (3.12)
\end{aligned}$$

From $\min_i \eta_t(\hat{\ell}_{t,i} - x)w_{ti}^{\frac{1}{2}} \geq -1$ it follows that $\forall i : \, (1 + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}})^{-1} \leq 2$, so

$$\begin{aligned}
\left(1 + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} \right)^{-1} &= 1 - \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} \left(1 + \frac{\eta_t}{2}(\hat{\ell}_{t,i} - x) w_{t,i}{}^{\frac{1}{2}} \right)^{-1} \\
&\leq 1 + \eta_t |x - \hat{\ell}_{t,i}|_+ w_{t,i}{}^{\frac{1}{2}} . \qquad (3.13)
\end{aligned}$$

Combining equations (3.12) and (3.13) completes the proof. □

Now we have all the tools to prove the main *stability* lemma.

Proof of Lemma 3.1. We begin by proving the first and the last part of the lemma followed by the second and third.

First part of the lemma. First we bound the *stability* by 1. By convexity of Φ_t, we have

$$\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}) \leq \ell_{t,I_t} - \langle \nabla\Phi_t(-\hat{L}_t), \hat{\ell}_t \rangle \leq 1\,,$$

where the second inequality uses the non-negativity of the loss estimator under importance sampling.

Recall that $w_t = \nabla\Phi_t(-\hat{L}_{t-1})$ and $\ell_{t,I_t} = \left\langle w_t, \hat{\ell}_t \right\rangle$. Furthermore, $\Phi_t(L + x\mathbf{1}_K) = \Phi_t(L) + x$, where $\mathbf{1}_K$ is a vector of K ones, since we take the argmax over probability distributions. Finally, from equation (3.9) follows the existence of a constant c_t, such that $\nabla\Psi_t(w_t) = -\hat{L}_{t-1} + c_t\mathbf{1}_K$. Hence, for any $x \in \mathbb{R}$

$$\begin{aligned}
&\mathbb{E}\left[\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \\
&\quad = \mathbb{E}\left[\left\langle w_t, \hat{\ell}_t \right\rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \\
&\quad = \mathbb{E}\left[\left\langle w_t, \hat{\ell}_t \right\rangle + \Phi_t(\nabla\Psi_t(w_t) - \hat{\ell}_t) - \Phi_t(\nabla\Psi_t(w_t))\right] \\
&\quad = \mathbb{E}\left[\left\langle w_t, \hat{\ell}_t - x\mathbf{1}_K \right\rangle + \Phi_t(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K) - \Phi_t(\nabla\Psi_t(w_t))\right] \\
&\quad \leq \mathbb{E}\left[\left\langle w_t, \hat{\ell}_t - x\mathbf{1}_K \right\rangle + \Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K) - \Psi_t^*(\nabla\Psi_t(w_t))\right] && (3.14)\\
&\quad = \mathbb{E}\left[D_{\Psi_t^*}(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(w_t))\right] \\
&\quad \leq \mathbb{E}\left[\max_{z\in\mathrm{conv}(\nabla\Psi_t(w_t), \nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K)} \frac{1}{2}||\hat{\ell}_t - x\mathbf{1}_K||^2_{\nabla^2\Psi_t^*(z)}\right] && (3.15)\\
&\quad = \mathbb{E}\left[\max_{w\in\mathrm{conv}(w_t, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K))} \frac{1}{2}||\hat{\ell}_t - x\mathbf{1}_K||^2_{\nabla^2\Psi_t(w)^{-1}}\right] && (3.16)\\
&\quad \leq \mathbb{E}\left[\sum_{i=1}^K \max_{w_i\in[w_{t,i}, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K)_i]} \frac{\eta_t\xi_i}{2}(\hat{\ell}_{t,i} - x)^2 w_i^{2-\alpha}\right],
\end{aligned}$$

where in equation (3.14) we have $\Phi_t(x) \leq \Psi_t^*(x)$, because Φ_t is a constrained version of Ψ_t^*, and $\Phi_t(\nabla\Psi_t(w_t)) = \Psi_t^*(\nabla\Psi_t(w_t))$, because $\arg\max_{w\in\mathbb{R}^K} \langle w, \nabla\Psi_t(w_t)\rangle - \Psi(w) = w_t$ is in the probability simplex and the constraint is inactive. Inequality (3.15) follows by equation (3.11), and (3.16) by equation (3.10).

In order to prove the first part of the Lemma, we set $x = 0$ and observe that $\nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t)_i \leq \nabla\Psi_t^*(\nabla\Psi_t(w_t))_i = w_{t,i}$ because of non-negativity of the losses and the fact that $\nabla\Psi_t^*$ is monotonically increasing, see (3.8). (The observation implies that the highest value of $w_i \in [w_{t,i}, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t)_i]$ is $w_{t,i}$.) Since the importance weighted losses are 0 for the arms that were not played, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^K \max_{w_i\in[w_{t,i}, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t)_i]} \frac{\eta_t\xi_i}{2}\hat{\ell}_{t,i}^2 w_i^{2-\alpha}\right] = \mathbb{E}\left[\sum_{i=1}^K \frac{\eta_t\xi_i}{2}\hat{\ell}_{t,i}^2 {w_{t,i}}^{2-\alpha}\right] \\
&\quad = \frac{\eta_t\xi_i}{2}\mathbb{E}\left[\sum_{i=1}^K \frac{\ell_{t,i}^2}{{w_{t,i}}^2}{w_{t,i}}^{2-\alpha}\mathbb{1}_t(i)\right] = \frac{\eta_t\xi_i}{2}\mathbb{E}\left[\sum_{i=1}^K \frac{\ell_{t,i}^2}{{w_{t,i}}^2}{w_{t,i}}^{3-\alpha}\right] \leq \sum_{i=1}^K \frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha},
\end{aligned}$$

where we use that $\mathbb{E}[\mathbb{1}_t(i)|\ell_1, \ldots, \ell_{t-1}, I_1, \ldots, I_{t-1}] = w_{t,i}$. The last inequality follows by Jensen's inequality.

Fourth part of the lemma. We set $x = \mathbb{1}_t(j)\ell_{t,j}$. In the calculation below, for the events $I_t \in \{1, \ldots, K\} \setminus j$, we have $x = 0$ and use the same derivation as in the previous case. When $I_t = j$, for $i \neq j$ we have $\hat{\ell}_{t,i} - x = -x \geq -1$ and for j we have $\hat{\ell}_{t,j} - x \geq 0$. For $i \neq j$

we use Lemma 3.6 to bound $\left(\nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K)_i\right)^{2-\alpha} \leq 2w_{t,i}^{2-\alpha}$ and for j we use $\nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t)_j \leq \nabla\Psi_t^*(\nabla\Psi_t(w_t))_j = w_{t,j}$.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{w}_i \in [w_{t,i}, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \hat{\ell}_t + x\mathbf{1}_K)_i]} \frac{\eta_t\xi_i}{2}(\hat{\ell}_{t,i} - x)^2 \tilde{w}_i^{2-\alpha}\right] \\
&\quad \leq \sum_{i \neq j} \frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha} + \mathbb{E}\left[\mathbb{1}_t(j)\left(\frac{\eta_t\xi_j}{2}\left(\frac{\ell_{t,j}}{w_{t,j}} - \ell_{t,j}\right)^2 w_{t,j}^{2-\alpha} + \sum_{i \neq j}\frac{\eta_t\xi_i}{2}\ell_{t,j}^2 2{w_{t,i}}^{2-\alpha}\right)\right] \\
&\quad \leq \sum_{i \neq j} \frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha} + \mathbb{E}\left[\frac{\eta_t\xi_j}{2}(1 - w_{t,j})^2 w_{t,j}^{1-\alpha} + \sum_{i \neq j}\eta_t\xi_i {w_{t,i}}^{2-\alpha} w_{t,j}\right] \\
&\quad \leq \sum_{i \neq j}\left(\frac{\eta_t\xi_i}{2}\mathbb{E}\left[w_{t,i}\right]^{1-\alpha} + \frac{\eta_t(\xi_j + 2\xi_i)}{2}\mathbb{E}\left[w_{t,i}\right]\right),
\end{aligned}
$$

where in the last step for the middle term we use $(1 - w_{t,j})^2 w_{t,j}^{1-\alpha} \leq 1 - w_{t,j} = \sum_{i \neq j} w_{t,i}$ and for the last term $w_{t,i}^{2-\alpha} \leq 1$.

Second part of the lemma. We set $x = \ell_{t,I_t}$ and first verify that Lemma 3.8 can be applied. We have for any i:

$$
\eta_t(\hat{\ell}_{t,i} - x)w_{t,i}^{\frac{1}{2}} \geq -\eta_t w_{t,i}^{\frac{1}{2}} \geq -\eta_t \geq -1\,,
$$

where the last inequality is by the assumption of the lemma. Since $\mathbb{E}[\ell_{t,I_t}] = \mathbb{E}[\langle w_t, \ell_t\rangle\rangle] = \mathbb{E}[\langle w_t, \hat{\ell}_t\rangle\rangle]$, applying Lemma 3.8 we have

$$
\begin{align}
\mathbb{E}\left[\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}}(\hat{\ell}_{t,i} - \ell_{t,I_t})^2 + \frac{\eta_t^2}{2}{w_{t,i}}^2\left|\ell_{t,I_t} - \hat{\ell}_{t,i}\right|_+^3\right] \notag \\
&\leq \frac{\eta_t}{2}\mathbb{E}\left[{w_{t,I_t}}^{-\frac{1}{2}}(1 - w_{t,I_t})^2 + \sum_{i \neq I_t}{w_{t,i}}^{\frac{3}{2}}\right] + \frac{\eta_t^2}{2} \tag{3.17} \\
&= \frac{\eta_t}{2}\mathbb{E}\left[\sum_{i=1}^{K}{w_{t,i}}^{\frac{1}{2}}(1 - w_{t,i})^2 + (1 - w_{t,i}){w_{t,i}}^{\frac{3}{2}}\right] + \frac{\eta_t^2}{2} \tag{3.18} \\
&= \frac{\eta_t}{2}\mathbb{E}\left[\sum_{i=1}^{K}{w_{t,i}}^{\frac{1}{2}}(1 - w_{t,i})\right] + \frac{\eta_t^2}{2} \notag \\
&\leq \frac{\eta_t}{2}\sum_{i=1}^{K}\mathbb{E}[w_{t,i}]^{\frac{1}{2}}(1 - \mathbb{E}[w_{t,i}]) + \frac{\eta_t^2}{2}\,, \tag{3.19}
\end{align}
$$

where equation (3.17) uses that the loss estimators are positive and hence $\left|\ell_{t,I_t} - \hat{\ell}_{t,i}\right|_+ \leq 1$ and that the losses are bounded in $[0,1]$; equation (3.18) uses that the conditional probability of $I_t = i$ is $w_{t,i}$; and equation (3.19) follows by concavity of the function $f(z) = z^{\frac{1}{2}}(1 - z)$ and Jensen's inequality.

Third part of the lemma. We set $x = \ell_{t,I_t}$ and first verify that Lemma 3.8 can be applied. Recall that $\mathbb{B}_t(i) := \frac{1}{2}\mathbb{1}(w_{t,i} \geq \eta_t^2)$. For $i \neq I_t$ we have $\hat{\ell}_{t,i} = \mathbb{B}_t(i)$ and

$$
\eta_t(\hat{\ell}_{t,i} - x)w_{ti}^{\frac{1}{2}} = \eta_t(\mathbb{B}_t(i) - \ell_{t,I_t})w_{ti}^{\frac{1}{2}} \geq -\eta_t \geq -1\,,
$$

while for I_t we have $\hat{\ell}_{t,I_t} = \frac{\ell_{t,I_t} - \mathbb{B}_t(I_t)}{w_{t,I_t}} + \mathbb{B}_t(I_t)$ and

$$\eta_t(\hat{\ell}_{t,I_t} - \ell_{t,I_t})w_{t,i}^{\frac{1}{2}} = \eta_t(\ell_{t,I_t} - \mathbb{B}_t(I_t))(\frac{1}{w_{t,I_t}} - 1){w_{t,I_t}}^{\frac{1}{2}} \geq -\eta_t \mathbb{B}_t(I_t){w_{t,I_t}}^{-\frac{1}{2}} \geq -1\,.$$

Since $\mathbb{E}[\ell_{t,I_t}] = \mathbb{E}[\langle w_t, \ell_t \rangle\rangle] = \mathbb{E}[\langle w_t, \hat{\ell}_t \rangle\rangle]$, applying Lemma 3.8 we have

$$\begin{aligned}
\mathbb{E}\left[\ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}}(\hat{\ell}_{t,i} - \ell_{t,I_t})^2 + \frac{\eta_t^2}{2}{w_{t,i}}^2 \left|\ell_{t,I_t} - \hat{\ell}_{t,i}\right|_+^3\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}}(\hat{\ell}_{t,i} - \ell_{t,I_t})^2 + \frac{\eta_t^2}{2}{w_{t,i}}^2 \left|\hat{\ell}_{t,i} - \ell_{t,I_t}\right|^3\right]. \quad (3.20)
\end{aligned}$$

For any $\ell \in [0,1]$ and any i, we have $|\ell - \mathbb{B}_t(i)| \leq |1 - \mathbb{B}_t(i)|$. For I_t, we have $|\hat{\ell}_{t,I_t} - \ell_{t,I_t}| = |\ell_{t,I_t} - \mathbb{B}_t(I_t)|\frac{1-w_{t,I_t}}{w_{t,I_t}} \leq |1 - \mathbb{B}_t(I_t)|\frac{1-w_{t,I_t}}{w_{t,I_t}}$, while for $i \neq I_t$ we have $|\hat{\ell}_{t,i} - \ell_{t,I_t}| \leq |1 - \mathbb{B}_t(i)|$. Let $\bar{\mathbb{B}}_t(i) = \frac{1}{2}\mathbb{1}(w_{t,i} < \eta_t^2) = \frac{1}{2} - \mathbb{B}_t(i)$, then $|1 - \mathbb{B}_t(i)| = |\frac{1}{2} + \bar{\mathbb{B}}_t(i)|$. We have

$$\begin{aligned}
|\frac{1}{2} + \bar{\mathbb{B}}_t(i)|^2 &= \frac{1}{4} + \frac{3}{2}\bar{\mathbb{B}}_t(i)\,, \\
|\frac{1}{2} + \bar{\mathbb{B}}_t(i)|^3 &\leq 1\,.
\end{aligned}$$

Plugging these into equation (3.20) leads to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}}(\hat{\ell}_{t,i} - \ell_{t,I_t})^2 + \frac{\eta_t^2}{2}{w_{t,i}}^2 \left|\hat{\ell}_{t,i} - \ell_{t,I_t}\right|^3\right] \\
&\qquad \leq \mathbb{E}\Bigg[\left(\frac{1}{4} + \frac{3}{2}\bar{\mathbb{B}}_t(I_t)\right)\frac{\eta_t}{2}{w_{t,I_t}}^{-\frac{1}{2}}(1 - w_{t,I_t})^2 + \frac{\eta_t^2}{4}{w_{t,I_t}}^{-1}(1 - w_{t,I_t})^3 \\
&\qquad\qquad + \sum_{i \neq I_t}\left(\frac{1}{4} + \frac{3}{2}\bar{\mathbb{B}}_t(i)\right)\frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}} + \frac{\eta_t^2}{4}{w_{t,i}}^2\Bigg] \\
&\qquad = \mathbb{E}\Bigg[\sum_{i=1}^{K}\Bigg(\left(\frac{1}{4} + \frac{3}{2}\bar{\mathbb{B}}_t(i)\right)\frac{\eta_t}{2}{w_{t,i}}^{\frac{1}{2}}(1 - w_{t,i})^2 + \frac{\eta_t^2}{4}(1 - w_{t,i})^3 \\
&\qquad\qquad + (1 - w_{t,i})\left(\frac{1}{4} + \frac{3}{2}\bar{\mathbb{B}}_t(i)\right)\frac{\eta_t}{2}{w_{t,i}}^{\frac{3}{2}} + (1 - w_{t,i})\frac{\eta_t^2}{4}{w_{t,i}}^2\Bigg)\Bigg] \\
&\qquad \leq \mathbb{E}\Bigg[\sum_{i=1}^{K}\left(\frac{\eta_t}{8}{w_{t,i}}^{\frac{1}{2}}(1 - w_{t,i}) + \frac{3}{2}\bar{\mathbb{B}}_t(i)\frac{\eta_t}{2}{w_{t,i}}^{\frac{1}{2}} + \frac{\eta_t^2}{4}\right)\Bigg] \\
&\qquad \leq \frac{5\eta_t^2}{8}K + \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_t}{8}{w_{t,i}}^{\frac{1}{2}}(1 - w_{t,i})\right] \\
&\qquad \leq \frac{5\eta_t^2}{8}K + \sum_{i=1}^{K}\frac{\eta_t}{8}\mathbb{E}[w_{t,i}]^{\frac{1}{2}}(1 - \mathbb{E}[w_{t,i}])\,,
\end{aligned}$$

where the last step is by concavity of $f(z) = z^{\frac{1}{2}}(1 - z)$ and Jensen's inequality. □

Controlling the *penalty* Term

We begin with a standard lemma to simplify the *penalty* term.

Lemma 3.9. *For any $\alpha \in [0,1]$, any positive learning rate, and any fixed $v, u \in \Delta^{K-1}$, the* penalty *term satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \le \mathbb{E}\Bigg[\frac{\Psi(v)-\Psi(w_1)}{\eta_1} + \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)(\Psi(v)-\Psi(w_t)) + \frac{\Psi(u)-\Psi(v)}{\eta_T}\Bigg] + \left\langle u - \mathbf{e}_{i_T^*}, L_T\right\rangle.$$

Proof. First, note that all the terms involving $\Psi(v)$ in the lemma sum up to 0. Then, recall that w_t is defined as $\arg\max_{w\in\Delta^{K-1}}\left\{\left\langle w, -\hat{L}_{t-1}\right\rangle - \frac{\Psi(w)}{\eta_t}\right\}$. Therefore,

$$\Phi_t(-\hat{L}_{t-1}) = -\left\langle w_t, \hat{L}_{t-1}\right\rangle - \frac{\Psi(w_t)}{\eta_t}.$$

Furthermore, by definition of the potential function, for any $\tilde{w} \in \Delta^{K-1}$ it holds that:

$$-\Phi_t(-\hat{L}_t) = -\max_{w\in\Delta^{K-1}}\left\{\left\langle w, -\hat{L}_t\right\rangle - \frac{\Psi(w)}{\eta_t}\right\} \le \left\langle \tilde{w}, \hat{L}_t\right\rangle + \frac{\Psi(\tilde{w})}{\eta_t}.$$

Setting $\tilde{w}$ to w_{t+1} for $t < T$ and to u for $t = T$, and using $\hat{L}_0 = \mathbf{0}_K$, where $\mathbf{0}_K$ is a vector of K zeros, the sum of potential differences can be bounded as follows:

$$\begin{aligned}\sum_{t=1}^{T}&\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right)\\ &\le \sum_{t=1}^{T}\left(-\left\langle w_t, \hat{L}_{t-1}\right\rangle - \frac{\Psi(w_t)}{\eta_t}\right) + \sum_{t=1}^{T-1}\left(\left\langle w_{t+1}, \hat{L}_t\right\rangle + \frac{\Psi(w_{t+1})}{\eta_t}\right) + \left\langle u, \hat{L}_T\right\rangle + \frac{\Psi(u)}{\eta_T}\\ &= -\frac{\Psi(w_1)}{\eta_1} - \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)\Psi(w_t) + \frac{\Psi(u)}{\eta_T} + \left\langle u, \hat{L}_T\right\rangle.\end{aligned}$$

The proof is finalised by taking the expectation and subtracting the optimal loss. Due to unbiasedness of the loss estimators, for a fixed u we have $\mathbb{E}[\langle u, \hat{L}_T\rangle] = \langle u, L_T\rangle$. □

Proof of Lemma 3.2. The proof of both parts of the lemma is based on Lemma 3.9.

Part 1: We set $v = w_1$. Since $w_1 = \arg\max_{w\in\Delta^{K-1}} -\Psi_1(w) = \arg\max_{w\in\Delta^{K-1}} -\Psi(w)$, we have $\Psi(w_1) - \Psi(w_t) \le 0$ for any t. Since the learning rate is non-increasing, the terms $({\eta_t}^{-1} - {\eta_{t-1}}^{-1})$ are all positive, so

$$\begin{aligned}\mathbb{E}&\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right]\\ &\le \mathbb{E}\Bigg[\frac{\Psi(w_1)-\Psi(w_1)}{\eta_1} + \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)(\Psi(w_1)-\Psi(w_t)) + \frac{\Psi(u)-\Psi(w_1)}{\eta_T}\Bigg]\\ &\qquad + \left\langle u - \mathbf{e}_{i_T^*}, L_T\right\rangle\\ &\le \mathbb{E}\left[\frac{\Psi(u)-\Psi(w_1)}{\eta_T}\right] + \left\langle u - \mathbf{e}_{i_T^*}, L_T\right\rangle.\end{aligned}$$

Following the trick of Agarwal et al. [7], we set $u_{i_T^*} = 1 - T^{-1}$ and $u_i = \frac{T^{-1}}{K-1}$ for $i \neq i_T^*$. The losses are bounded in $[0, T]$, so this choice of u implies $\langle u - \mathbf{e}_{i_T^*}, L_T\rangle \le 1$. Since we assume that the regulariser is symmetric, the explicit form of w_1 is $w_{1,i} = K^{-1}$ and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \\
&\qquad \le \frac{K^{1-\alpha}}{\alpha(1-\alpha)\eta_T} - \frac{(K-1)^{1-\alpha}T^{-\alpha} + (1-T^{-1})^{\alpha}}{\alpha(1-\alpha)\eta_T} + 1.
\end{aligned}$$

It remains to bound $K^{1-\alpha} - (K-1)^{1-\alpha}T^{-\alpha} - (1-T^{-1})^{\alpha}$. Since x^{α} and $x^{1-\alpha}$ are concave functions, by Taylor's expansion around $X-1$ we have $X^{1-\alpha} \le (X-1)^{1-\alpha} + (1-\alpha)(X-1)^{-\alpha}$ and $X^{\alpha} \le (X-1)^{\alpha} + \alpha(X-1)^{\alpha-1}$ for any $X > 1$, thus

$$\begin{aligned}
K^{1-\alpha} + T^{\alpha} &\le (K-1)^{1-\alpha} + (1-\alpha)(K-1)^{-\alpha} + (T-1)^{\alpha} + \alpha(T-1)^{\alpha-1} \\
&\le (K-1)^{1-\alpha} + (T-1)^{\alpha} + 1,
\end{aligned}$$

where the last line uses $(T-1)^{\alpha-1}, (K-1)^{-\alpha} \le 1$. Therefore,

$$\begin{aligned}
K^{1-\alpha} - (K-1)^{1-\alpha}T^{-\alpha} - (1-T^{-1})^{\alpha} &= K^{1-\alpha} + T^{-\alpha}(-(K-1)^{1-\alpha} - (T-1)^{\alpha}) \\
&\le K^{1-\alpha} + T^{-\alpha}(-K^{1-\alpha} - T^{\alpha} + 1) \\
&= (K^{1-\alpha} - 1)(1 - T^{-\alpha}).
\end{aligned}$$

Part 2: Set

$$\begin{aligned}
\tilde{w} &= \operatorname*{arg\,max}_{w \in \Delta^{K-1}} -\Psi\left((1-T^{-x})\mathbf{e}_{i_T^*} + T^{-x}w\right), \\
v = u &= (1-T^{-x})\mathbf{e}_{i_T^*} + T^{-x}\tilde{w}, \\
v_t &= (1-T^{-x})\mathbf{e}_{i_T^*} + T^{-x}w_t.
\end{aligned}$$

By definition, $\Psi(v) \le \Psi(v_t)$ for all t. So

$$\Psi(v) - \Psi(w_t) \le \Psi(v_t) - \Psi(w_t) \le \sum_{i \ne i_T^*} \frac{({w_{t,i}}^{\alpha} - \alpha w_{t,i})(1-T^{-\alpha x})}{\alpha(1-\alpha)\xi_i}.$$

In the last inequality we have used the fact that the contribution of the optimal arm i_T^* is non-positive, since $v_{t,i_T^*} \ge w_{t,i_T^*}$ and $w^{\alpha} - \alpha w$ is monotonically increasing in w over $[0,1]$. The choice of u ensures that $\langle u - \mathbf{e}_{i_T^*}, L_T\rangle \le T^{1-x}$. Starting again with Lemma 3.9, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \\
&\le \mathbb{E}\left[\frac{\Psi(v) - \Psi(w_1)}{\eta_1} + \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)(\Psi(v) - \Psi(w_t)) + \frac{\Psi(u) - \Psi(v)}{\eta_T}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \left\langle u - \mathbf{e}_{i_T^*}, L_T\right\rangle \\
&\le \frac{1 - T^{-\alpha x}}{\alpha} \sum_{i \ne i_T^*} \mathbb{E}\left[\frac{{w_{1,i}}^{\alpha} - \alpha w_{1,i}}{(1-\alpha)\eta_1\xi_i} + \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)\frac{{w_{t,i}}^{\alpha} - \alpha w_{t,i}}{(1-\alpha)\xi_i}\right] + T^{1-x} \\
&\le \frac{1 - T^{-\alpha x}}{\alpha} \sum_{i \ne i_T^*} \frac{\mathbb{E}[w_{1,i}]^{\alpha} - \alpha\mathbb{E}[w_{1,i}]}{(1-\alpha)\eta_1\xi_i} + \sum_{t=2}^{T}\left({\eta_t}^{-1} - {\eta_{t-1}}^{-1}\right)\frac{\mathbb{E}[w_{t,i}]^{\alpha} - \alpha\mathbb{E}[w_{t,i}]}{(1-\alpha)\xi_i} + T^{1-x}.
\end{aligned}$$

□

Proof of Theorem 3.3

We follow the same strategy as outlined in Section 3.7. In order to cover the limit cases $\alpha \in \{0, 1\}$, the proof is significantly more technical than the proof of Theorem 3.1.

Proof of Theorem 3.3. Recall that the learning rate is $\eta_t = \frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}$, where $\bar{t} = \max\{e, t\}$, and regularisation parameters are $\xi_i = \Delta_i^{1-2\alpha}$, where $\Delta_{\min} = \min_{i \neq i^*} \Delta_i$.

Bounding the *stability* term We start by bounding the *stability* term. When $t \leq T_0$ we use the first part of Lemma 3.1 and otherwise the second with $j = i^*$. T_0 will be later chosen, so that $\eta_{T_0}\xi_i \leq \frac{1}{4}$.

$$\begin{aligned}
\textit{stability} &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] \\
&\leq \underbrace{\sum_{t=1}^{T}\sum_{i \neq i^*} \frac{\eta_t \xi_i \mathbb{E}[w_{t,i}]^{1-\alpha}}{2}}_{\textit{concave}} + \underbrace{\sum_{t=1}^{T_0} \frac{\eta_t \xi_{i^*} \mathbb{E}[w_{t,i^*}]^{1-\alpha}}{2}}_{\textit{constant}} + \underbrace{\sum_{t=T_0+1}^{T}\sum_{i \neq i^*} \frac{\eta_t(\xi_i + 2\xi_{i^*})}{2}\mathbb{E}[w_{t,i}]}_{\textit{linear}}.
\end{aligned}$$

Bounding the concave part Since $w^{1-\alpha}$ is a concave function of w, it can be upper bounded by the first order Taylor's approximation for all w^*:

$$\begin{aligned}
{w_{t,i}}^{1-\alpha} &\leq {w^*}^{1-\alpha} + (1-\alpha){w^*}^{-\alpha}(w_{t,i} - w^*) \\
&= \alpha {w^*}^{1-\alpha} + (1-\alpha){w^*}^{-\alpha} w_{t,i}.
\end{aligned}$$

Taking $w^* = \frac{16}{\Delta_i^2 t}$ (with $\eta_t = \frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}$, $\xi_i = \Delta_i^{1-2\alpha}$):

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\eta_t \xi_i}{2}\mathbb{E}[w_{t,i}]^{1-\alpha} &\leq \sum_{t=1}^{T} \frac{\Delta_i^{1-2\alpha}\frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}}{2}\left(\alpha\left(\frac{16}{\Delta_i^2 t}\right)^{1-\alpha} + (1-\alpha)\left(\frac{16}{\Delta_i^2 t}\right)^{-\alpha}\mathbb{E}[w_{t,i}]\right) \\
&= \sum_{t=1}^{T} \frac{1-\bar{t}^{-1+\alpha}}{1-\alpha}\left(\frac{2\alpha}{\Delta_i t} + \frac{1-\alpha}{8}\Delta_i \mathbb{E}[w_{t,i}]\right) \\
&\leq \frac{1-T^{-1+\alpha}}{1-\alpha}\frac{2(\log(T)+1)}{\Delta_i} + \sum_{t=1}^{T}\frac{\Delta_i \mathbb{E}[w_{t,i}]}{8}. \qquad (3.21)
\end{aligned}$$

Finally, we bound the leading factor of the log term with Lemma 3.3:

$$\frac{1-T^{-1+\alpha}}{1-\alpha} \leq \min\{\frac{1}{1-\alpha}, \log(T)\}.$$

Bounding the linear part We first show that all $t > T_0 = \frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2})$ satisfy $\eta_t \xi_i \leq \frac{\Delta_i}{4}$.

$$\begin{aligned}
\eta_t \xi_i = \Delta_i^{1-2\alpha}\frac{16^\alpha}{4}\frac{1-\bar{t}^{-1+\alpha}}{(1-\alpha)t^\alpha} &< \frac{\Delta_i}{4}\left(\frac{16}{\Delta_{\min}^2 T_0}\right)^\alpha \frac{1-\overline{T_0}^{-1+\alpha}}{1-\alpha} \\
&\leq \frac{\Delta_i}{4}\frac{1-(\frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2}))^{-1+\alpha}}{(1-\alpha)(\log(\frac{16}{\Delta_{\min}^2}))^{2\alpha}}.
\end{aligned}$$

It remains to show that $\frac{1-(\frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2}))^{-1+\alpha}}{(1-\alpha)(\log(\frac{16}{\Delta_{\min}^2}))^{2\alpha}} \leq 1$. By Lemma 3.5 we have

$$\begin{aligned}\frac{1-(\frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2}))^{-1+\alpha}}{1-\alpha} &\leq \left(\log\left(\frac{16}{\Delta_{\min}^2}\log^2(\frac{16}{\Delta_{\min}^2})\right)\right)^{\alpha}\\ &\leq \left(2\log(\frac{16}{\Delta_{\min}^2})\right)^{\alpha} \leq \left(\log(\frac{16}{\Delta_{\min}^2})\right)^{2\alpha},\end{aligned}$$

which concludes the proof. Therefore,

$$\sum_{i\neq i^*}\sum_{t=T_0+1}^{T}\frac{\eta_t(\xi_i+2\xi_{i^*})}{2}\mathbb{E}[w_{t,i}] \leq \sum_{i\neq i^*}\sum_{t=1}^{T}\frac{\Delta_i+2\Delta_{\min}}{8}\mathbb{E}[w_{t,i}] \leq \sum_{i\neq i^*}\sum_{t=1}^{T}\frac{3\Delta_i\mathbb{E}[w_{t,i}]}{8}. \tag{3.22}$$

Bounding the constant part Recall $T_0 = \frac{16}{\Delta_{\min}^2}\log^2\left(\frac{16}{\Delta_{\min}^2}\right) \geq 16$. We can use the estimation $\sum_{t=1}^{T_0} t^{-\alpha} \leq 1 + \int_1^{T_0} t^{-\alpha}\,dt = 1 + \frac{T_0^{1-\alpha}-1}{1-\alpha} \leq 2\frac{T_0^{1-\alpha}-1}{1-\alpha}$:

$$\begin{aligned}\sum_{t=1}^{T_0}\frac{\eta_t\xi_{i^*}}{2} &\leq \sum_{t=1}^{T_0}\frac{\Delta_{\min}^{1-2\alpha}16^{\alpha}(1-T_0^{-1+\alpha})}{8(1-\alpha)t^{\alpha}} = \frac{\Delta_{\min}^{1-2\alpha}16^{\alpha}(1-T_0^{-1+\alpha})}{8(1-\alpha)}\sum_{t=1}^{T_0}\frac{1}{t^{\alpha}}\\ &\leq \frac{\Delta_{\min}^{1-2\alpha}16^{\alpha}(1-T_0^{-1+\alpha})(T_0^{1-\alpha}-1)}{4(1-\alpha)^2}\\ &= \frac{16^{\alpha}T_0^{1-\alpha}(1-T_0^{-1+\alpha})^2}{4\Delta_{\min}^{2\alpha-1}(1-\alpha)^2} = \frac{4\log^{2-2\alpha}(\frac{16}{\Delta_{\min}^2})}{\Delta_{\min}}\left(\frac{1-T_0^{-1+\alpha}}{1-\alpha}\right)^2\\ &\leq \frac{4\log^2(\frac{16}{\Delta_{\min}^2})\log^2(T_0)}{\Delta_{\min}} \leq \frac{4\log^4(T_0)}{\Delta_{\min}},\end{aligned} \tag{3.23}$$

where the last line uses Lemma 3.3. Combining (3.21), (3.22), and (3.23) we obtain:

$$\textit{stability} \leq \sum_{i\neq i^*}\left(\min\left\{\frac{1}{1-\alpha},\log(T)\right\}\frac{2(\log(T)+1)}{\Delta_i} + \sum_{t=1}^{T}\frac{\Delta_i\mathbb{E}[w_{t,i}]}{2}\right) + \frac{4\log^4(T_0)}{\Delta_{\min}}. \tag{3.24}$$

Bounding the *penalty* term For the *penalty* term we start with the second part of Lemma 3.2 with $x = 1$. We have

$$\begin{aligned}\textit{penalty} &= \mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i_T^*}\right]\\ &\leq \frac{1-T^{-\alpha}}{\alpha}\sum_{i\neq i^*}\left(\frac{\mathbb{E}[w_{1,i}]^{\alpha}-\alpha\mathbb{E}[w_{1,i}]}{\eta_1\xi_i(1-\alpha)} + \sum_{t=2}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\frac{\mathbb{E}[w_{t,i}]^{\alpha}-\alpha\mathbb{E}[w_{t,i}]}{(1-\alpha)\xi_i}\right) + 1\\ &\leq \sum_{i\neq i^*}\left(\underbrace{\frac{\mathbb{E}[w_{1,i}]^{\alpha}-\alpha\mathbb{E}[w_{1,i}]}{\eta_1\xi_i(1-\alpha)}}_{\textit{constant}}\log(T) + \underbrace{\sum_{t=2}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\frac{\mathbb{E}[w_{t,i}]^{\alpha}-\alpha\mathbb{E}[w_{t,i}]}{(1-\alpha)\alpha\xi_i}}_{\textit{concave}}\right) + 1,\end{aligned}$$

where in the first term we use $\frac{1-T^{-\alpha}}{\alpha} \leq \log(T)$ and in the second $\frac{1-T^{-\alpha}}{\alpha} \leq \frac{1}{\alpha}$, both bounds following from Lemma 3.3.

Bounding the concave term Since w^α is a concave function of w it can be upper bounded by the first order Taylor's approximation:

$$w_{t,i}{}^\alpha \leq w^{*\alpha} + \alpha w^{*\alpha-1}(w_{t,i} - w^*) = (1-\alpha)w^{*\alpha} + \alpha w^{*\alpha-1} w_{t,i}.$$

Taking $w^* = \frac{16}{\Delta_i^2 t}$ (with $\eta_t = \frac{16^\alpha}{4} \frac{1-\overline{t}^{-1+\alpha}}{(1-\alpha)t^\alpha}$ and $\xi_i = \Delta_i^{1-2\alpha}$):

$$\begin{aligned}
&\sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \frac{\mathbb{E}[w_{t+1,i}]^\alpha - \alpha \mathbb{E}[w_{t+1,i}]}{(1-\alpha)\alpha\xi_i} && (3.25)\\
&\quad \leq \sum_{t=1}^{T-1} \left(\frac{(t+1)^\alpha}{1-\overline{(t+1)}^{-1+\alpha}} - \frac{t^\alpha}{1-\overline{t}^{-1+\alpha}} \right) \\
&\qquad\quad \cdot \frac{4\Delta_i^{2\alpha-1}}{\alpha 16^\alpha} \left((1-\alpha) \left(\frac{16}{\Delta_i^2 t}\right)^\alpha + \alpha \left(\frac{16}{\Delta_i^2 t}\right)^{\alpha-1} \mathbb{E}[w_{t+1,i}] - \alpha \mathbb{E}[w_{t+1,i}] \right) \\
&\quad \leq \sum_{t=1}^{T-1} \left(\frac{(t+1)^\alpha - t^\alpha}{1-\overline{t}^{-1+\alpha}} \right) \left(\frac{4(1-\alpha)}{\alpha \Delta_i t^\alpha} + \frac{\Delta_i \mathbb{E}[w_{t+1,i}]}{4t^{\alpha-1}} \left(1 - \left(\frac{16}{\Delta_i^2 t}\right)^{1-\alpha} \right) \right) \\
&\quad \text{(by Taylor's approximation } (t+1)^\alpha \leq t^\alpha + \alpha t^{\alpha-1} \text{ and also use } \frac{16}{\Delta_i^2} > 1) \\
&\quad \leq \sum_{t=1}^{T-1} \left(\frac{\alpha t^{\alpha-1}}{1-\overline{t}^{-1+\alpha}} \right) \left(\frac{4(1-\alpha)}{\alpha \Delta_i t^\alpha} + \frac{\Delta_i \mathbb{E}[w_{t+1,i}]}{4t^{\alpha-1}} \left(1 - \left(\frac{1}{t}\right)^{1-\alpha} \right) \right) \\
&\quad \leq \sum_{t=1}^{T-1} \left(\frac{1-\alpha}{1-e^{-1+\alpha}} \frac{4}{\Delta_i t} + \frac{\Delta_i \mathbb{E}[w_{t+1,i}]}{4} \frac{1-t^{-1+\alpha}}{1-\overline{t}^{-1+\alpha}} \right) \\
&\quad \leq \frac{1-\alpha}{1-e^{-1+\alpha}} \frac{4(\log(T)+1)}{\Delta_i} + \sum_{t=1}^{T} \frac{\Delta_i \mathbb{E}[w_{t,i}]}{4} \\
&\quad \leq \frac{8(\log(T)+1)}{\Delta_i} + \sum_{t=1}^{T} \frac{\Delta_i \mathbb{E}[w_{t,i}]}{4}. && (3.26)
\end{aligned}$$

The last step follows from the leading factor $\frac{1-\alpha}{1-e^{-1+\alpha}}$ being bounded by 2.

Bounding the constant term Since $\frac{w^\alpha - \alpha w}{1-\alpha}$ is monotonically decreasing in α and ξ_i is monotonically increasing in α, we have

$$\frac{\mathbb{E}[w_{1,i}]^\alpha - \alpha \mathbb{E}[w_{1,i}]}{(1-\alpha)\eta_1 \xi_i} \leq \frac{4}{\Delta_i(1-e^{-1})} \leq \frac{8}{\Delta_i}. \tag{3.27}$$

Combining (3.26) and (3.27) we obtain

$$\mathit{penalty} \leq \sum_{i \neq i^*} \left(\frac{16(\log(T)+1)}{\Delta_i} + \sum_{t=1}^{T} \frac{\Delta_i \mathbb{E}[w_{t,i}]}{4} \right) + 1. \tag{3.28}$$

Finishing the proof Finally, we combine (3.24) and (3.28), rearrange the terms to get

$$\begin{aligned}\overline{Reg}_T &\leq \sum_{t=1}^{T}\sum_{i\neq i^*}\frac{3\Delta_i\mathbb{E}[w_{t,i}]}{4}+\sum_{i\neq i^*}\left(\frac{\left(2\min\{\frac{1}{1-\alpha},\log(T)\}+16\right)(\log(T)+1)}{\Delta_i}\right)\\&\qquad+\frac{4\log^4(T_0)}{\Delta_{\min}}+1.\\&=\frac{3\overline{Reg}_T}{4}+\sum_{i\neq i^*}\left(\frac{\left(2\min\{\frac{1}{1-\alpha},\log(T)\}+16\right)(\log(T)+1)}{\Delta_i}\right)+\frac{4\log^4(T_0)}{\Delta_{\min}}+1.\end{aligned}$$

Rearranging and multiplying by 4 finishes the proof. □

Chapter 4

Combinatorial semi-bandits

The work presented in this chapter is based on a paper that has been accepted as [109].

[109] Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*

Abstract

We develop the first general semi-bandit algorithm that simultaneously achieves $\mathcal{O}(\log T)$ regret for stochastic environments and $\mathcal{O}(\sqrt{T})$ regret for adversarial environments without prior knowledge of the regime or the number of rounds T. The leading problem-dependent constants of our bounds are not only optimal in a certain worst-case sense studied previously, but also optimal for two concrete instances of semi-bandit problems. Our algorithm and analysis extend the recent work of Zimmert and Seldin [111] for the special case of multi-armed bandits, but importantly requires a novel hybrid regularizer designed specifically for semi-bandit. Experimental results on synthetic data show that our algorithm indeed performs well over different environments. Finally, we provide a preliminary extension of our results to the full bandit feedback.

4.1 Introduction

The multi-armed bandit is one of the most fundamental online learning problems with partial information feedback. In this problem a learner repeatedly selects one of d arms and observes its loss generated by the environment, with the goal of minimizing her *regret*, the difference between her total loss and the loss of the best fixed arm in hindsight. It is well known that in the stochastic environment where each arm's loss is drawn independently from a fixed distribution, the minimax optimal regret is of order $\mathcal{O}(\log T)$ where T is the number of rounds (dependence on all other parameters is omitted) [67], while in the adversarial environment where each arm's loss can be completely arbitrary, the minimax optimal regret is of order $\mathcal{O}(\sqrt{T})$ [17].

Several recent works [18, 31, 90, 91, 103, 111] develop "best-of-both-worlds" results for multi-armed bandits and propose adaptive algorithms that achieve $\mathcal{O}(\log T)$ regret in stochastic environments while simultaneously ensuring worst-case robustness, that is, $\mathcal{O}(\sqrt{T})$ regret even for adversarial environments. Importantly, this is achieved without any prior knowledge of the nature of the environment.

In this work, we extend such best-of-both-worlds results to the combinatorial bandit problem, a generalization of multi-armed bandits, where the learner has to pick a subset of arms (called a combinatorial action) at each time (see Section 4.2 for formal definitions). In particular, we consider the *semi-bandit* feedback, meaning that the learner observes the loss of each arm in the selected subset. Our main contributions include the following:

1. We propose a simple and general semi-bandit algorithm based on the *Follow-the-Regularized-Leader* (FTRL) framework with a novel regularizer (Section 4.2.1).

2. For any combinatorial action set, we prove that our algorithm achieves $\mathcal{O}(C_{sto} \log T)$ regret for stochastic environments and $\mathcal{O}(C_{adv}\sqrt{T})$ regret for adversarial environments, where C_{sto} and C_{adv} are problem-dependent factors (that do not depend on T) and are optimal in some worst-case sense. This is the first best-of-both-worlds result for combinatorial bandit to the best of our knowledge (Section 4.3.1).

3. For two common special cases of combinatorial action sets: the set of all subsets of arms and the set of all subsets with a fixed size m (so called m-set), we further derive refined bounds for the problem-dependent constants C_{sto} and C_{adv}, which are optimal for each of these special cases. As a side result, our bounds imply that for the m-set with $m > d/2$, semi-bandit feedback is no harder than full-information feedback in the adversarial case (Sections 4.3.2 and 4.3.3).

4. We conduct experiments with synthetic data to show that our algorithm indeed adapts well to the nature of the environment. Additionally, we present a simple intermediate

setting where our algorithm outperforms all baselines (Section 4.4).

5. We also provide a preliminary extension of our results to a special case of the more challenging bandit feedback (Section 4.6).

Our techniques are close to those of [111]: we make use of the Ftrl algorithm, a well-known framework for adversarial environments, and show that with a simple time-decaying learning rate schedule (that is, $1/\sqrt{t}$ for time t), the regret admits a certain *self-bounding* property under the stochastic environment which eventually leads to logarithmic regret in this case. Importantly, however, our results require the use of a novel *hybrid* regularizer, designed specifically for semi-bandit. Roughly speaking, the idea is that for arms outside of the optimal subset, the problem of identifying their suboptimality is analogous to the multi-armed bandit problem, and we apply the regularizer of Zimmert and Seldin [111] to these arms; and on the other hand for arms in the optimal subset, the problem behaves like the full-information expert problem [50], and we thus apply the classical Shannon entropy as the regularizer to these arms.

4.1.1 Related work

Semi-bandits. The combinatorial semi-bandit problem is a natural generalization of multi-armed bandits and captures many real-life applications. There are many algorithms for stochastic semi-bandits based on the well-known optimistic principle [39, 43, 51, 65]. Optimistic algorithms are provably not instance-optimal [69] and a recent work developed a general instance-optimal algorithm for any structured stochastic bandits (including semi-bandit as a special case [42]). Specifically, they obtain the optimal regret $\mathcal{O}(C \log T)$ where C is an instance-dependent term expressed as the solution of a certain optimization problem. The constant C_{sto} in our stochastic bound $\mathcal{O}(C_{sto} \log T)$ is also expressed as an optimization problem (see Theorem 4.1), but it is not clear how it compares to the instance-optimal constant C in general, except for the two special cases we discuss in Section 4.3. Two advantages of our algorithm compared to prior work are: a) our stochastic assumption is weaker than others (see Section 4.2) and b) our algorithm ensures worst-case robustness even when the stochastic assumption does not hold.

Algorithms with $\mathcal{O}(\sqrt{T})$ regret for the adversarial semi-bandit setting are also well-studied [15, 43, 79, 80, 103]. These algorithms are either based on Follow-the-Regularized-Leader (equivalently Online Mirror Descent) or Follow-the-Perturbed-Leader, both of which are standard frameworks for designing adversarial online learning algorithms (see Hazan [56] for an introduction). It is easy to show that even if the environment is stochastic, the regret of these algorithms is still $\Theta(\sqrt{T})$, indicating the lack of adaptivity. Moreover, even for the adversarial case the leading constant in previous bounds is only worst-case optimal but not instance-optimal. In contrast, our adversarial regret bound $\mathcal{O}(C_{adv}\sqrt{T})$ is instance-dependent through the term C_{adv}, again expressed as the solution of a certain optimization problem (see Theorem 4.1). To the best of our knowledge, there is no known general instance-dependent lower bound for this term, but again we show the optimality of our bound in two special cases in Section 4.3.

Best-of-both-worlds. Algorithms that are optimal for both stochastic and adversarial environments were studied for multi-armed bandits [18, 31, 90, 91, 103, 111], and also for the easier full-information (the expert problem) [52, 63, 74] and intermediate version [97]. Notably, among these works the recent two [103, 111] discovered that sophisticated hypothesis testing or gap estimations used in earlier works are in fact not needed for such adaptivity. Instead, their algorithms are based on the Ftrl framework with special regularizers. As mentioned, our work also follows this route by designing a new regularizer for the more general semi-bandit setting.

Hybrid regularizers. The idea of using hybrid regularizers for FTRL was first proposed by Bubeck et al. [27] for sparse bandit and bandit with a specific form of adaptive regret bound, and also recently used by Luo et al. [75] for the online portfolio selection problem. The form of the hybrid regularizers and the way they are used in the analysis, however, are different both among these two prior works and with ours.

4.2 Problem Setting and Algorithm

The semi-bandit problem is a sequential game between a learner and an environment with d fixed arms. We call a subset of arms a combinatorial action,[1] and the learner is given a fixed set of combinatorial actions $\mathcal{X} \subset \{0,1\}^d$. At any time $t = 1, 2, \dots$, the learner chooses an action $X_t \in \mathcal{X}$ and at the same time the environment chooses a loss vector $\ell_t \in [-1,1]^d$. The learner suffers the loss $\langle X_t, \ell_t \rangle$ and receives the feedback $o_t = X_t \circ \ell_t$, where $\circ$ stands for the element-wise multiplication. In other words, the learner only observes the loss of each arm in the selected subset (the so-called semi-bandit feedback).

The environment can be either *stochastic* or *adversarial*. In the stochastic case, we adopt and extend the broader "stochastically constrained adversarial setting" [103, 111] and assume that there is a fixed action $x^* \in \mathcal{X}$ such that for any $x \in \mathcal{X} \backslash \{x^*\}$ there exists a constant $\Delta_x > 0$, such that $\mathbb{E}[\langle x - x^*, \ell_t \rangle] \geq \Delta_x$ for all t. Note that this clearly subsumes the traditional stochastic setting where $\ell_1, \dots, \ell_T$ are i.i.d. samples from a fixed unknown distribution, and is much more general since neither independence nor identical distributions are required. In the adversarial case, on the other hand, ℓ_t is chosen in an arbitrary way based on the history $\ell_1, X_1, \dots, \ell_{t-1}, X_{t-1}$ and possibly an internal randomization by the environment.

The performance of a learner is measured by *pseudo-regret*:

$$\mathfrak{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t - x^*, \ell_t \rangle\right],$$

where $x^* = \arg\min_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \langle x, \ell_t \rangle\right]$ is the best action in hindsight and the expectation is with respect to the randomness of both the learner and the environment. Note that in the stochastic case we are overloading the notation x^* since clearly they are the same action.

It is well known that in terms of the dependence on T, the optimal regret is $\Theta(\log T)$ in the stochastic case and $\Theta(\sqrt{T})$ in the adversarial case (see, for example, Audibert et al. [15], Combes et al. [42]).

Notations. We denote by $\mathbb{E}_t[\cdot]$ the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ where $\mathcal{F}_t$ is the filtration $\sigma(X_1, o_1, \dots, X_t, o_t)$. We also use a shorthand $\mathbb{I}_t(i)$ for the indicator function $\mathbb{I}\{X_{ti} = 1\}$ (X_{ti} is the i-th component of the vector $X_t \in \mathcal{X} \subset \{0,1\}^d$) and write the characteristic function of a set A as $\mathcal{I}_A(x)$ which is 0 if $x \in A$ and $+\infty$ otherwise. We denote the d-dimensional vector with all 1s as $\mathbf{1}_d$.

4.2.1 Our algorithm

Our algorithm is based on the general FTRL framework.[2] In this framework, each time the algorithm computes the regularized leader $x_t = \arg\min_{x \in \mathrm{conv}(\mathcal{X})} \langle x, \hat{L}_{t-1} \rangle + \eta_t^{-1} \Psi(x)$, where $\mathrm{conv}(\mathcal{X})$ is the convex hull of $\mathcal{X}$, $\hat{L}_{t-1} = \sum_{s=1}^{t-1} \hat{\ell}_s$ is the cumulative estimated loss, $\eta_t > 0$ is a

[1] In some works a combinatorial action is also referred to as "an arm", but here we exclusively use the term "arm" for one of the d elements and "combinatorial action" for a subset of these elements.

[2] For linear objectives and Legendre regularizers, FTRL is equivalent to Online Mirror Descent as defined in [81]. The same framework is also known under the names OMD, OSMD, or INF.

Algorithm 6: FTRL with hybrid regularizer for semi-bandits

```
Input: $0 < \gamma \leq 1$, sampling scheme $P$
Initialize: $\hat{L}_0 = (0, \dots, 0), \eta_t = 1/\sqrt{t}$
for $t = 1, 2, \dots$ do
    compute $x_t = \arg\min_{x \in \text{conv}(\mathcal{X})} \langle x, \hat{L}_{t-1} \rangle + \eta_t^{-1} \Psi(x)$, where $\Psi(\cdot)$ is defined in Eq. (4.1)
    sample $X_t \sim P(x_t)$
    observe $o_t = X_t \circ \ell_t$
    construct estimator $\hat{\ell}_t$, $\forall i : \hat{\ell}_{ti} = \frac{(o_{ti}+1)\mathbb{I}_t(i)}{x_{ti}} - 1$
    update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
```

learning rate, and $\Psi(x) : \text{conv}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ is a regularizer. Then the algorithm samples $X_t \sim P(x_t)$ for a sampling rule P that provides a distribution over $\mathcal{X}$ satisfying $\mathbb{E}_{X \sim P(x)}[X] = x$. As long as $\text{conv}(\mathcal{X})$ can be described by a polynomial number of constraints, one can always find an efficient sampling rule P (see concrete examples in Section 4.3). Finally, the algorithm constructs a loss estimator $\hat{\ell}_t$ based on the observed information and proceeds to the next round.

The novelty of our algorithm lies in the use of the hybrid regularizer

$$\Psi(x) = \sum_{i=1}^{d} -\sqrt{x_i} + \gamma(1 - x_i)\log(1 - x_i) \tag{4.1}$$

with a parameter $0 < \gamma \leq 1$ to be chosen later based on the action set $\mathcal{X}$ (in most cases we use $\gamma = 1$). This is a combination of the Tsallis entropy (with power 1/2) $\sum_i -\sqrt{x_i}$, and the Shannon entropy $\sum_i (1 - x_i)\log(1 - x_i)$ on the *complement* of x. The $\sum_i -\sqrt{x_i}$ regularizer was first implicitly introduced by Audibert and Bubeck [14], and later discovered as a member of the Tsallis entropy regularizers by Abernethy et al. [6]. It was also recently shown to be optimal for both stochastic and adversarial multi-armed bandits [111].

In addition, similar to Zimmert and Seldin [111], our algorithm uses a very simple time-decaying learning rate schedule $\eta_t = \eta$. The loss estimators $\hat{\ell}_t$ are defined as $\hat{\ell}_{ti} = \frac{(o_{ti}+1)\mathbb{I}_t(i)}{x_{ti}} - 1$ for all i. It is clear the estimators are unbiased, $\mathbb{E}_t[\hat{\ell}_t] = \ell_t$, just as common importance weighted estimators. The shift by 1 is used to ensure that the range of the loss estimates is bounded from one side, $\hat{\ell}_{t,i} \geq -1$. See Algorithm 6 for a complete pseudocode.

Intuition behind the new regularizer. It is known that the classical Shannon entropy regularizer [50] is optimal for both adversarial and stochastic environments in the full-information setting. In fact, the Shannon entropy on the *complement* of x is also optimal for full-information. This can be verified by considering the complementary problem: the problem with action set $\mathbf{1}_d - \mathcal{X}$ and reversed losses $-\ell_t$. Both problems describe the exact same game with the same information, and using Shannon entropy in the complementary problem is the same as using it on the complement of x in the original problem.

The intuition behind combining Tsallis and Shannon entropy is that when x_i is close to 0, the learner is starved of information and has to act similarly to a regular bandit problem. The magnitude of the gradient and its slope in that regime are dominated by the Tsallis entropy, which again is known to be optimal for bandits.

On the other hand, when x_i is close to 1, the game resembles a full-information game, and Shannon entropy on the complement becomes the dominating part of the regularizer in that regime. Effectively, this allows us to regularize arms in the optimal combinatorial set differently than arms outside the optimal set, without the need to know which arms are in the optimal set.

4.3 Main Results

In this section we present general regret guarantees for our algorithm, followed by concrete instantiations in two special cases.

4.3.1 Arbitrary action set

To state the general regret bound for our algorithm for any arbitrary action set $\mathcal{X}$, we define the following two functions:

$$f(x) = \sum_{i:x_i^*=0} \sqrt{x_i}$$
$$g(x) = \sum_{i:x_i^*=1} (\gamma^{-1} - \gamma \log(1 - x_i))(1 - x_i)$$

and the instantaneous regret function $r : [0,\infty)^{|\mathcal{X}|} \to \mathbb{R}$ as

$$r(\alpha) = \sum_{x \in \mathcal{X} \setminus \{x^*\}} \alpha_x \Delta_x$$

(recall the definition of x^* and Δ_x from Section 4.2). We also define $\overline{\alpha} = \sum_{x\in\mathcal{X}} \alpha_x x$ for any $\alpha \in [0,\infty)^{|\mathcal{X}|}$, and let $\Delta(\mathcal{X})$ denote the simplex of distributions over $\mathcal{X}$.

Theorem 4.1. *For any $\gamma \leq 1$ the pseudo regret of Algorithm 6 is upper bounded by*

$$\mathfrak{R}_T \leq \mathcal{O}\left(C_{sto} \log T\right) + \mathcal{O}\left(C_{add}\right)$$

in the stochastic case and

$$\mathfrak{R}_T \leq \mathcal{O}\left(C_{adv}\sqrt{T}\right)$$

in the adversarial case, where C_{sto}, C_{add} and C_{adv} are defined as

$$C_{sto} := \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} f(\overline{\alpha}) - r(\alpha),$$
$$C_{add} := \sum_{t=1}^{\infty} \max_{\alpha\in\Delta(\mathcal{X})} \left(\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha)\right),$$
$$C_{adv} := \max_{x\in\mathrm{conv}(\mathcal{X})} f(x) + g(x).$$

Moreover, it always holds that $C_{sto} = \mathcal{O}\left(\frac{md}{\Delta_{\min}}\right)$, $C_{add} = \mathcal{O}\left(\frac{m^2}{\gamma^2\Delta_{\min}}\right)$, and $C_{adv} = \mathcal{O}\left(\frac{1}{\gamma}\sqrt{md}\right)$, where $m = \max_{x\in\mathcal{X}} ||x||_1$ and $\Delta_{\min} = \min_{x\in\mathcal{X}\setminus\{x^*\}} \Delta_x$.

We defer the proof to Section 4.5. The dependence of our bounds on T is optimal in both cases. The leading problem-dependent constants C_{sto} and C_{adv} are expressed as solutions to optimization problems. Recent works [42, 43, 69] also expressed the instance-optimal leading constant in the stochastic case in a similar way, but it is not clear how to compare the results.

The explicit upper bounds on these constants stated at the end of the theorem immediately imply that for $\gamma = 1$ our bounds are worst-case optimal according to [65] and [15]. Here, worst-case optimality refers to the minimax regret over all environments with the same value m of $\max_{x\in\mathcal{X}} ||x||_1$ and also the same value $\Delta_{\min}$ of $\min_{x\in\mathcal{X}\setminus\{x^*\}} \Delta_x$ in the stochastic case.

However, for explicit instances, one can hope to achieve even better bounds. By exploiting the structure of the problem and providing better bounds on the constants C_{sto}, C_{add} and C_{adv}, we show in the next two sections that our algorithm is optimal in two special cases. For better interpretability, in the stochastic case we consider the more traditional setting where $\ell_1, \ldots, \ell_T$ are i.i.d. samples from an unknown distribution $\mathcal{D}$. It is clear that we can define $\Delta_x = \mathbb{E}_{\ell\sim\mathcal{D}}[\langle x - x^*, \ell\rangle]$ in this case.

4.3.2 Special case: full combinatorial set

The simplest semi-bandit problem is when $\mathcal{X} = \{0,1\}^d$, that is, the learner can pick any subset of arms. In this case $\operatorname{conv}(\mathcal{X}) = [0,1]^d$ and a trivial sampling rule is $P(x) = \bigotimes_{i=1}^d \operatorname{Ber}(x_i)$ where $\operatorname{Ber}(\cdot)$ stands for Bernoulli distribution.

It is clear that in this case each dimension/arm can be treated completely independently. Note, however, that the problem of each dimension is not exactly a two-armed bandit problem since the loss of "not choosing the arm" is known to be 0, and the problem is asymmetric between positive and negative losses. Specifically, we prove the following regret guarantee for our algorithm, where in the stochastic case with a slight abuse of notation we define $\Delta_i = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell_i]$.

Theorem 4.2. *If $\mathcal{X} = \{0,1\}^d$, the pseudo-regret of Algorithm 6 with $\gamma = 1$ is*

$$\mathfrak{R}_T \le \mathcal{O}\left(\sum_{\Delta_i > 0} \frac{\log(T)}{\Delta_i} \right) + \mathcal{O}\left(\sum_{\Delta_i < 0} \frac{1}{|\Delta_i|} \right)$$

in the stochastic case and

$$\mathfrak{R}_T \le \mathcal{O}\left(d\sqrt{T} \right)$$

in the adversarial case. Moreover, both bounds are optimal.

Proof. Note that in this case the algorithm is equivalent to the following: for each coordinate, run a copy of Algorithm 6 for a one-dimensional problem with $\mathcal{X} = \{0,1\}$ as the action set. We can thus apply Theorem 4.1 to such one-dimensional problems and finally sum up the regret along each coordinate. Below we focus on a fixed coordinate i.

In particular, in the stochastic case, if $\Delta_i > 0$, it implies $x_i^* = 0$ and thus $g(\cdot) \equiv 0$ and $C_{add} = \sum_t \max_{\alpha \in [0,1]} -\alpha \Delta_i = 0$. For C_{sto} we apply the general bound from Theorem 4.1 and obtain $C_{sto} = \mathcal{O}(1/\Delta_i)$ (since $m = d = 1$ and $\Delta_{\min} = \Delta_i$). This gives the bound $\mathcal{O}\left(\frac{\log(T)}{\Delta_i}\right)$ for $\Delta_i > 0$.

On the other hand if $\Delta_i < 0$ then $x_i^* = 1$ and $f(\cdot) \equiv 0$, so $C_{sto} = \max_{\alpha \ge 0} \alpha \Delta_i = 0$. For C_{add} we apply the general bound from Theorem 4.1 and obtain $C_{add} = \mathcal{O}(1/\Delta_i)$ (since $m = \gamma = 1$ and $\Delta_{\min} = \Delta_i$). This gives the bound $\mathcal{O}\left(\frac{1}{\Delta_i}\right)$ for $\Delta_i < 0$.

In the adversarial case, we apply the general bound of Theorem 4.1 and obtain $C_{adv} = \mathcal{O}(1)$. This finishes the proof for the regret upper bounds. The optimality of the adversarial bound is trivial since it matches the full-information lower bound. Obtaining a matching lower bound in the stochastic regime is a simple adaptation of the regular two-armed bandit lower bound. We believe this result is well known, but provide a proof in the appendix in absence of a reference. □

4.3.3 Special case: m-set

Another common instance of semi-bandit is when the learner can only select subsets of a fixed size. Specifically, let $m \in \{1, \dots, d-1\}$ be a fixed parameter and define the m-set as

$$\mathcal{X} = \left\{ x \in \{0,1\}^d \;\middle|\; \sum_{i=1}^d x_i = m \right\}. \tag{4.2}$$

Note that we are overloading the notation $m = \max_{x \in \mathcal{X}} ||x||_1$ since clearly they are the same in this case. It is well-known that the convex hull of m-set is $\operatorname{conv}(\mathcal{X}) = \left\{ x \in [0,1]^d \mid \sum_{i=1}^d x_i = m \right\}$, and in the appendix we provide a simple sampling rule P with $\mathcal{O}(d \log(d))$ time complexity. This improves over previous work that requires $\mathcal{O}(d^2)$ time complexity [95, 102].

In the stochastic case, we assume without loss of generality that the expected losses of arms are increasing in i. Overloading the notation again we define the stochastic gaps as $\Delta_i = \mathbb{E}_{\ell\sim\mathcal{D}}\left[\ell_i - \ell_m\right]$ for all i. Note that the uniqueness of x^* also implies $\Delta_i \neq 0$ for all $i > m$. The next theorem shows that our algorithm is optimal for both environments. As a side result, we also show that when $m > d/2$, semi-bandit feedback is no harder than full-information feedback in the adversarial case. To the best of our knowledge, this was previously unknown.

Theorem 4.3. *If $\mathcal{X}$ is the m-set defined by Eq.* (4.2)*, then the pseudo-regret of Algorithm 6 with*

$$\gamma = \begin{cases} 1 & \text{if } m \le d/2 \\ \min\{1, 1/\sqrt{\log(d/(d-m))}\} & \text{otherwise,} \end{cases}$$

satisfies

$$\mathfrak{R}_T \le \mathcal{O}\left(\sum_{i=m+1}^{d} \frac{\log(T)}{\Delta_i}\right) + \mathcal{O}\left(\sum_{i=m+1}^{d} \frac{(\log d)^2}{\Delta_i}\right)$$

in the stochastic case and

$$\mathfrak{R}_T \le \begin{cases} \mathcal{O}\left(\sqrt{mdT}\right) & \text{if } m \le d/2 \\ \mathcal{O}\left((d-m)\sqrt{\log(\frac{d}{d-m})T}\right) & \text{otherwise} \end{cases}$$

in the adversarial case. Moreover, both bounds are optimal.

Proof sketch. We provide a proof sketch here and defer some details to Section 4.7.3.

$\mathbf{C_{adv}}$: The optimization problem is concave in x and symmetric for all i with the same value of x_i^*. Therefore the optimal solution takes the form

$$\left(\operatorname*{arg\,max}_{x\in\operatorname{conv}(\mathcal{X})} f(x) + g(x)\right)_i = \begin{cases} \lambda & \text{if } x_i^* = 0 \\ 1 - \frac{d-m}{m}\lambda & \text{if } x_i^* = 1 \end{cases}$$

for some $\lambda \in [0, \min\{1, \frac{m}{d-m}\}]$. In Section 4.7.3 we show that the function is increasing in λ, and that inserting $\lambda = \min\{1, \frac{m}{d-m}\}$ leads to the stated adversarial bound.

$\mathbf{C_{sto}}$: With the definitions of the gaps, we can express $\Delta_x = \sum_{i:x_i\neq x_i^*} |\Delta_i|$, which is lower bounded by $\sum_{i:x_i^*=0,x_i=1}\Delta_i = \sum_{i:x_i^*=0}\Delta_i x_i$. So the immediate regret function $r(\alpha)$ can be bounded as

$$\begin{aligned} r(\alpha) &= \sum_{x\neq x^*} \Delta_x \alpha_x \ge \sum_{x\neq x^*}\sum_{i:x_i^*=0} \Delta_i \alpha_x x_i \\ &= \sum_{i:x_i^*=0} \Delta_i \left(\sum_{x\neq x^*} \alpha_x x_i\right) = \sum_{i:x_i^*=0} \Delta_i \overline{\alpha}_i. \end{aligned}$$

The optimization problem can now be bounded as

$$\begin{aligned} C_{sto} &= \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \sum_{i:x_i^*=0} \sqrt{\overline{\alpha}_i} - \sum_{x\neq x^*} \alpha_x \Delta_x \\ &\le \max_{\overline{\alpha}\in[0,\infty)^d} \sum_{i:x_i^*=0} \left(\sqrt{\overline{\alpha}_i} - \Delta_i \overline{\alpha}_i\right) = \sum_{i:x_i^*=0} \frac{1}{4\Delta_i}, \end{aligned}$$

which is the same as $\sum_{i=m+1}^{d} \frac{1}{4\Delta_i}$.

$\mathbf{C_{add}}$: We bound the function g as follows:

$$
\begin{aligned}
g(\overline{\alpha}) &= \sum_{i:x_i^*=1} (\gamma^{-1} - \gamma \log(1-\overline{\alpha}_i))(1-\overline{\alpha}_i) \\
&\le \left(\gamma^{-1} - \gamma \log\left(\sum_{i:x_i^*=1} \frac{1-\overline{\alpha}_i}{m}\right)\right) \sum_{i:x_i^*=1} (1-\overline{\alpha}_i) \\
&= \left(\gamma^{-1} - \gamma \log\left(\sum_{i:x_i^*=0} \frac{\overline{\alpha}_i}{m}\right)\right) \sum_{i:x_i^*=0} \overline{\alpha}_i \\
&\le \sum_{i:x_i^*=0} \left(\gamma^{-1} - \gamma \log\left(\frac{\overline{\alpha}_i}{m}\right)\right) \overline{\alpha}_i
\end{aligned}
$$

where the first inequality is by the concavity of g; the second equality is by the fact $\sum_{i:x_i^*=1} 1 - \overline{\alpha}_i = \sum_{i:x_i^*=0} \overline{\alpha}_i$ since $\overline{\alpha}$ is in the convex hull of m-set.

Recall the lower bound $r(\alpha) \ge \sum_{i:x_i^*=0} \Delta_i \overline{\alpha}_i$ as derived previously. We can thus bound C_{add} as

$$
\sum_{i:x_i^*=0} \sum_{t=1}^{\infty} \max_{A\in[0,1]} \frac{100}{\sqrt{t}} \left(\gamma^{-1} - \gamma \log\left(\frac{A}{m}\right)\right) A - \Delta_i A
$$

Solving the one-dimensional optimization problems above independently for each i (see Section 4.7.3) proves $C_{add} \le \mathcal{O}\left(\sum_{i:x_i^*=0} \frac{(\log d)^2}{\Delta_i}\right)$.

Optimality: The optimality for the stochastic case is implied by [11, 42]. For the adversarial case, only a matching lower bound $\Omega(\sqrt{mdT})$ for $m \le d/2$ is known (Theorem 2 of [68]). We close this gap by making a simple observation that when $m > d/2$, our bound in fact matches the lower bound of the same problem with full-information feedback. This clearly implies the optimality of our bound since semi-bandit feedback is harder.

Indeed, Koolen et al. [64] prove the lower bound $\Omega(m\sqrt{T \log(d/m)})$ for full-information m-set when $m \le d/2$. When $m > d/2$, one can simply work on the complementary problem with action set $\mathbf{1}_d - \mathcal{X}$ and reversed losses. This is exactly a $(d-m)$-set problem and thus a lower bound $\Omega((d-m)\sqrt{T \log(d/(d-m))})$ applies. This exactly matches our upper bound. □

4.4 Empirical Comparisons

We compare our novel algorithm with four baselines from the literature. For stochastic algorithms, we choose CombUCB [65] and Thompson Sampling [53]; for adversarial algorithms, we choose Exp2 [15] and LogBarrier [103], which are respectively Ftrl with generalized Shannon entropy and log-barrier regularizer. For each adversarial algorithm, we tune the time-independent part of the learning rate by choosing from the grid of $\{2^i | i \in \{-5, -4, \ldots, 5\}\}$, and the optimal value happens to be identical for both adversarial and stochastic environment in our experiments. Specifically the final learning rates η_t for our algorithm, Exp2 and LogBarrier are respectively $1/\sqrt{t}$, $1/(4\sqrt{t})$ and $4\sqrt{\log(t)/t}$.

We test the algorithms on concrete instances of the m-set problem with parameters: $d = 10$, $m = 5$, $T = 10^7$. Below, we specify the mean of each arm's loss at each time. With mean μ_{ti} the actual loss of arm i at time t will be -1 with probability $(1-\mu_{ti})/2$ and $+1$ with probability $(1+\mu_{ti})/2$, independent of everything else. We create the following two environments:

Stochastic environment. In this case the losses are drawn from a fixed distribution with $\mu_{ti} = -\Delta$ if $i \le 5$ and $\mu_{ti} = \Delta$ otherwise, where $\Delta = 1/8$.

"Adversarial" environment. Since it is difficult to create truly adversarial data, here we in fact use a stochastically constrained adversarial setting defined in Section 4.2. The construction is similar to that of Zimmert and Seldin [111]. Specifically, the time is split into phases

$$\underbrace{1, \ldots, t_1}_{T_1}, \underbrace{t_1 + 1, \ldots, t_2}_{T_2}, \ldots, \underbrace{t_{n-1}, \ldots, T}_{T_n}.$$

The length of phase s is $T_s = 1.6^s$, and the means of the losses are set to

$$\mu_{ti} = \begin{cases} -\Delta/2 \pm (1 - \Delta/2) & \text{if } i \leq 5, \\ +\Delta/2 \pm (1 - \Delta/2) & \text{otherwise,} \end{cases},$$

where ± represents + if t belongs to an odd phase and − otherwise. This model is not only a nice toy example, but could also be justified by real world applications. For example, in a network routing problem, an adversary might periodically attack the network, making the delay of every edge increase by roughly the same amount.

We measure the performance of the algorithms by the average pseudo-regret over at least 20 runs. For CombUCB and Thompson Sampling in the adversarial environment, we increase the number of runs to 500 and 1000 respectively due to the high variance of the pseudo-regret. Fig. 4.1 shows the average pseudo-regret of all algorithms at each time, where plot (a) uses the stochastic data and plot (b) uses the adversarial data. We use log-log scale after 10^4 rounds. Shaded areas in the plot show the confidence intervals.

The plots clearly confirm our theoretical results. Our algorithm outperforms Exp2 and LogBarrier (in the later stage) in both environments. In the stochastic case our algorithm is competitive with CombUCB, while Thompson Sampling has the best performance (a well-known phenomenon). However, these two stochastic algorithms clearly fail in the adversarial case and exhibit nearly-linear regret.

4.5 Proof of Main Theorem

We provide the key steps of the proof for our general result (Theorem 4.1) in this section. Define $\Psi_t(\cdot) = \eta_t^{-1}\Psi(\cdot)$ and potential function $\Phi_t(\cdot) = \max_{x \in \mathrm{conv}(\mathcal{X})} \langle x, \cdot \rangle - \Psi_t(x)$, which is the convex conjugate of $\Psi_t + \mathcal{I}_{\mathrm{conv}(\mathcal{X})}$.

Following a standard analysis of FTRL, we decompose the regret

$$\mathfrak{R}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, \ell_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{\mathfrak{R}_{\mathrm{stab}}}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \langle x^*, \ell_t \rangle\right]}_{\mathfrak{R}_{\mathrm{pen}}}, \tag{4.3}$$

into terms corresponding to the *stability* and the *regularization penalty* of the algorithm.

We then further bound these two terms respectively in the following two lemmas using mostly standard Ftrl analysis (see Section 4.7 for the proofs).

Lemma 4.1. *The regularization penalty is bounded as*

$$\mathfrak{R}_{\mathrm{pen}} \leq \sum_{t=1}^{T} \frac{3}{2\sqrt{t}} \Bigg(\sum_{i: x_i^* = 0} \sqrt{\mathbb{E}[x_{ti}]}$$
$$- \sum_{i: x_i^* = 1} \gamma(1 - \mathbb{E}[x_{ti}]) \log(1 - \mathbb{E}[x_{ti}]) \Bigg).$$

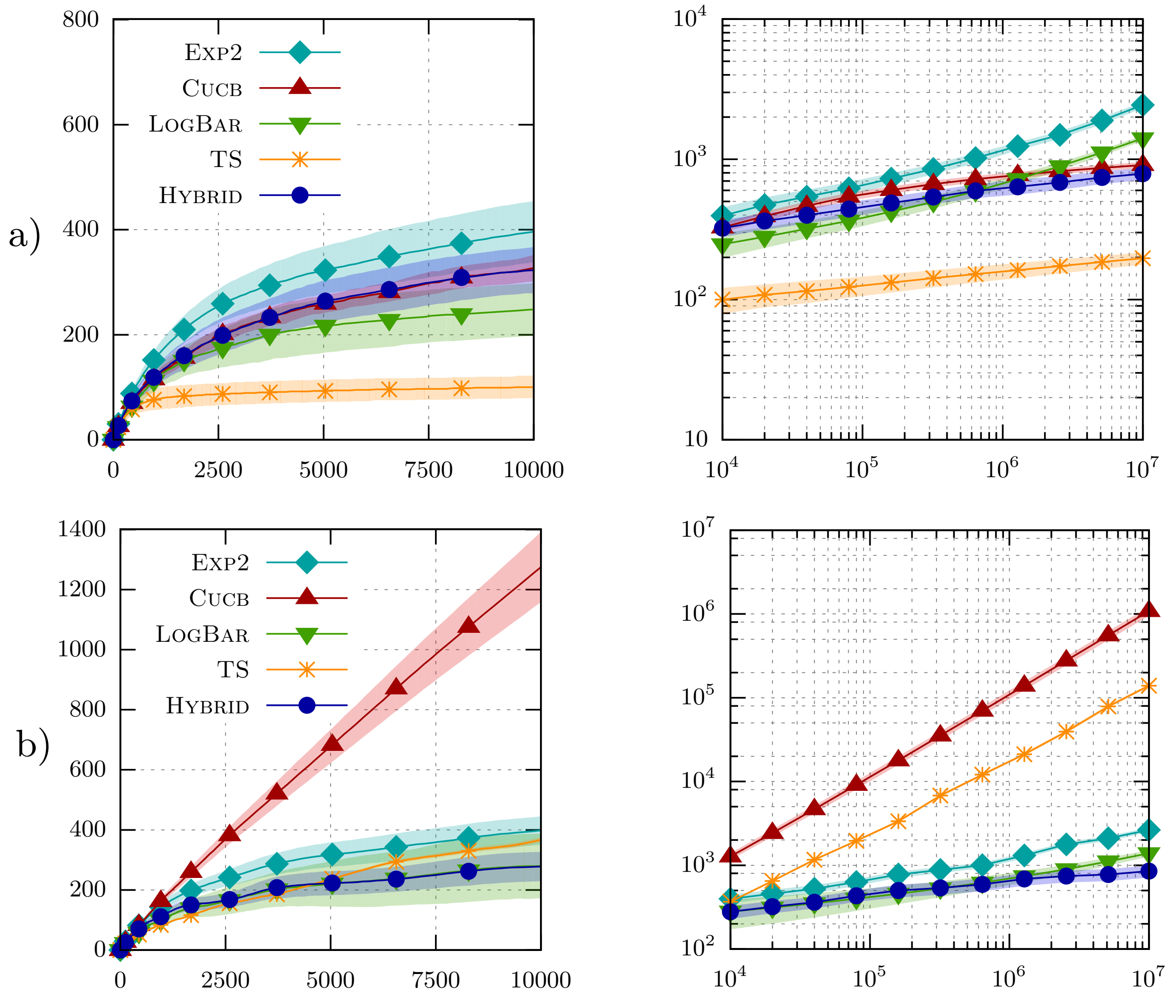


Figure 4.1: Comparisons of our new algorithm (Hybrid) and several existing algorithms with $d = 10, m = 5$ and $T = 10^7$ under a) stochastic and b) stochastically constrained adversarial setting. The left side is in linear scale and the right is in log-log scale.

Lemma 4.2. *The stability term is bounded as*

$$\begin{aligned}\mathfrak{R}_{\text{stab}} \leq \sum_{t=1}^{T} \frac{16\sqrt{2}}{\sqrt{t}} \Bigg(& \sum_{i:x_i^*=0} \sqrt{\mathbb{E}[x_{ti}]} \\ & + \sum_{i:x_i^*=1} \gamma^{-1}(1-\mathbb{E}[x_{ti}]) \Bigg) + c.\end{aligned}$$

where $c = 58m/\gamma^2$ (recall that $m = \max_{x\in\mathcal{X}} ||x||_1$).

We now proceed to the proof of Theorem 4.1.

Proof of Theorem 4.1. Using Lemma 4.1 and Lemma 4.2 in Eq. (4.3) and the definition of functions f and g, we can bound the regret by

$$\begin{aligned}\mathfrak{R}_T &\leq \sum_{t=1}^{T} \frac{25}{\sqrt{t}} \left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c \qquad (4.4)\\ &\leq 50\sqrt{T} \max_{x\in\text{conv}(\mathcal{X})} \left(f(x)+g(x)\right) + c \\ &= \mathcal{O}\left(C_{adv}\sqrt{T}\right),\end{aligned}$$

which concludes the adversarial case.

For the stochastic case we use a *self-bounding* technique similar to Wei and Luo [103], Zimmert and Seldin [111]. First, by the definition of the function r and the stochastic assumption we have

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbb{E}[x_t] - x^*, \ell_t\rangle\right] \geq \sum_{t=1}^{T} r(P(\mathbb{E}[x_t])).$$

Together with Eq. (4.4) we have

$$\sum_{t=1}^{T} \frac{25}{\sqrt{t}} \left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c - \sum_{i=1}^{T} r(P(\mathbb{E}[x_t])) \geq 0.$$

Combining the above with Eq. (4.4) again we bound $\mathfrak{R}_T$ by

$$\sum_{t=1}^{T} \left(\frac{50}{\sqrt{t}} \left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) - r(P(\mathbb{E}[x_t])) \right) + 2c.$$

We next decompose the summation above into two terms and upper bound them as $C_{sto}\log T$ and C_{add} respectively:

$$\begin{aligned}&\sum_{t=1}^{T} \frac{50}{\sqrt{t}} f(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t])) \\ &\leq \sum_{t=1}^{T} \max_{\alpha\in\Delta(\mathcal{X})} \frac{50}{\sqrt{t}} f(\overline{\alpha}) - \frac{1}{2} r(\alpha) \\ &\leq \sum_{t=1}^{T} \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \frac{50}{\sqrt{t}} f\left(\frac{10^4}{t}\overline{\alpha}\right) - \frac{1}{2} r\left(\frac{10^4}{t}\alpha\right) \\ &\overset{(\star)}{=} \sum_{t=1}^{T} \frac{10^4}{2t} \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} f(\overline{\alpha}) - r(\alpha) = \mathcal{O}\left(C_{sto}\log(T)\right)\end{aligned}$$

where ($\star$) follows since r is linear and f satisfies for any scalar $a \geq 0$: $f(ax) = \sqrt{a} f(x)$. On the other hand,

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{50}{\sqrt{t}} g(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t])) \\
&\leq \frac{1}{2} \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(\mathcal{X})} \left(\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha) \right) = \mathcal{O}(C_{add}),
\end{aligned}$$

where the last inequality uses the fact: for all $t > 0$, $\max_{\alpha \in \Delta(\mathcal{X})} \left(\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha) \right) \geq 0$. This is because a particular α that puts all the weight on x^* attains the value of 0.

The above finishes the proof of the general regret bounds. Due to space limitations we defer the derivation of upper bounds on the constants C_{sto}, C_{add} and C_{adv} to Section 4.7. □

4.6 Extensions to Bandit Feedback

The most natural extension of our work is to consider the full bandit feedback setting, where each time after playing an action X_t the learner only observes $\langle X_t, \ell_t \rangle$. Again, both stochastic and adversarial versions of the problem are well-studied in the literature, but there is no best-of-both-worlds result. Here, we provide a preliminary result for the simplest case $\mathcal{X} = \{0, 1\}^d$. Following convention for this setting we also restrict ℓ_t to be such that $\|\ell_t\|_1 \leq 1$. Similar to Section 4.3.2, in the stochastic case we assume $\ell_t \sim \mathcal{D}$ and define $\Delta_i = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell_i]$.

Theorem 4.4. *For the full bandit feedback setting with $\mathcal{X} = \{0, 1\}^d$ and $\|\ell_t\|_1 \leq 1$, FTRL with regularizer $\Psi(x) = \sum_{i=1}^{d} \sqrt{x_i} + \sqrt{1 - x_i}$, learning rate $\eta_t = 1/\sqrt{t}$ and loss estimators $\hat{\ell}_{ti} = \frac{\langle X_t, \ell_t \rangle X_{ti}}{x_{ti}} - \frac{\langle X_t, \ell_t \rangle (1 - X_{ti})}{1 - x_{ti}}$ ensures:*

$$\mathfrak{R}_T \leq \mathcal{O}\left(\sum_{i : \Delta_i \neq 0} \frac{\log(T)}{|\Delta_i|} \right)$$

in the stochastic case and

$$\mathfrak{R}_T \leq \mathcal{O}\left(d\sqrt{T} \right)$$

in the adversarial case. Moreover, both bounds are optimal.

Proof sketch. In this case, the optimization of FTRL decomposes over the coordinates and it is clear that the stated algorithm is equivalent to the following: for each coordinate i, apply the algorithm of Zimmert and Seldin [111] to a two-armed bandit problem where the loss of arm 1 at time t is $\ell_{ti} + \sum_{j \neq i} X_{tj} \ell_{tj}$ and the loss of arm 2 is $\sum_{j \neq i} X_{tj} \ell_{tj}$.[3] In the stochastic case this exactly fits into the stochastically constrained adversarial setting of Zimmert and Seldin [111] with gap $|\Delta_i|$ and, therefore, applying their Theorem 2 and summing up the regret over each coordinate finishes the proof for the stated regret bounds. The optimality of the stochastic bound follows from Combes et al. [42] and the optimality of the adversarial bound follows from Dani et al. [45]. □

For general action sets, however, the problem becomes significantly harder, because all known adversarial algorithms, e.g. Cesa-Bianchi and Lugosi [36], require implicit or explicit exploration of order $1/\sqrt{T}$, which prohibits $\log(T)$ regret in the stochastic case. We leave this as question for future work.

[3] The losses are well defined since they do not depend on X_{ti}.

4.7 Conclusions

We provide the first best-of-both-worlds results for combinatorial bandits, via an Ftrl-based algorithm with a novel hybrid regularizer. Our bounds are worst-case optimal and also optimal in two particular instances of the problem. Empirical evaluations also confirm our theory.

Other than the open problem under bandit feedback mentioned in Section 4.6, another open question is whether our stochastic bound is instance-optimal as in Combes et al. [42], and if not, whether there is a best-of-both-worlds algorithm that is instance-optimal in the stochastic case. One can also ask the same question for the adversarial case, however, next to nothing is known regarding the instance-optimality of the adversarial case, let alone best-of-both-worlds results.

Acknowledgments HL and CYW are supported by NSF Grant #1755781. We thank Yevgeny Seldin for valuable feedback and discussions, and Shinji Ito for pointing us to missing references and unclarities in Section 4.6.

Appendix

Omitted details for the Proof of Theorem 4.1

In this section we provide omitted details for the proof of Theorem 4.1. We first prove Lemma 4.1 Lemma 4.2, then continue on Section 4.5 and prove the upper bounds for C_{sto}, C_{add} and C_{adv}.

4.7.1 Regularization penalty

In order to bound the regularization penalty, we make use of the following standard result for Ftrl.

Lemma 4.3. *The penalty term defined in Eq.* (4.3) *is upper bounded by*

$$\mathfrak{R}_{\text{pen}} \leq \mathbb{E}\left[\frac{-\Psi(x_1)+\Psi(x^*)}{\eta_1}+\sum_{t=2}^{T}(\eta_t^{-1}-\eta_{t-1}^{-1})\left(-\Psi(x_t)+\Psi(x^*)\right)\right].$$

Proof. We proceed as follows:

$$\begin{aligned}
&\sum_{t=1}^{T}\left(-\Phi_t(-\hat{L}_t)+\Phi_t(-\hat{L}_{t-1})-\langle x^*,\hat{\ell}_t\rangle\right)\\
&=\sum_{t=1}^{T}\left(\min_{x\in\text{conv}(\mathcal{X})}\left\{\langle x,\hat{L}_t\rangle+\eta_t^{-1}\Psi(x)\right\}-\left(\langle x_t,\hat{L}_{t-1}\rangle+\eta_t^{-1}\Psi(x_t)\right)\right\}-\sum_{t=1}^{T}\langle x^*,\hat{\ell}_t\rangle\\
&\hspace{20em}\text{(by the definitions of } \Phi_t \text{ and } x_t\text{)}\\
&\leq\langle x^*,\hat{L}_T\rangle+\eta_T^{-1}\Psi(x^*)+\sum_{t=1}^{T-1}\left(\langle x_{t+1},\hat{L}_t\rangle+\eta_t^{-1}\Psi(x_{t+1})\right)-\sum_{t=1}^{T}\left(\langle x_t,\hat{L}_{t-1}\rangle+\eta_t^{-1}\Psi(x_t)\right)-\langle x^*,\hat{L}_T\rangle\\
&=\eta_T^{-1}\Psi(x^*)+\sum_{t=2}^{T}\eta_{t-1}^{-1}\Psi(x_t)-\sum_{t=1}^{T}\eta_t^{-1}\Psi(x_t) \qquad \text{(by telescoping and } \hat{L}_0=\mathbf{0}\text{)}\\
&=\frac{-\Psi(x_1)+\Psi(x^*)}{\eta_1}+\sum_{t=2}^{T}(\eta_t^{-1}-\eta_{t-1}^{-1})\left(-\Psi(x_t)+\Psi(x^*)\right).
\end{aligned}$$

Finally using $\mathbb{E}[\ell_t]=\mathbb{E}[\hat{\ell}_t]$ and plugging in the definition of $\mathfrak{R}_{\text{pen}}$ finish the proof. □

Proof of Lemma 4.1. We directly plug into Lemma 4.3 the learning rate $\eta_t = \eta$ and the regularizer $\Psi(x) = \sum_{i=1}^d -\sqrt{x_i} + \gamma(1-x_i)\log(1-x_i)$. Since $\gamma \leq 1$ and $-(1-x)\log(1-x) \leq \frac{\sqrt{x}}{2}$ for $x \in [0,1]$, we get

$$\begin{aligned} -\Psi(x_t) + \Psi(x^*) &= \sum_{i=1}^d \sqrt{x_{ti}} - \gamma(1-x_{ti})\log(1-x_{ti}) - \sum_{i:x_i^*=1} \sqrt{1} \\ &\leq \sum_{i:x_i^*=0} \frac{3}{2}\sqrt{x_{ti}} - \sum_{i:x_i^*=1} \gamma(1-x_{ti})\log(1-x_{ti}) \\ &\leq \frac{3}{2}\left(\sum_{i:x_i^*=0} \sqrt{x_{ti}} - \sum_{i:x_i^*=1} \gamma(1-x_{ti})\log(1-x_{ti})\right). \end{aligned}$$

It further holds that $\eta_1 = \eta_1^{-1}$ and

$$\eta_t^{-1} - \eta_{t-1}^{-1} = \sqrt{t} - \sqrt{t-1} \leq \frac{1}{2\sqrt{t-1}} \leq \frac{1}{\sqrt{t}} = \eta_t.$$

Inserting everything into Lemma 4.3:

$$\begin{aligned} \mathfrak{R}_{\text{pen}} &\leq \mathbb{E}\left[\frac{-\Psi(x_1) + \Psi(x^*)}{\eta_1} + \sum_{t=2}^T (\eta_t^{-1} - \eta_{t-1}^{-1})\left(-\Psi(x_t) + \Psi(x^*)\right)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^T \eta_t \left(-\Psi(x_t) + \Psi(x^*)\right)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^T \frac{3}{2\sqrt{t}}\left(\sum_{i:x_i^*=0} \sqrt{x_{ti}} - \sum_{i:x_i^*=1} \gamma(1-x_{ti})\log(1-x_{ti})\right)\right] \\ &\leq \sum_{t=1}^T \frac{3}{2\sqrt{t}}\left(\sum_{i:x_i^*=0} \sqrt{\mathbb{E}[x_{ti}]} - \sum_{i:x_i^*=1} \gamma(1-\mathbb{E}[x_{ti}])\log(1-\mathbb{E}[x_{ti}])\right). \end{aligned}$$

where the last step follows from Jensen's inequality and the concavity of functions $\sqrt{x}$ and $-(1-x)\log(1-x)$. □

4.7.2 Stability term

Bounding the stability term defined in Eq. (4.3) requires tools from convex analysis. First we extend the domain of Ψ to $\mathbb{R}^d$ by setting $\Psi(x) = \infty$, $\forall x \in \mathbb{R}^d \setminus [0,1]^d$. Recall the convex conjugate of a convex function f is defined as

$$f^*(\cdot) = \max_{x \in \mathbb{R}^d} \langle x, \cdot \rangle - f(x),$$

and the *Bregman divergence* associated with f is defined as

$$D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

By the above definition, Φ_t can be written as $(\Psi_t + \mathcal{I}_{\text{conv}}(\mathcal{X}))^*$. Note that Ψ_t^* differs from Φ_t because it does not constrain its maximizer to be within $\text{conv}(\mathcal{X})$. The following properties hold (see, e.g., Chapter 7 of [22]):

$$\nabla \Phi_t(\cdot) = \operatorname*{arg\,max}_{x \in \text{conv}(\mathcal{X})} \langle x, \cdot \rangle - \Psi_t(x), \tag{4.5}$$

$$\nabla \Psi_t^*(\cdot) = \operatorname*{arg\,max}_{x \in [0,1]^d} \langle x, \cdot \rangle - \Psi_t(x). \tag{4.6}$$

For Ψ_t and Ψ_t^*, we have

$$\nabla \Psi_t = (\nabla \Psi_t^*)^{-1}, \tag{4.7}$$

$$\nabla^2 \Psi_t(x) = \left(\nabla^2 \Psi_t^*(\nabla \Psi_t(x))\right)^{-1}. \tag{4.8}$$

Furthermore, by Taylor's theorem, for any $x, y \in \mathbb{R}^d$ there exists a $z \in \operatorname{conv}(\{x, y\})$ such that

$$D_{\Psi_t^*}(x, y) = \frac{1}{2}||x - y||^2_{\nabla^2 \Psi_t^*(z)}. \tag{4.9}$$

The explicit expressions for $\nabla \Psi_t, \nabla^2 \Psi_t$ and a convenient upper bound for $(\nabla^2 \Psi_t)^{-1}$ in the domain $(0,1)^d$ are

$$\Psi_t(x) = \eta_t^{-1} \left(\sum_{i=1}^d -\sqrt{x_i} + \gamma(1 - x_i)\log(1 - x_i) \right),$$

$$\nabla \Psi_t(x) = \eta_t^{-1} \left(-\frac{1}{2\sqrt{x_i}} - \gamma \log(1 - x_i) - \gamma \right)_{i=1,\dots,d},$$

$$\nabla^2 \Psi_t(x) = \eta_t^{-1} \operatorname{diag}\left[\left(\frac{1}{4\sqrt{x_i^3}} + \frac{\gamma}{1 - x_i} \right)_{i=1,\dots,d} \right], \tag{4.10}$$

$$\left(\nabla^2 \Psi_t(x)\right)^{-1} \preceq \eta_t \operatorname{diag}\left[\left(\min\left\{ 4\sqrt{x_i^3}, \gamma^{-1}(1 - x_i) \right\} \right)_{i=1,\dots,d} \right], \tag{4.11}$$

where $(v_i)_{i=1,\dots,d}$ denotes $(v_1, \dots, v_d)$, $\operatorname{diag}[(v_i)_{i=1,\dots,d}]$ denotes a diagonal matrix with $(v_i)_{i=1,\dots,d}$ on the diagonal, and $A \preceq B$ for two matrices A and B means $B - A$ is positive semidefinite. Note $\nabla \Psi_t$ is a bijection from $(0,1)^d$ to $\mathbb{R}^d$. Therefore $\nabla \Psi_t^*(L) \in (0,1)^d$ for any $L \in \mathbb{R}^d$, and all x_t's we consider here are in the domain $(0,1)^d$.

The following Lemma will be useful to show that the stability term can be bounded independently of the action set $\mathcal{X}$.

Lemma 4.4. *For any L, let $\tilde{L} = \nabla \Psi_t(\nabla \Phi_t(L))$. Then it holds for any $\ell \in \mathbb{R}^d$:*

$$D_{\Phi_t}(L + \ell, L) \leq D_{\Psi_t^*}(\tilde{L} + \ell, \tilde{L}).$$

Proof. First we state two equalities that follow from the previously stated properties.

$$\nabla \Psi_t^*(\tilde{L}) = \nabla \Psi_t^*(\nabla \Psi_t(\nabla \Phi_t(L))) \overset{Eq.~(4.7)}{=} \nabla \Phi_t(L), \tag{4.12}$$

$$\begin{aligned} \Psi_t^*(\tilde{L}) &\overset{Eq.~(4.6)}{=} \langle \nabla \Psi_t^*(\tilde{L}), \tilde{L} \rangle - \Psi_t(\nabla \Psi_t^*(\tilde{L})) \\ &= \langle \nabla \Phi_t(L), \tilde{L} \rangle - \Psi_t(\nabla \Phi_t(L)) \\ &\overset{Eq.~(4.5)}{=} \Phi_t(L) + \langle \nabla \Phi_t(L), \tilde{L} - L \rangle. \end{aligned} \tag{4.13}$$

We then proceed as follows:

$$
\begin{aligned}
&D_{\Psi_t^*}(\tilde{L}+\ell,\tilde{L})\\
&=\Psi_t^*(\tilde{L}+\ell)-\Psi_t^*(\tilde{L})-\left\langle\nabla\Psi_t^*(\tilde{L}),\ell\right\rangle && \text{(definition of Bregman divergence)}\\
&=\Psi_t^*(\tilde{L}+\ell)-\Phi_t(L)-\left\langle\nabla\Phi_t(L),\tilde{L}-L\right\rangle-\langle\nabla\Phi_t(L),\ell\rangle && \text{(by Eqs. (4.12) and (4.13))}\\
&=\Psi_t^*(\tilde{L}+\ell)-\Phi_t(L)-\left\langle\nabla\Phi_t(L),\tilde{L}-L+\ell\right\rangle\\
&\geq\langle\nabla\Phi_t(L+\ell),\tilde{L}+\ell\rangle-\Psi_t(\nabla\Phi_t(L+\ell))-\Phi_t(L)-\left\langle\nabla\Phi_t(L),\tilde{L}-L+\ell\right\rangle\\
& && \text{(Ψ_t^* is defined as the maximum)}\\
&=\langle\nabla\Phi_t(L+\ell),L+\ell\rangle-\Psi_t(\nabla\Phi_t(L+\ell))+\left\langle\nabla\Phi_t(L+\ell),\tilde{L}-L\right\rangle-\Phi_t(L)-\left\langle\nabla\Phi_t(L),\tilde{L}-L+\ell\right\rangle\\
&=\Phi_t(L+\ell)+\left\langle\nabla\Phi_t(L+\ell),\tilde{L}-L\right\rangle-\Phi_t(L)-\left\langle\nabla\Phi_t(L),\tilde{L}-L+\ell\right\rangle\\
& && \text{(by the definition of Φ_t and Eq. (4.5))}\\
&=D_{\Phi_t}(L+\ell,L)+\left\langle\nabla\Phi_t(L+\ell)-\nabla\Phi_t(L),\tilde{L}-L\right\rangle\\
&=D_{\Phi_t}(L+\ell,L)+\langle\nabla\Phi_t(L+\ell)-\nabla\Phi_t(L),\nabla\Psi_t(\nabla\Phi_t(L))-L\rangle\\
&\geq D_{\Phi_t}(L+\ell,L).
\end{aligned}
$$

The last step is by the first-order optimality condition: for the maximizer $\nabla\Phi_t(L):=\arg\max_{x\in\mathrm{conv}(\mathcal{X})}\langle x,L\rangle-\Psi_t(x)$ it must hold that $\langle y-\nabla\Phi_t(L),L-\nabla\Psi_t(\nabla\Phi_t(L))\rangle\leq 0$ for any $y\in\mathrm{conv}(\mathcal{X})$. □

The next Lemma will be useful to bound the eigenvalues of the Hessian of Ψ_t^*.

Lemma 4.5. *If $\eta_t\leq\min\{\frac{\sqrt{2}-1}{2},\frac{\gamma\log(2)}{4}\}$, then for any $x\in(0,1)^d$ and $\hat{\ell}$ such that $-1\leq\hat{\ell}_i\leq\frac{2}{x_i}$ for all i, we have*

$$2x_i-1\leq\nabla\Psi_t^*(\nabla\Psi_t(x)-\hat{\ell})_i\leq 2x_i.$$

Proof. The functions $\nabla\Psi_t$ and $\nabla\Psi_t^*$ are symmetric and independent in each dimension. Therefore it is sufficient to consider $d=1$ and drop the index i.

For the upper bound we can assume $x<\frac{1}{2}$; otherwise the statement is trivial since the range of $\nabla\Psi_t^*$ is $(0,1)^d$. Now assume the opposite holds: $\nabla\Psi_t^*(\nabla\Psi_t(x)-\hat{\ell})>2x$, then we have

$$
\begin{aligned}
\hat{\ell}&=\nabla\Psi_t(x)-\nabla\Psi_t(x)+\hat{\ell}=\nabla\Psi_t(x)-\nabla\Psi_t(\nabla\Psi_t^*(\nabla\Psi_t(x)-\hat{\ell}))\\
&<\nabla\Psi_t(x)-\nabla\Psi_t(2x) && \text{($\nabla\Psi_t(x)$ is strictly increasing in $(0,1)$)}\\
&=\eta_t^{-1}\left(-\frac{1}{2\sqrt{x}}-\gamma\log(1-x)+\frac{1}{2\sqrt{2x}}+\gamma\log(1-2x)\right)\\
&<-\eta_t^{-1}\left(\frac{\sqrt{2}-1}{2\sqrt{2}}\right)\frac{1}{\sqrt{x}}\\
&<-\eta_t^{-1}\left(\frac{\sqrt{2}-1}{2}\right). && (x\leq\tfrac{1}{2})
\end{aligned}
$$

The last line is a contradiction to the conditions $\eta_t\leq\frac{\sqrt{2}-1}{2}$ and $\hat{\ell}\geq-1$.

For the lower bound we can assume $x>\frac{1}{2}$, otherwise the statement is again trivial. Assume the opposite holds: $\nabla\Psi_t^*(\nabla\Psi_t(x) + \hat{\ell}) < 2x - 1$, then we have

$$\begin{aligned}
\hat{\ell} &= \nabla\Psi_t(x) - \nabla\Psi_t(\nabla\Psi_t^*(\nabla\Psi_t(x) - \hat{\ell}))\\
&> \nabla\Psi_t(x) - \nabla\Psi_t(2x-1) && (\nabla\Psi_t(x) \text{ is strictly increasing in } (0,1))\\
&= \eta_t^{-1}\left(-\frac{1}{2\sqrt{x}} - \gamma\log(1-x) + \frac{1}{2\sqrt{2x-1}} + \gamma\log(2-2x)\right)\\
&> \eta_t^{-1}\log(2) > \eta_t^{-1}\frac{\gamma\log(2)}{4}\frac{2}{x}. && (\gamma \le 1 \text{ and } x > 1/2)
\end{aligned}$$

which again leads to a contradiction to the conditions $\eta_t \le \frac{\gamma\log(2)}{4}$ and $\hat{\ell} \le \frac{2}{x}$. This finishes the proof. □

Finally we are ready to prove Lemma 4.2.

Proof of Lemma 4.2. Let $\tilde{x}_t = \nabla\Psi^*(\nabla\Psi(x_t) - \hat{\ell}_t)$. Define $\mathcal{A}_t = \bigotimes_{i=1}^d [x_{ti}, \tilde{x}_{ti}]$. For any $t_0 \ge 0$, we bound the stability term by

$$\begin{aligned}
\mathfrak{R}_{\text{stab}} &= \mathbb{E}\left[\sum_{t=1}^T \langle X_t, \ell_t\rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]\\
&\overset{(1)}{\le} \mathbb{E}\left[\sum_{t=t_0}^T \langle X_t, \ell_t\rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right] + 2t_0 m\\
&\overset{(2)}{=} \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[\langle x_t, \hat{\ell}_t\rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]\right] + 2t_0 m\\
&= \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[D_{\Phi_t}(-\hat{L}_t, -\hat{L}_{t-1})\right]\right] + t_0 m\\
&\overset{(3)}{\le} \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[D_{\Psi_t^*}(\nabla\Psi_t(x_t) - \hat{\ell}_t, \nabla\Psi_t(x_t))\right]\right] + 2t_0 m\\
&= \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[D_{\Psi_t^*}(\nabla\Psi_t(\tilde{x}_t), \nabla\Psi_t(x_t))\right]\right] + 2t_0 m\\
&\overset{(4)}{=} \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[\frac{1}{2}||\hat{\ell}_t||^2_{\nabla^2\Psi_t^*(z_t)}\right]\right] + 2t_0 m\\
&\overset{(5)}{\le} \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[\max_{x\in\mathcal{A}_t}\frac{1}{2}||\hat{\ell}_t||^2_{\nabla^2\Psi_t(x)^{-1}}\right]\right] + 2t_0 m\\
&= \mathbb{E}\left[\sum_{t=t_0}^T \mathbb{E}_t\left[\max_{x\in\mathcal{A}_t}\frac{\eta_t}{2}||\hat{\ell}_t||^2_{\nabla^2\Psi(x)^{-1}}\right]\right] + 2t_0 m. \qquad (4.14)
\end{aligned}$$

(1) The difference of potentials for each step is bounded by $\Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}) \le \langle\nabla\Phi_t(-\hat{L}_t), -\hat{\ell}\rangle \le ||\nabla\Phi_t(-\hat{L}_t)||_1 \le m$, and the loss $\langle X_t, \ell_t\rangle$ is bounded by $m = \max_{x\in\mathcal{X}} ||x||_1$. (2) By the tower rule of conditional expectation, the unbiaseness of $\hat{\ell}$ and the sampling assumption, it holds that

$$\mathbb{E}\left[\langle X_t, \ell_t\rangle\right] = \mathbb{E}\left[\mathbb{E}_t\left[\langle X_t, \ell_t\rangle\right]\right] = \mathbb{E}\left[\mathbb{E}_t\left[\langle x_t, \ell_t\rangle\right]\right] = \mathbb{E}\left[\mathbb{E}_t\left[\langle x_t, \hat{\ell}_t\rangle\right]\right].$$

(3) Applyication of Lemma 4.4.

(4) Eq. (4.9) ensures that some $z_t \in \operatorname{conv}(\{\nabla\Psi_t(x), \nabla\Psi_t(\tilde{x})\})$ exists that satisfies the equality.

(5) By property (4.8) and the coordinate-wise monotonicity of $\nabla\Psi_t^*$ so that $\nabla\Psi_t^*(z_t) \subset \bigotimes_{i=1}^d [x_{ti}, \tilde{x}_{ti}] = \mathcal{A}_t$.

We choose $t_0 = 58\gamma^{-2}$ such that $\eta_t \leq \min\{\frac{\sqrt{2}-1}{2}, \frac{\gamma\log(2)}{4}\}$ for any $t \geq t_0$. By the construction of $\hat{\ell}_t$ we clearly have $-1 \leq \hat{\ell}_{ti} \leq \frac{2}{x_{ti}}$. We can then apply Lemma 4.5 to conclude that $\tilde{x}_{ti} \in [2x_{ti}-1, 2x_{ti}]$. Therefore, with the form of Hessian Eq. (4.11) we have:

$$\forall x \in \mathcal{A}_t : \qquad \nabla^2\Psi(x)^{-1} \leq \operatorname{diag}\left[\left(\min\left\{4\sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti})\right\}\right)_{i=1,\dots,d}\right],$$

and therefore,

$$\begin{aligned}\sum_{t=t_0}^T \mathbb{E}_t\left[\max_{x\in\mathcal{A}_t}\frac{\eta_t}{2}||\hat{\ell}_t||^2_{\nabla^2\Psi(x)^{-1}}\right] &\leq \sum_{t=t_0}^T \mathbb{E}_t\left[\frac{\eta_t}{2}\sum_{i=1}^d(\hat{\ell}_{ti})^2\min\{4\sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti})\}\right]\\ &\overset{(1)}{\leq} \sum_{t=t_0}^T \frac{\eta_t}{2}\sum_{i=1}^d \frac{4}{x_{ti}}\min\{4\sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti})\}\\ &\overset{(2)}{\leq} \sum_{t=t_0}^T 16\sqrt{2}\eta_t\sum_{i=1}^d\min\{\sqrt{x_{ti}}, \gamma^{-1}(1-x_{ti})\}\\ &\leq \sum_{t=1}^T\frac{16\sqrt{2}}{\sqrt{t}}\left(\sum_{i:x_i^*=0}\sqrt{x_{ti}} + \sum_{i:x_i^*=1}\gamma^{-1}(1-x_{ti})\right). \end{aligned} \tag{4.15}$$

(1) Conditioned on $\mathcal{F}_{t-1}$, only $(\hat{\ell}_{ti})^2$ is random and its expectation is

$$\mathbb{E}_t\left[(\hat{\ell}_{ti})^2\right] = x_{ti}\left(\frac{\ell_{ti}+1}{x_{ti}}-1\right)^2 + 1 - x_{ti} = \frac{4}{x_{ti}}\frac{(\ell_{ti}+1)^2 - 2(\ell_{ti}+1)x_{ti} + x_{ti}}{4} \leq \frac{4}{x_{ti}}\frac{(4-4x_{ti}+x_{ti})}{4} \leq \frac{4}{x_{ti}}.$$

(2) Note that it always holds

$$\frac{4}{x_{ti}}\min\left\{4\sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti})\right\} \leq \frac{16}{x_{ti}}\sqrt{(2x_{ti})^3} = 32\sqrt{2x_{ti}}.$$

So it suffices to prove $\frac{4}{x_{ti}}\min\{4\sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti})\} \leq 32\sqrt{2}\gamma^{-1}(1-x_{ti})$. We consider two cases: (A) If $4\sqrt{(2x_{ti})^3} \leq 2\gamma^{-1}(1-x_{ti})$, then we need to prove $\sqrt{x_{ti}} \leq \gamma^{-1}(1-x_{ti})$. This is true since either $x_{ti} \geq 1/\sqrt{32}$ and thus $\sqrt{x_{ti}} \leq 2\sqrt{(2x_{ti})^3} \leq \gamma^{-1}(1-x_{ti})$, or $x_{ti} < 1/\sqrt{32}$ in which case $\sqrt{x_{ti}} \leq 1 - x_{ti} \leq \gamma^{-1}(1-x_{ti})$. (B) If $4\sqrt{(2x_{ti})^3} \geq 2\gamma^{-1}(1-x_{ti})$, then x_{ti} must be larger than 1/4. In this case we bound $\frac{1}{x_{ti}}$ by 4 and the desired inequality follows.

The proof is concluded by inserting Eq. (4.15) into Eq. (4.14) and using Jensen's inequality to move the expectation into the concave functions. □

4.7.3 General upper bounds for C_{sto}, C_{add} and C_{adv}

We now finish the proof of Theorem 4.1 on the upper bounds of the three constants.

Bounding $\mathbf{C_{adv}}$:

$$\begin{aligned}
C_{adv} &= \max_{x\in\operatorname{conv}(\mathcal{X})} \sum_{i:x_i^*=0} \sqrt{x_i} + \sum_{i:x_i^*=1} (\gamma^{-1} - \gamma\log(1-x_i))(1-x_i) \\
&\le \max_{x\in\operatorname{conv}(\mathcal{X})} \sum_{i:x_i^*=0} \sqrt{x_i} + \sum_{i:x_i^*=1} \gamma\sqrt{1-x_i} + \sum_{i:x_i^*=1} \gamma^{-1}(1-x_i) \\
&\hphantom{\le} \qquad\qquad\qquad\qquad\qquad\qquad (-y\log y \le \sqrt{y} \text{ for } y\in[0,1]) \\
&\le \max_{x\in\operatorname{conv}(\mathcal{X})} \sqrt{\left(\sum_{i:x_i^*=0} 1\right)\left(\sum_{i:x_i^*=0} x_i\right)} + \gamma\sqrt{\left(\sum_{i:x_i^*=1} 1\right)\left(\sum_{i:x_i^*=1} (1-x_i)\right)} + \gamma^{-1}m \\
&\hphantom{\le} \qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy-Schwarz)} \\
&\le \sqrt{dm} + \gamma m + \gamma^{-1}m \\
&\le \mathcal{O}\left(\gamma^{-1}\sqrt{md}\right).
\end{aligned}$$

Bounding $\mathbf{C_{sto}}$: C_{sto} is defined as $\max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} f(\overline{\alpha}) - r(\alpha)$. First we bound $f(\overline{\alpha})$:

$$f(\overline{\alpha}) = \sum_{i:x_i^*=0} \sqrt{\overline{\alpha}_i} = \sum_{i:x_i^*=0} \sqrt{\sum_{x\in\mathcal{X}} \alpha_x x_i} = \sum_{i:x_i^*=0} \sqrt{\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x x_i} \le \sqrt{d\sum_{i:x_i^*=0}\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x x_i} \le \sqrt{dm\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x}.$$

On the other hand,

$$r(\alpha) = \sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x\Delta_x \ge \Delta_{\min}\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x.$$

Combining them we get

$$\begin{aligned}
C_{sto} &\le \max_{\alpha\in[0,\infty)} \sqrt{dm\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x} - \Delta_{\min}\sum_{x\in\mathcal{X}\backslash\{x^*\}} \alpha_x \\
&\le \max_{A\ge 0} \sqrt{dmA} - \Delta_{\min}A \\
&\le \max_{A\ge 0} \Delta_{\min}A + \frac{dm}{4\Delta_{\min}} - \Delta_{\min}A \qquad \text{(AM-GM inequality)} \\
&= \frac{dm}{4\Delta_{\min}}.
\end{aligned}$$

Bounding $\mathbf{C_{add}}$: Recall C_{add} is defined as $\sum_{t=1}^{\infty}\max_{\alpha\in\Delta(\mathcal{X})}\left(\frac{100}{\sqrt{t}}g(\overline{\alpha}) - r(\alpha)\right)$. We will give a upper bound for $g(\overline{\alpha})$ and lower bound for $r(\alpha)$ below.

We first prove the following property: for any $y\in\mathbb{R}_+^N$, $\sum_{i=1}^N y_i\log\frac{1}{y_i} \le \|y\|_1\log\frac{N}{\|y\|_1}$. Indeed, by the concavity of the log function and Jensen's inequality,

$$\sum_{i=1}^N \frac{y_i}{\|y\|_1}\log\frac{1}{y_i} \le \log\left(\sum_{i=1}^N \frac{y_i}{\|y\|_1}\frac{1}{y_i}\right) = \log\frac{N}{\|y\|_1}.$$

Therefore, for any $\alpha\in\Delta(\mathcal{X})$ we have

$$\begin{aligned}
g(\overline{\alpha}) &= \sum_{i:x_i^*=1}\left(\gamma^{-1} + \gamma\log\left(\frac{1}{1-\bar{\alpha}_i}\right)\right)(1-\bar{\alpha}_i) \\
&\le \left(\sum_{i:x_i^*=1}(1-\overline{\alpha}_i)\right)\left(\gamma^{-1} + \gamma\log\frac{m}{\sum_{i:x_i^*=1}(1-\overline{\alpha}_i)}\right). \qquad \text{(using the above property)}
\end{aligned}$$

Then consider the following two facts. First, the function of y defined by $y(\gamma^{-1} + \gamma \log \frac{m}{y})$ is increasing in $y \in [0, m]$. This can be verified by

$$\frac{\partial}{\partial y}\left(y\left(\gamma^{-1} + \gamma \log \frac{m}{y}\right)\right) = \gamma^{-1} + \gamma \log m - \gamma \log y - \gamma \geq 0. \qquad (\gamma \leq 1)$$

Second, we have $\sum_{i:x_i^*=1}(1-\overline{\alpha}_i) = \sum_{i:x_i^*=1}\sum_{\alpha\in\mathcal{X}}\alpha_x(1-x_i) = \sum_{i:x_i^*=1}\sum_{\alpha\in\mathcal{X}\backslash\{x^*\}}\alpha_x(1-x_i) \leq \|x^*\|_1\left(\sum_{\alpha\in\mathcal{X}\backslash\{x^*\}}\alpha_x\right) \leq m\left(\sum_{\alpha\in\mathcal{X}\backslash\{x^*\}}\alpha_x\right)$. Combining these two facts with the above bound for $g(\overline{\alpha})$, we get

$$g(\overline{\alpha}) \leq m\left(\sum_{\alpha\in\mathcal{X}\backslash\{x^*\}}\alpha_x\right)\left(\gamma^{-1} + \gamma \log \frac{1}{\sum_{\alpha\in\mathcal{X}\backslash\{x^*\}}\alpha_x}\right).$$

On the other hand, we have the lower bound for $r(\alpha)$:

$$r(\alpha) = \sum_{x\in\mathcal{X}\backslash\{x^*\}}\alpha_x\Delta_x \geq \Delta_{\min}\sum_{x\in\mathcal{X}\backslash\{x^*\}}\alpha_x.$$

Therefore,

$$\begin{aligned} C_{add} &= \sum_{t=1}^{\infty}\max_{\alpha\in\Delta(|\mathcal{X}|)}\left(\frac{100}{\sqrt{t}}g(\overline{\alpha}) - r(\alpha)\right) \\ &\leq \sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{100}{\sqrt{t}}mA\left(\gamma^{-1} + \gamma\log\frac{1}{A}\right) - \Delta_{\min}A\right). \end{aligned}$$

We further bound it by the sum of the following two summations:

- $\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{100}{\sqrt{t}}mA\gamma^{-1} - \frac{1}{2}\Delta_{\min}A\right)$
- $\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{100}{\sqrt{t}}mA\gamma\log\frac{1}{A} - \frac{1}{2}\Delta_{\min}A\right)$

Lemma 4.6 Lemma 4.7 below respectively bound these two as $\mathcal{O}\left(\frac{m^2\gamma^{-2}}{\Delta_{\min}}\right)$ and $\mathcal{O}\left(\frac{m^2\gamma^2}{\Delta_{\min}}\right)$, which finishes the proof.

Lemma 4.6. *For any $C > 0$ and $\Delta > 0$, we have* $\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{C}{\sqrt{t}}A - \Delta A\right) \leq \mathcal{O}\left(\frac{C^2}{\Delta}\right)$.

Proof. Let T_0 be the largest t such that $\frac{C}{\sqrt{t}} - \Delta > 0$, then

$$\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{C}{\sqrt{t}}A - \Delta A\right) \leq \sum_{t=1}^{T_0}\frac{C}{\sqrt{t}} \leq 2C\sqrt{T_0} = \mathcal{O}\left(\frac{C^2}{\Delta}\right).$$

□

Lemma 4.7. *For any $C > 0$ and $\Delta > 0$, we have* $\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{C}{\sqrt{t}}A\log\frac{1}{A} - \Delta A\right) \leq \frac{C^2}{\Delta}$.

Proof. We first solve the inner optimization with respect to a specific t. Taking the derivative with respect to A, and setting it to zero:

$$\frac{C}{\sqrt{t}}\log\frac{1}{A^*} - \frac{C}{\sqrt{t}} - \Delta = 0, \qquad (4.16)$$

we get the solution

$$A^* = \exp\left(-1 - \frac{\sqrt{t}\Delta}{C}\right).$$

And thus,

$$\begin{aligned}\max_{A\in[0,1]}\left(\frac{C}{\sqrt{t}}A\log\frac{1}{A} - \Delta A\right) &= \frac{C}{\sqrt{t}}A^*\log\frac{1}{A^*} - \Delta A^* \overset{Eq.\ (4.16)}{=} A^*\left(\frac{C}{\sqrt{t}} + \Delta\right) - \Delta A^* \\ &= \frac{C}{\sqrt{t}}\exp\left(-1 - \frac{\sqrt{t}\Delta}{C}\right)\end{aligned}$$

Finally we have

$$\begin{aligned}\sum_{t=1}^{\infty}\max_{A\in[0,1]}\left(\frac{C}{\sqrt{t}}A\log\frac{1}{A} - \Delta A\right) &\leq \sum_{t=1}^{\infty}\frac{C}{\sqrt{t}}\exp\left(-1 - \frac{\sqrt{t}\Delta}{C}\right) \leq \int_{t=0}^{\infty}\frac{C}{\sqrt{t}}\exp\left(-1 - \frac{\sqrt{t}\Delta}{C}\right)dt \\ &= \frac{C^2}{\Delta}\int_{\tau=0}^{\infty}\frac{1}{\sqrt{\tau}}\exp(-1 - \sqrt{\tau})d\tau \leq \frac{C^2}{\Delta}.\end{aligned}$$

□

Omitted Details for Section 4.3.2 Section 4.3.3

In this section we provide omitted details for the two special cases: full combinatorial set and m-set.

4.7.4 Optimality of the stochastic bound when $\mathcal{X} = \{0,1\}^d$

As mentioned in the proof of Theorem 4.2, we provide here for completeness a proof showing that when $d = 1$ and $\Delta > 0$, the regret is at least $\Omega(\frac{\log T}{\Delta})$.

Assume that there exists an algorithm that is at least as good as ours asymptotically, which implies $\lim_{T\to\infty}\frac{\log(\mathfrak{R}_T)}{\log(T)} \leq \lim_{T\to\infty}\frac{\log(\mathcal{O}(\log(T))}{\log(T)} = 0$ for any problem. For some $\Delta > 0$ we consider two problems: $\mathbb{E}[\ell] = \Delta$ and $\mathbb{E}[\ell] = -\Delta$. For simplicity we assume that the losses are drawn from i.i.d. Gaussian with variance $\sigma^2 = 1$, but the proof can be easily transferred to Bernoulli noise as well. For the problem with positive loss, we denote the regret as $\mathfrak{R}_T^+$ and the probability space induced by an algorithm by $\mathbb{P}^+$. Equivalently we define $\mathfrak{R}_T^-$ and $\mathbb{P}^-$. The relative entropy between $\mathbb{P}^+$ and $\mathbb{P}^-$ is

$$\mathrm{KL}(\mathbb{P}^+,\mathbb{P}^-) = \sum_{t=1}^{T}\mathbb{P}^+(X_t = 1)(2\Delta)^2 = 4\mathfrak{R}_T^+\Delta.$$

Also we have by the definition of regret:

$$\mathbb{P}^+\left(\sum_{t=1}^{T}X_t \geq \frac{T}{2}\right) + \mathbb{P}^-\left(\sum_{t=1}^{T}X_t < \frac{T}{2}\right) \leq \frac{2(\mathfrak{R}_T^+ + \mathfrak{R}_T^-)}{\Delta T}.$$

Using the high probability Pinsker inequality (included after the proof for completeness), we get

$$\frac{2(\mathfrak{R}_T^+ + \mathfrak{R}_T^-)}{\Delta T} \geq \frac{1}{2}\exp\left(-4\mathfrak{R}_T^+\Delta\right).$$

Rearranging gives

$$\frac{\mathfrak{R}_T^+}{\log(T)} = \frac{1}{4\Delta} - \frac{1}{4\Delta}\frac{\log(4(\mathfrak{R}_T^+ + \mathfrak{R}_T^-))}{\log(T)} + \frac{\log(\Delta)}{4\Delta\log(T)}.$$

Taking the limit on both sides shows $\lim_{T\to\infty}\frac{\mathfrak{R}_T^+}{\log(T)} = \Omega(\frac{1}{\Delta})$, which finishes the proof.

Lemma 4.8 (High Probability Pinsker, e.g. [29])**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space (Ω, F) and let $A \in F$ be an arbitrary event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P},\mathbb{Q})),$$

where A^c is the complement of A and $\mathrm{KL}(\mathbb{P},\mathbb{Q})$ the relative entropy.

4.7.5 Sampling rule for m-set

In this section $\mathcal{X}$ represents the m-set. We first define the following auxiliary vectors for $0 \leq i \leq m$, $0 \leq j \leq d-m$.

$$\beta_{i,j} = \left(\underbrace{1,\dots,1}_{i}, \frac{m-i}{d-i-j}, \dots, \frac{m-i}{d-i-j}, \underbrace{0,\dots,0}_{j}\right) \in \mathrm{conv}(\mathcal{X}).$$

It is trivial to sample with mean $\beta_{i,j}$ with the sampling rule:

$$P_{i,j} = \mathrm{Uniform}\left(\{x \in \mathcal{X} \mid x_{1,\dots,i} = \mathbf{1} \wedge x_{d-j+1,\dots,d} = \mathbf{0}\}\right).$$

This requires uniform sampling of a $(m-i)$-sized subset of $(d-i-j)$ elements, which can be done in $\mathcal{O}(d)$ time.

Now for a given $x_t \in \mathrm{conv}(\mathcal{X})$, one sampling rule P such that $\mathbb{E}_{X\sim P}[X] = x_t$ is the following: First we sort the entries of x_t so that x is the sorted version with $x_1 \geq \dots \geq x_d$. This takes $\mathcal{O}(d\log(d))$ time. Next we decompose $x = \sum_{s=0}^d p_{x,s}\beta_{i_s,j_s}$ such that $p_{x,s} \in [0,1]$, $\sum_{s=0}^d p_{x,s} = 1$, $(i_0, j_0) = (0,0)$ and $(i_{s+1}, j_{s+1}) - (i_s, j_s) \in \{(1,0),(0,1)\}$. In other words, either i or j increases by one from s to $s+1$. This decomposition is unique and can be computed in a greedy manner in time $\mathcal{O}(d)$. Finally the full sampling scheme is $\sum_{s=0}^d p_{x,s}P_{i_s,j_s}$ (in terms of permuted coordinates). The runtime is dominated by the sorting and hence is $\mathcal{O}(d\log(d))$ overall.

4.7.6 Complete proof for Theorem 4.3

Bounding $\mathbf{C_{adv}}$:

$$C_{adv} = \max_{x\in\mathrm{conv}(\mathcal{X})}(f(x)+g(x)) = \max_{x\in\mathrm{conv}(\mathcal{X})}\sum_{i:x_i^*=0}\sqrt{x_i} + \sum_{i:x_i^*=1}(\gamma^{-1} - \gamma\log(1-x_i))(1-x_i).$$

The optimization problem is concave in x and symmetric for all i with the same value of x_i^*. This implies that the arg max solution must take the following form:

$$\left(\operatorname*{arg\,max}_{x\in\mathrm{conv}(\mathcal{X})} f(x)+g(x)\right)_i = \begin{cases}\lambda & \text{if } x_i^*=0\\ 1-\frac{d-m}{m}\lambda & \text{if } x_i^*=1\end{cases}$$

for some $\lambda \in [0, \min\{1, \frac{m}{d-m}\}]$.

Therefore,

$$\begin{aligned}C_{adv} &= \max_{\lambda\in[0,\min(1,\frac{m}{d-m})]}(d-m)\sqrt{\lambda}+m\left(\gamma^{-1}-\gamma\log\left(\frac{d-m}{m}\lambda\right)\right)\frac{d-m}{m}\lambda\\ &= \max_{\lambda\in[0,\min(1,\frac{m}{d-m})]}(d-m)\left(\sqrt{\lambda}+\left(\gamma^{-1}-\gamma\log\left(\frac{d-m}{m}\lambda\right)\right)\lambda\right). \qquad (4.17)\end{aligned}$$

Since $\frac{d-m}{m}\lambda \leq 1$ and $\gamma \leq 1$, the derivative is always positive:

$$\begin{aligned}&\frac{\partial}{\partial\lambda}\left(\sqrt{\lambda}+\left(\gamma^{-1}-\gamma\log\left(\frac{d-m}{m}\lambda\right)\right)\lambda\right)\\ &=\left(\frac{1}{2\sqrt{\lambda}}+\gamma^{-1}-\gamma\log\left(\frac{d-m}{m}\lambda\right)-\gamma\right)\geq\frac{1}{2\sqrt{\lambda}}>0.\end{aligned}$$

Therefore we can simply plug in the upper border of λ in Eq. (4.17):

Case $m \leq d/2$ (for which $\gamma = 1$ and the optimal λ is $m/(d-m)$):

$$C_{adv} = (d-m)\left(\sqrt{\frac{m}{d-m}}+\frac{m}{d-m}\right)\leq 2\sqrt{(d-m)m}=\mathcal{O}\left(\sqrt{md}\right).$$

Case $m > d/2$ (for which $\gamma = \min\left\{1, 1/\sqrt{\log\left(\frac{d}{d-m}\right)}\right\}$ and the optimal λ is 1):

Note that $\gamma \leq \frac{1}{\sqrt{\log\left(\frac{d}{d-m}\right)}}$ and thus $\gamma^{-1} = \max\left\{1, \sqrt{\log\left(\frac{d}{d-m}\right)}\right\} \leq \frac{\sqrt{\log\left(\frac{d}{d-m}\right)}}{\sqrt{\log(2)}}$ and $-\gamma \leq -\frac{\sqrt{\log(2)}}{\sqrt{\log\left(\frac{d}{d-m}\right)}}$. Therefore

$$\begin{aligned}C_{adv} &\leq (d-m)\left(1+\frac{1}{\sqrt{\log(2)}}\sqrt{\log\left(\frac{d}{d-m}\right)}+\frac{\sqrt{\log(2)}}{\sqrt{\log\left(\frac{d}{d-m}\right)}}\log\left(\frac{m}{d-m}\right)\right)\\ &\leq (d-m)\left(1+\left(\frac{1}{\sqrt{\log(2)}}+\sqrt{\log(2)}\right)\sqrt{\log\left(\frac{d}{d-m}\right)}\right)=\mathcal{O}\left((d-m)\sqrt{\log\left(\frac{d}{d-m}\right)}\right).\end{aligned}$$

Bounding $\mathbf{C_{sto}}$: With our definitions of Δ_i, for any $x \in \mathcal{X}$, we have

$$\Delta_x = \mathbb{E}\left[\sum_i (x_i - x_i^*)\ell_{ti}\right] = \mathbb{E}\left[\sum_i (x_i - x_i^*)(\ell_{ti}-\ell_{tm})\right] = \sum_{i:x_i^*=1}(1-x_i)|\Delta_i| + \sum_{i:x_i^*=0}x_i\Delta_i \geq \sum_{i:x_i^*=0}x_i\Delta_i, \qquad (4.18)$$

and thus for any $\alpha \in [0,\infty)^{|\mathcal{X}|}$

$$r(\alpha) = \sum_{x\in\mathcal{X}\setminus\{x^*\}}\alpha_x\Delta_x \geq \sum_{x\in\mathcal{X}\setminus\{x^*\}}\sum_{i:x_i^*=0}\alpha_x x_i\Delta_i = \sum_{i:x_i^*=0}\overline{\alpha}_i\Delta_i. \qquad (4.19)$$

Therefore,

$$\begin{aligned}C_{sto} &= \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}}\sum_{i:x_i^*=0}\sqrt{\overline{\alpha}_i}-r(\alpha)\\ &\leq \max_{\overline{\alpha}\in[0,\infty)^d}\sum_{i:x_i^*=0}\left(\sqrt{\overline{\alpha}_i}-\overline{\alpha}_i\Delta_i\right)\\ &\overset{\text{AM-GM}}{\leq} \max_{\overline{\alpha}\in[0,\infty)^d}\sum_{i:x_i^*=0}\left(\overline{\alpha}_i\Delta_i+\frac{1}{4\Delta_i}-\overline{\alpha}_i\Delta_i\right)=\sum_{i:x_i^*=0}\frac{1}{4\Delta_i}.\end{aligned}$$

Bounding $\mathbf{C_{add}}$: Similar to the "Bounding C_{add}" part in the proof of Theorem 4.1 (earlier in Section 4.7), we can bound for any $\alpha \in \Delta(\mathcal{X})$:

$$\begin{aligned}
g(\overline{\alpha}) &= \sum_{i:x_i^*=1} \left(\gamma^{-1} + \gamma \log\left(\frac{1}{1-\overline{\alpha}_i}\right)\right)(1-\overline{\alpha}_i) \\
&\leq \left(\gamma^{-1} + \gamma \log\left(\frac{m}{\sum_{i:x_i^*=1}(1-\overline{\alpha}_i)}\right)\right) \sum_{i:x_i^*=1}(1-\overline{\alpha}_i) \qquad \text{(by the concavity of } g\text{)} \\
&= \left(\gamma^{-1} + \gamma \log\left(\frac{m}{\sum_{i:x_i^*=0}\overline{\alpha}_i}\right)\right) \sum_{i:x_i^*=0}\overline{\alpha}_i \\
&\leq \sum_{i:x_i^*=0} \left(\gamma^{-1} + \gamma \log\left(\frac{m}{\overline{\alpha}_i}\right)\right)\overline{\alpha}_i.
\end{aligned}$$

where in the second equality we use an property of m-set: $\sum_{i:x_i^*=1}(1-\overline{\alpha}_i) = \sum_{i:x_i^*=0}\overline{\alpha}_i$, which follows from the fact that $\overline{\alpha}$ is in the convex hull of m-set. In the last inequality, we simply lower bound $\sum_{i:x_i^*=0}\overline{\alpha}_i$ by one of its summands.

Using the same lower bound

$$r(\alpha) \geq \sum_{i:x_i^*=0} \Delta_i \overline{\alpha}_i, \qquad \text{(by Eq. (4.19))}$$

we have an upper bound for C_{add}:

$$\begin{aligned}
C_{add} &= \sum_{t=1}^{\infty} \max_{\alpha\in\Delta(\mathcal{X})} \frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha) \\
&\leq \sum_{i:x_i^*=0} \sum_{t=1}^{\infty} \max_{\overline{\alpha}_i\in[0,1]} \frac{100}{\sqrt{t}} \left(\gamma^{-1} + \gamma \log\frac{m}{\overline{\alpha}_i}\right)\overline{\alpha}_i - \Delta_i\overline{\alpha}_i \\
&\leq \sum_{i:x_i^*=0} \left(\sum_{t=1}^{\infty} \max_{\overline{\alpha}_i\in[0,1]} \left(\frac{100}{\sqrt{t}}\left(\gamma^{-1} + \gamma\log m\right)\overline{\alpha}_i - \frac{\Delta_i}{2}\overline{\alpha}_i\right) + \sum_{t=1}^{\infty} \max_{\overline{\alpha}_i\in[0,1]} \left(\frac{100}{\sqrt{t}}\gamma\overline{\alpha}_i \log\frac{1}{\overline{\alpha}_i} - \frac{\Delta_i}{2}\overline{\alpha}_i\right)\right).
\end{aligned}$$

Invoking Lemma 4.6 Lemma 4.7 on the above two terms, we get

$$C_{add} \leq \mathcal{O}\left(\sum_{i:x_i^*=0} \frac{(\gamma^{-1} + \gamma\log m)^2}{\Delta_i}\right).$$

This can be further upper bounded by $\mathcal{O}\left(\sum_{i:x_i^*=0} \frac{(\log d)^2}{\Delta_i}\right)$ by our selection of γ in either regime.

Chapter 5

Delayed feedback

The work presented in this chapter is based on a paper that has been accepted as [112].

[112] Zimmert, J. and Seldin, Y. (2020a). An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*

Table 5.1: Overview of state-of-the-art regret bounds for multi-armed bandits with delayed feedback. (*) requires oracle knowledge of the time horizon n and the total delay D; the result appeared independently in two papers. (**) requires advance knowledge of the delays d_t "at action time" t.

Setting	Regret upper and lower bounds		Reference
Uniform delays d	$\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$		Cesa-Bianchi et al. [34]
	$\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{dn\log(k)})$		Cesa-Bianchi et al. [34]
	$\mathcal{O}(\sqrt{kn} + \sqrt{dn\log(k)})$		This paper
Arbitrary delays,	$\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{D\log(k)})$	(*)	Thune et al. [98] Bistritz et al. [24]
non-adaptive bounds	$\mathcal{O}(\sqrt{k^2 n\log(k)} + \sqrt{D\log(k)})$		Bistritz et al. [24]
	$\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$		This paper
Arbitrary delays,	$\mathcal{O}(\min_\beta \lvert S_\beta\rvert + \beta\log(k) + \beta^{-1}(kn + D_{\bar{S}_\beta}))$	(**)	Thune et al. [98]
adaptive bounds	$\mathcal{O}(\sqrt{kn} + \min_S(\lvert S\rvert + \sqrt{D_{\bar{S}}\log(k)}))$		This paper

Abstract

We propose a new algorithm for adversarial multi-armed bandits with unrestricted delays. The algorithm is based on a novel hybrid regularizer applied in the Follow the Regularized Leader (FTRL) framework. It achieves $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$ regret guarantee, where k is the number of arms, n is the number of rounds, and D is the total delay. The result matches the lower bound within constants and requires no prior knowledge of n or D. Additionally, we propose a refined tuning of the algorithm, which achieves $\mathcal{O}(\sqrt{kn} + \min_S(|S| + \sqrt{D_{\bar{S}}\log(k)}))$ regret guarantee, where S is a set of rounds excluded from delay counting, $\bar{S} = [n] \setminus S$ are the counted rounds, and $D_{\bar{S}}$ is the total delay in the counted rounds. If the delays are highly unbalanced, the latter regret guarantee can be significantly tighter than the former. The result requires no advance knowledge of the delays and resolves an open problem of Thune et al. [98]. The new FTRL algorithm and its refined tuning are anytime and require no doubling, which resolves another open problem of Thune et al. [98].

5.1 Introduction

Multi-armed bandits are a fundamental sequential decision making problem with an increasing number of industrial applications. In the multi-armed bandit setting, a learner repeatedly chooses an action from a finite set of actions and immediately observes a loss for that specific action. The action might be, for example, a choice of an advertisement layout out of a finite set of layouts. The loss could be the response of a user to the layout, for example, a lack of a click on the advertisement. In practice, it is often required to make decisions for new users before observing the feedback of all previous users, either due to response latency or parallel interaction with multiple users. This can be modeled by introducing a *delay* between the action and observation.

We focus on the oblivious adversarial (a.k.a. non-stochastic) bandit setting, meaning that the sequence of losses and the delays are fixed before the start of the game. The setting was first studied by Cesa-Bianchi et al. [34] under the assumption of uniform delays, which are all equal to d. They proved a lower bound of $\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$ for $d \le n/\log(k)$ (they do not report the $\log(k)$ term) and an almost matching upper bound of $\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{dn\log(k)})$. By translating individual delays into the total delay $D = dn$ the lower bound for uniform delays is $\Omega(\max\{\sqrt{kn}, \sqrt{D\log(k)}\})$. Thune et al. [98] and Bistritz et al. [24] independently derived an algorithm that can handle non-uniform delays and achieves an $\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{D\log(k)})$ regret bound under the assumption that n and D are known in advance. Thune et al. further provide a doubling scheme that achieves the same regret bound under the assumption that the delays d_t are known "at action time", i.e., at time t, but n and D are unknown, whereas Bistritz et al. provide a doubling scheme that achieves an $\mathcal{O}(\sqrt{k^2n\log(k)} + \sqrt{D\log(k)})$ regret bound when n and D are unknown and the delays d_t are observed together with the observations, i.e., at time $t + d_t$.

Thune et al. further observe that if the delays are highly unbalanced it may be worth "skipping" rounds with excessively large delays. "Skipping" means that the regret in the corresponding round is trivially bounded by 1 and the observation is ignored by the algorithm. The skipping approach of Thune et al. requires knowledge of the delays "at action time". Under the assumption that this information is available, Thune et al. provide an algorithm that achieves $\mathcal{O}(\min_\beta |S_\beta| + \beta\log(k) + \beta^{-1}(kn + D_{\bar{S}_\beta}))$ regret guarantee, where β is the skipping threshold (the rounds with delays $d_t \geq \beta$ are skipped), S_β is the set of skipped rounds and $|S_\beta|$ is their number, $\bar{S}_\beta = [n] \setminus S_\beta$ are the remaining rounds (where $[n] = \{1, \ldots, n\}$), and $\mathrm{D}_{\bar{S}_\beta} = \sum_{t\in\bar{S}_\beta} d_t$ is their total delay. Thune et al. provide an example, where the first $\lfloor\sqrt{kn/\log(k)}\rfloor$ rounds have delays of order n and the remaining rounds have zero delays. By skipping the first rounds, the dependence of the regret bound on n improves from order $n^{3/4}$ to $n^{1/2}$. The skipping procedure of Thune et al. crucially depends on availability of delays "at action time" in order to make the skipping decision and the skipping threshold β is tuned by doubling. Relaxation of the assumption on early availability of delays, as well as replacement of doubling with anytime strategies (i.e., algorithms without resets) were left as open questions.

We resolve both open questions and make the following contributions:

1. We provide an anytime FTRL algorithm based on a novel hybrid regularizer. The regularizer combines $\frac{1}{2}$-Tsallis entropy and negative entropy, each with its own learning rate. The algorithm requires no advance knowledge of the delays and achieves a regret bound of $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$, which matches the lower bound within constants.

2. We provide a novel "skipping" technique, which allows to "ignore" rounds with excessively large delays with no advance knowledge of the delays. We put "skipping" and "ignore" in quotation marks, because the observations are still used by the algorithm and the "skipped" rounds are only excluded from updates of the learning rate. We prove an $\mathcal{O}(\sqrt{kn} + \min_S(|S| + \sqrt{D_{\bar{S}}\log(k)}))$ regret bound for the refined algorithm. The bound is slightly tighter than the refined regret bound of Thune et al. [98], but most importantly it requires no advance knowledge of the delays. [1]

In Table 5.1 we provide a comparison of state-of-the art bounds with our new results. Additional prior work in other online learning settings with delayed feedback includes the full information setting studied by Joulani et al. [58], who derived a general reduction to a non-delayed problem. To the best of our knowledge, no similar reduction under bandit feedback has been found yet. Another related setting are bandits with anonymous composite feedback, where the learner is not informed about the round from which the delayed observation is coming from, neither the identity of the action it corresponds to, and delayed observations from several rounds may be composed together with no possibility to separate them. This harder setting was studied by Cesa-Bianchi et al., who derived an $\mathcal{O}(\sqrt{kd_{max}n\log(k)})$ regret bound, where

[1] We note that the new skipping technique could also be combined with the doubling scheme of Thune et al. to eliminate the need in advanced knowledge of delays there as well. However, the anytime FTRL algorithm presented here is much more elegant than doubling.

d_{max} is a known upper bound on the delays. We refer the reader to Thune et al. [98] for further review of prior work in related settings.

The paper is structured in the following way: Section 5.2 provides a formal definition of the problem setting. Section 5.3 explains in detail our algorithm and two versions of learning rate tuning. Section 5.4 contains our main theorems, as well as an intuition behind the refined learning rate tuning. Section 5.5 presents a general analysis of FTRL for multi-armed bandits with delays and formally proves the theorems from the previous section. Finally, Section 5.6 provides a summary and directions for future work.

5.2 Problem setting

Adversarial bandits with delay is a sequential game between a learner and an environment with k fixed actions. At time steps $t = 1, \ldots, n$ the learner picks actions $A_t \in [k]$ and immediately suffers the loss ℓ_{t,A_t}, where $(\ell_t)_{t=1,\ldots,n}$ are vectors in $[0,1]^k$. Unlike in the regular bandit problem, the learner does not necessarily observe the loss ℓ_{t,A_t} at the end of round t. Instead, the environment chooses a sequence of delays $(d_t)_{t=1,\ldots,n}$ and the player observes the tuples (s, ℓ_{s,A_s}) for each s such that $s + d_s = t$ at the end of round t. Without loss of generality, we assume that all outstanding tuples are observed at the end of the game, i.e., $t + d_t \leq n$ for all t. We focus on the oblivious adversarial setting (sometimes called "non-stochastic"), which means that both the sequence of losses $(\ell_t)_{t=1,\ldots,n}$ and the sequence of delays $(d_t)_{t=1,\ldots,n}$ are chosen by the environment at the beginning of the game. We use $D = \sum_{t=1}^{n} d_t$ to denote the total delay. The learner has no prior knowledge of the quantities n, D, or $(d_t)_{t=1,\ldots,n}$. The performance of the algorithm is measured by its expected regret

$$\mathfrak{R}_n := \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t}\right] - \min_{i \in [k]} \sum_{t=1}^{n} \ell_{t,i}\,.$$

Some technical definitions We use $\Delta([k]) = \{x \in \mathbb{R}_+^k \| \|x\|_1 = 1\}$ to denote the $(k-1)$-simplex. For a set $S \subset [n] = \{1, \ldots, n\}$, we denote its complement by $\bar{S} = [n] \setminus S$. For a convex function F we use F^* to denote its convex conjugate (a.k.a. Fenchel conjugate) and $\overline{F}^*$ to denote the constrained convex conjugate. They are defined, respectively, by

$$\begin{aligned} F^*(y) &= \max_{x \in \mathbb{R}^k} \langle x, y \rangle - F(x), \\ \overline{F}^*(y) &= \max_{x \in \Delta([k])} \langle x, y \rangle - F(x)\,. \end{aligned}$$

5.3 Algorithm

Our Algorithm 7 is a standard Follow the Regularized Leader (FTRL) algorithm that works with importance weighted loss estimators of all observations available up to the current point in time. The loss estimators are defined by

$$\hat{\ell}_s = \frac{\ell_{s,A_s}}{x_{s,A_s}} \mathbf{e}_{A_s}\,,$$

where x_{s,A_s} is the algorithm's probability of selecting action A_s at round s and $\mathbf{e}_{A_s}$ is a standard basis vector. We define the cumulative observed loss estimator at time t by

$$\hat{L}_t^{obs} = \sum_{s: s+d_s<t} \hat{\ell}_s\,.$$

Given a convex regularizer $F_t : \mathbb{R}^k \to \mathbb{R}$, FTRL samples action A_t according to the distribution

$$x_t = \arg\min_{x \in \Delta([k])} \langle x, \hat{L}_t^{obs} \rangle + F_t(x) .$$

x_t can be equivalently expressed as $x_t = \nabla \overline{F}_t^*(-\hat{L}_t^{obs})$.

We are using a hybrid regularizer $F_t = F_{t,1} + F_{t,2}$, where in contrast to most prior work each of the two parts of the regularizer has its own learning rate.

$$\underbrace{F_t(x)}_{=\sum_{i=1}^k f_t(x_i)} = \underbrace{-\sum_{i=1}^k 2\sqrt{t} x_i^{1/2}}_{F_{t,1}(x)=\sum_{i=1}^k f_{t,1}(x_i)} + \underbrace{\eta_t^{-1} \sum_{i=1}^k x_i \log(x_i)}_{F_{t,2}(x)=\sum_{i=1}^k f_{t,2}(x_i)} .$$

The first part of the regularizer $F_{t,1}(x) = \sqrt{t}F_1(x)$ is the $\frac{1}{2}$-Tsallis entropy $F_1(x) = -2\sum_{i=1}^k \sqrt{x_i}$ with learning rate $\frac{1}{\sqrt{t}}$, which is non-adaptive to the problem. The second part of the regularizer $F_{t,2}(x) = \eta_t^{-1}F_2(x)$ is the negative entropy $F_2(x) = \sum_{i=1}^k x_i \log(x_i)$ with adaptive learning rate η_t. We call a sequence of learning rates $(\eta_t)_{t=1,\dots,n}$ *proper* if it is non-increasing and can be defined using information available at the beginning of round t.

5.3.1 Intuition behind the regularizer

Hybrid regularizers have been successfully used in adaptive regret bounds for sparse bandits, online portfolio selection, adversarially robust semi-bandits, and adaptive first order bounds for multi-armed bandits [27, 75, 82, 109]. They are useful for targeting multiple objectives. In our case, the regret lower bound for bandits with fixed delay d is $\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$ [34]. The first part of the bound is the standard regret lower bound for multi-armed bandits with no delays, which is clearly also a lower bound for the problem with delays. The second part of the bound is achieved by grouping the game rounds into batches of size d and reducing the game to a full information game over n/d rounds with loss range $[0, d]$. The second part is then a lower bound on the regret in the full information game.

Our regularizer uses the same decomposition of the problem. We combine the optimal regularizer for the standard bandit problem with no delay, the $\frac{1}{2}$-Tsallis Entropy, with the optimal regularizer for the full information problems, the negative entropy. We further tune the learning rate for the second part to the actual delay sequence $(d_t)_{t=1,\dots,n}$.

5.3.2 Tuning of the learning rate

We propose and analyze two versions of learning rate tuning. The *simple tuning* is given in Algorithm 7. For *advanced tuning*, replace the colored blocks **Initialize** and **determine** η_t in Algorithm 7 with the corresponding blocks from Algorithm 8.

Simple tuning We define the key quantity, which is used for tuning the learning rate.

Definition 5.1. *The* number of outstanding observations *at round t is defined by*

$$\mathfrak{d}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\},$$

where $\mathbb{I}$ is the indicator function.

Algorithm 7: FTRL for bandits with delay

```
Input: Proper learning rate rule $\eta_t$
Initialize $\hat{L}_1^{obs} = 0$
Initialize $\mathfrak{D}_0 = 0$                    (simple tuning)
for $t = 1, \dots, n$ do
    determine $\eta_t$
        Set $\mathfrak{D}_t = \mathfrak{D}_{t-1} + \mathfrak{d}_t$                 } (simple tuning)
        Set $\eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)}$          }
    Set $x_t = \arg\min_{x \in \Delta([k])} \langle x, \hat{L}_t^{obs} \rangle + F_t(x)$
    Sample $A_t \sim x_t$
    for $s : s + d_s = t$ do
        Observe $(s, \ell_{s,A_s})$
        Construct $\hat{\ell}_s$ and update $\hat{L}_t^{obs}$
```

$\mathfrak{d}_t$ counts how many observations from rounds $s < t$ are still missing at the beginning of round t. Note that $\mathfrak{d}_t$ is an observable quantity, unlike the delays d_t. Therefore, $\mathfrak{d}_t$ can be used for online tuning of the learning rate. The learning rate under the *simple tuning* is given by

$$\mathfrak{D}_t = \sum_{s=1}^{t} \mathfrak{d}_t \qquad , \qquad \eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)}\,.$$

The algorithm only uses the inverse of the learning rate. If $\mathfrak{D}_t = 0$, then $\eta_t^{-1} = 0$ and the algorithm is well-defined, even though $\eta_t = \infty$.

Advanced tuning In the advanced tuning, we maintain a running estimate $\tilde{\mathfrak{D}}_t$ of the optimal truncated delay $D_{\bar{S}}$. To achieve that, we modify the quantity $\mathfrak{d}_t$ by "skipping" some outstanding observations. To be precise, we keep indicator variables $a_s^t \in \{0,1\}$, where a_s^t indicates whether an outstanding observation from round s should still be counted at round t:

$$\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} a_s^t \mathbb{I}\{s + d_s \geq t\}.$$

We define

$$\tilde{\mathfrak{D}}_t = \sum_{s=1}^{t} \tilde{\mathfrak{d}}_t \qquad , \qquad \eta_t^{-1} = \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}\,.$$

The algorithm initially waits for all observations, but as soon as the waiting time exceeds a threshold the round is "skipped". If we observe a delay such that $d_s > \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}$, we set $(a_s^{t'})_{t' > t}$ to 0. The indicators are not changed retrospectively, which means that the initial waiting time still counts toward $\tilde{\mathfrak{D}}_t$. The intuition behind advanced tuning is explained in Section 5.4.3.

5.4 Main results

In this section, we present regret upper bounds for Algorithm 7 with *simple tuning* and *advanced tuning*. The first result confirms the conjecture of Cesa-Bianchi et al. [34] that an upper bound of $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$ is achievable with a simple algorithm. The second result shows that it is possible to obtain a refined bound of $\mathcal{O}(\sqrt{kn} + \min_{S \subset [n]}(|S| + \sqrt{D_{\bar{S}} \log(k)}))$ by a more careful tuning of the learning rate.

Algorithm 8: Advanced tuning of η_t for Alg. 7

```
Initialize $\tilde{\mathfrak{D}}_0 = 0$ and $(a_s^t)_{s=1,\dots,n;t=1,\dots,n} = 1$
determine $\eta_t$
    Set $\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\} a_s^t$
    Update $\tilde{\mathfrak{D}}_t = \tilde{\mathfrak{D}}_{t-1} + \tilde{\mathfrak{d}}_t$
    Set $\eta_t^{-1} = \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}$
    for $s = 1, \dots, t-1$ do
        if $\min\{d_s, t-s\} > \eta_t^{-1}$ then
            $(a_s^{t'})_{t'>t} = 0$     (At most one index $s$ satisfies the if-condition, see
             Lemma 5.7)
```

5.4.1 Adaptation to the total delay D

The following theorem provides a regret bound for Algorithm 7 with *simple tuning.*

Theorem 5.1. *The regret of Algorithm 7 with any non-increasing positive sequence of learning rates $(\eta_t)_{t=1,\dots,n}$ satisfies*

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \eta_n^{-1} \log(k) + \sum_{t=1}^{n} \eta_t \mathfrak{d}_t \,.$$

In particular, the simple tuning $\eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)} = \sqrt{2(\sum_{s=1}^{t} \mathfrak{d}_s)/\log(k)}$ is proper and leads to a regret bound of

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \sqrt{8D\log(k)} \,.$$

Proof. The first statement is a special case of Theorem 5.3, which is presented in Section 5.5. For the second statement we use a standard summation lemma, by which for a sequence of positive $\mathfrak{d}_1, \dots, \mathfrak{d}_n$ we have $\sum_{t=1}^{n} \left(\mathfrak{d}_t / \sqrt{\sum_{s=1}^{t} \mathfrak{d}_s} \right) \leq 2\sqrt{\sum_{s=1}^{t} \mathfrak{d}_t}$ [89, Lemma 8] and the convention that if $\mathfrak{d}_t = 0$ then $\eta_t \mathfrak{d}_t = 0$ (so that zero terms naturally fall out of the summation). By substituting the definition of the learning rate in the second statement into the first statement and using the summation lemma we obtain

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \sqrt{8\mathfrak{D}_n \log(k)} \,.$$

Finally, note that an observation from round t with delay d_t contributes 1 to each of $\mathfrak{d}_t, \dots, \mathfrak{d}_{t+d_t}$, i.e., it contributes d_t to the total sum of the number of outstanding observations $\sum_{t=1}^{n} \mathfrak{d}_t$. Since we have assumed that $t + d_t \leq n$ for all t, we have $\sum_{t=1}^{n} \mathfrak{d}_t = \sum_{t=1}^{n} d_t = D$. □

The main advantage of Algorithm 7 and Theorem 5.1 compared to the work of Thune et al. [98] is that the tuning requires neither the knowledge of D and n, nor doubling.

5.4.2 Refined bounds for unbalanced delays

Thune et al. [98] observed that if the delays are highly unbalanced it may be worth skipping rounds with overly large delays rather than keeping them in the analysis. Let S denote the set of skipped rounds and $|S|$ their number. The regret in every skipped round is trivially bounded by 1 and, assuming we knew which rounds to skip, we could reduce the regret bound to $\mathcal{O}\left(\sqrt{kn} + |S| + \sqrt{D_{\bar{S}} \log(k)}\right)$. As shown by Thune et al., this could potentially be much smaller than the regret bound in Theorem 5.1. For example, if the delay in the first $\theta(\sqrt{kn})$ rounds is of order n and the delay in the remaining rounds is zero, then the regret bound in Theorem 5.1 is of order $n^{3/4}$, whereas the refined regret bound is of order $n^{1/2}$ (ignoring the dependence on k). The challenge faced by Thune et al. was that they had to know the delays in advance (more precisely, "at action time") in order to tune the parameters of their algorithm and make the skipping decision. Since we have an anytime algorithm, we are able to obtain the refinement with no need in advance knowledge of the delay information. Strictly speaking, we even do not need to skip observations and we can obtain the refinement by using all observations and only adjusting the learning rate appropriately, although technically the "no-skipping" solution yields the same regret bound as skipping.

The following theorem provides our adaptive bound.

Theorem 5.2. *Algorithm 7 with advanced learning rate tuning provided in Algorithm 8 satisfies*

$$\begin{aligned}\mathfrak{R}_n &\leq 4\sqrt{kn}\\ &+ 10\max\begin{cases}\min_{S\subset[n]}|S| + \sqrt{D_{\bar{S}}\log(k)},\\ 2\log(k).\end{cases}\end{aligned}$$

The proof is postponed to Section 5.5

5.4.3 Intuition behind the "skipping" procedure

In order to give an intuition behind the refined algorithm we provide a simple back-of-the-envelope calculation. If we skip $|S|$ rounds and trivially bound their regret by 1 and apply Theorem 5.1 to the remaining rounds, then the regret bound is $\mathcal{O}(\sqrt{kn} + \sqrt{D_{\bar{S}}\log(k)} + |S|)$. Thus, the number of skipped rounds can be as large as $\sqrt{D_{\bar{S}}\log(k)}$ without significantly impacting the bound. Obviously, we want to skip rounds with the largest delays, but how should we determine the skipping threshold X? If we want to achieve a significant reduction in the regret bound, the skipped delay $D_S = \sum_{t\in S} d_t \geq X|S|$ should be at least as large as the remaining delay $D_{\bar{S}}$, because $D = D_S + D_{\bar{S}}$ and our aim is to reduce the $\sqrt{D\log k}$ term. Thus, if we put a threshold at X and skip $\sqrt{D_{\bar{S}}\log(k)}$ rounds we want to have $X\sqrt{D_{\bar{S}}\log(k)} \geq D_{\bar{S}}$. Therefore, we aim at $X = \sqrt{D_{\bar{S}}/\log(k)}$. However, there are two challenges: (a) we do not know the delays d_t in advance and, therefore, we do not know which rounds to skip, and (b) the threshold definition is recursive: X depends on $D_{\bar{S}}$ and $D_{\bar{S}}$ depends on X.

The strategy that we take in Algorithm 8 is the following: we keep a running estimate $\tilde{\mathfrak{D}}_t$ of $D_{\bar{S}}$. For an observation from round s we initially start waiting and count it in the number of outstanding observations $\tilde{\mathfrak{d}}_t$ for the initial rounds. However, we constantly monitor the waiting time and if the observation has not arrived within $\sqrt{\tilde{\mathfrak{D}}_t/\log(k)}$ rounds we stop waiting. The initial rounds we have been waiting for still count for the estimate $\tilde{\mathfrak{D}}_t$. Another quick back-of-the-envelope calculation shows that if $\tilde{\mathfrak{D}}_t$ is indeed a good approximation of $D_{\bar{S}}$, then the extra delay from the initial waiting rounds is of order $\sqrt{D_{\bar{S}}\log(k)}\sqrt{D_{\bar{S}}/\log(k)} = D_{\bar{S}}$, where the first term is a rough estimate of the number of rounds that we skip and the second term is a rough estimate of the initial waiting time for each of the observations. Thus, the initial waiting time has no significant impact on the final bound.

Algorithm 8 follows this intuitive approach. We use indicator variables $(a_s^t)_{(s,t)\in[n]^2}$ to keep track of which observations ℓ_{s,A_s} we are still waiting for at round t (expressed by $a_s^t = 1$) and which not (expressed by $a_s^t = 0$). We use $\tilde{\mathfrak{d}}_t$ to count the truncated number of outstanding observations, where those observations we are no longer waiting for at round t are excluded from counting. We provide a detailed analysis in Section 5.5.2, but before we get there we provide a refined version of Theorem 5.1, which allows us to use all observations and only use skipping in the tuning of the learning rate. (Though, as already mentioned, complete skipping of the observations would lead to the same regret bound as in Theorem 5.2.)

5.5 Analysis of FTRL for bandits with delays

In this section we develop a novel analysis of FTRL-style algorithms and present a generalization of the first part of Theorem 5.1. The analysis is based on a permuted counting of losses, similar to the techniques used by Joulani et al. [57] and Thune et al. [98]. Afterward, we use the general regret bound to prove Theorem 5.2.

5.5.1 Dependency preserving permutations

Reordering of losses by a permutation $\rho : [n] \to [n]$ is a useful tool in the analysis of online learning with delays. Joulani et al. [57] have used "ordering by arrival", where the losses $\hat{\ell}_s$ are sorted by the time of arrival $s + d_s$ with ties broken arbitrarily. We generalize this type of analysis by studying a general class of admissible permutations. This also provides insights into why it is useful to consider permutations.

Definition 5.2. *A permutation $\rho : [n] \to [n]$ is* dependency preserving *if it satisfies:*

$$\forall\, s, t \in [n] :\ s + d_s < t \ \Rightarrow\ \rho(s) < \rho(t)\,.$$

*It means that if at the beginning of round t the loss ℓ_{s,A_s} has been already observed (and thus can influence the selection of A_t), then s must come before t under the permutation. Furthermore, we define the ρ-*number of outstanding observations *at time t by*

$$\mathfrak{d}_t^\rho = \sum_{s:\rho(s)<\rho(t)} \mathbb{I}\{s + d_s \geq t\}\,.$$

Example 5.1. *The identity function* $\mathrm{id}(t) = t$ *is dependency preserving, since ℓ_{s,A_s} being observed before t implies* $\mathrm{id}(s) = s < t = \mathrm{id}(t)$*.*

Example 5.2. *"Ordering by arrival" is dependency preserving, since $s + d_s < t \Rightarrow s + d_s < t + d_t \Rightarrow \rho(s) < \rho(t)$.*

The ρ-number of outstanding observations extends the previous definition of the number of outstanding observations in the sense that $\mathfrak{d}_t = \mathfrak{d}_t^{id}$. Furthermore, the property $\sum_{t=1}^n \mathfrak{d}_t^\rho = D$ holds for any dependency preserving permutation ρ (refer to Lemma 5.6 in the supplementary material, Section 5.6.1).

Next we present a general regret bound which holds for any dependency preserving permutation ρ.

Theorem 5.3. *For any dependency preserving permutation ρ, the regret of Algorithm 7 with non-increasing positive learning rates $(\eta_t)_{t=1,\dots,n}$ satisfies*

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \eta_n^{-1}\log(k) + \sum_{t=1}^n \min\{1, \eta_t \mathfrak{d}_t^\rho\}\,.$$

Remark 5.1. *The first part of Theorem 5.1 is a direct corollary using $\rho = \mathrm{id}$.*

The proof uses Lemmas 5.1, 5.2, and 5.3. In order to motivate them we first present the proof and then the lemmas.

Proof. We define cumulative losses $\hat{L}_t^\rho = \sum_{s:\rho(s)<\rho(t)} \hat{\ell}_s$ (and by convention $\hat{L}_{n+1}^\rho = \sum_{s=1}^n \hat{\ell}_s$) and $i^* = \arg\min \sum_{t=1}^n \ell_{t,i}$. We decompose the regret into three terms:

$$\begin{aligned}
\mathfrak{R}_n &= \mathbb{E}\left[\sum_{t=1}^n \ell_{t,A_t} - \ell_{t,i^*}\right] = \mathbb{E}\left[\sum_{t=1}^n \langle x_t, \hat{\ell}_t\rangle - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right]\\
&= \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^n \left(\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t\rangle\right)}_{(A)}\\
&\qquad + \underbrace{\begin{aligned}\sum_{t=1}^n \Bigg(&\overline{F}_t^*(-\hat{L}_t^{obs}) - \overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t)\\
&-\overline{F}_t^*(-\hat{L}_t^\rho) + \overline{F}_t^*(-\hat{L}_{t+1}^\rho)\Bigg)\end{aligned}}_{(B)}\\
&\qquad + \underbrace{\sum_{t=1}^n \left(\overline{F}_t^*(-\hat{L}_t^\rho) - \overline{F}_t^*(-\hat{L}_{t+1}^\rho) - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right)}_{(C)}\Bigg].
\end{aligned}$$

Term (A) is a typical Bregman divergence term from the classical FTRL/OMD analysis and depends on the local norm of the regularizer. Lemma 5.1 directly gives

$$\mathbb{E}[(A)] \leq \sum_{t=1}^n \sqrt{k}/\sqrt{t} \leq 2\sqrt{kn}\,.$$

Term (C) can also be bounded by standard techniques. Lemma 5.2 gives us

$$(C) \leq 2\sqrt{kn} + \eta_n^{-1}\log(k)\,.$$

Term (B) requires a novel analysis, which is presented in Lemma 5.3. This allows to bound the second term by

$$\mathbb{E}[(B)] \leq \sum_{t=1}^n \min\{1, \eta_t \mathfrak{d}_t^\rho\}\,.$$

Combining everything finishes the proof. □

Support lemmas for the proof of Theorem 5.3 The proofs for all the support lemmas are given in the supplementary material, Section 5.6.4. The first Lemma is a small modification of the classical result that bounds the Bregman divergence by the local norm of the regularizer. We show that we can bound the local norm by the contribution of the Tsallis entropy.

Lemma 5.1. *For any t it holds that*

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t\rangle\right] \leq \frac{\sqrt{k}}{\sqrt{t}}\,.$$

The second Lemma bounds the so-called “penalty” term coming from the regularization penalty. It appears in almost identical form in the literature [70, Exercise 28.12].

Lemma 5.2. *For any non-increasing learning rate η_t, it holds that*

$$\sum_{t=1}^{n}\left(\overline{F}_t^*(-\hat{L}_t^{\rho}) - \overline{F}_t^*(-\hat{L}_{t+1}^{\rho}) - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right) \\ \leq 2\sqrt{kn} + \eta_n^{-1}\log(k)\,.$$

The third quantity does not show up in the regular analysis without delays. We show that similarly to the Bregman divergence, it depends on the local norm of the regularizer. However, it is beneficial to use the norm of the negative entropy instead of the Tsallis entropy.

Lemma 5.3. *For any t it holds that*

$$\mathbb{E}\Bigg[\overline{F}_t^*(-\hat{L}_t^{obs}) - \overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) \\ - \overline{F}_t^*(-\hat{L}_t^{\rho}) + \overline{F}_t^*(-\hat{L}_{t+1}^{\rho})\Bigg] \leq \min\{1, \eta_t\tilde{\mathfrak{d}}_t^{\rho}\}\,.$$

5.5.2 Refined regret bound

The reason why it is beneficial to consider permutations in the analysis is the following lemma.

Lemma 5.4. *For any $S \subset [n]$ there exists a dependency preserving permutation ρ, such that*

$$\forall t \in \bar{S}: \quad \mathfrak{d}_t^{\rho} = \sum_{s:s<t}\mathbb{I}\{s\in\bar{S}\}\mathbb{I}\{s+d_s\geq t\}\,.$$

Furthermore, this implies $\sum_{t\in\bar{S}}\mathfrak{d}_t^{\rho} \leq \sum_{t\in\bar{S}} d_t$.

An iterative procedure for construction of ρ is given in the supplementary material, Section 5.6.1. The lemma allows to split the rounds into sets S and $\bar{S}$ and construct a permutation, so that the number of outstanding delays for rounds in $\bar{S}$ only depends on the delays in other rounds in $\bar{S}$, but not on rounds in S. Fig. 5.1 provides an example of construction of such a permutation. The lemma is particularly useful for splitting the rounds into a set S containing large delays and the complementary set $\bar{S}$ containing small delays. Then the lemma allows to "push" the contributions to the ρ-number of outstanding observations away from the elements in $\bar{S}$ to the elements in S. Skipping the rounds in S yields the highest benefit.

Combining Lemma 5.4 with Theorem 5.3 and a suitable learning rate leads directly to the bound

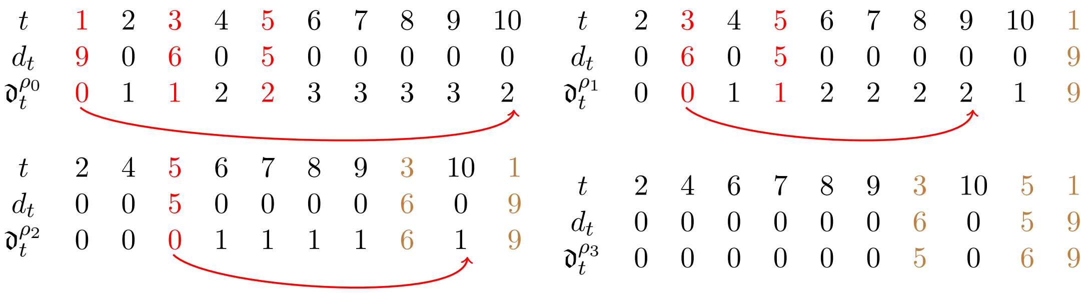
$$\mathfrak{R}_n \leq 4\sqrt{kn} + |S| + 2\sqrt{\sum_{t\in\bar{S}} d_s\log(k)}\,.$$

In the following proof, we show that the learning rate in Algorithm 8 brings us within a constant of the minimum of the above bound, $4\sqrt{kn} + \min_S(|S| + 2\sqrt{D_{\bar{S}}\log(k)})$.

From now on, let S be the set

$$S = \{t \in [n]\,|\,a_t^n = 0\}\,,$$

which is the set of rounds "skipped" by Algorithm 8, and let ρ be the associated permutation from Lemma 5.4. Since $(a_s^t)_{t=1,\dots,n}$ is non-increasing, we have for any $t\in\bar{S}$: $\mathfrak{d}_t^{\rho} \leq \tilde{\mathfrak{d}}_t$. Furthermore, the following lemma bounds the magnitude of $|S|$:



t	1	2	3	4	5	6	7	8	9	10
d_t	9	0	6	0	5	0	0	0	0	0
$\mathfrak{d}_t^{\rho_0}$	0	1	1	2	2	3	3	3	3	2

t	2	3	4	5	6	7	8	9	10	1
d_t	0	6	0	5	0	0	0	0	0	9
$\mathfrak{d}_t^{\rho_1}$	0	0	1	1	2	2	2	2	1	9

t	2	4	5	6	7	8	9	3	10	1
d_t	0	0	5	0	0	0	0	6	0	9
$\mathfrak{d}_t^{\rho_2}$	0	0	0	1	1	1	1	6	1	9

t	2	4	6	7	8	9	3	10	5	1
d_t	0	0	0	0	0	0	6	0	5	9
$\mathfrak{d}_t^{\rho_3}$	0	0	0	0	0	0	5	0	6	9

Figure 5.1: An iterative construction of the permutation in Lemma 5.4. Colored columns are elements in S.

Lemma 5.5. *For any sequence of delays d_t, Algorithm 8 satisfies*

$$|S| = \sum_{t=1}^{n} \mathbb{I}\{a_t^n = 0\} \leq 2\sqrt{\tilde{\mathfrak{D}}_n \log(k)}\,.$$

The proof is provided in the supplementary material, Section 5.6.2.

Finally we have all the prerequisites to prove Theorem 5.2 .

Proof of Theorem 5.2. Using Theorem 5.3 and Lemma 5.5 with ρ constructed for S, we have

$$\begin{aligned}
\mathfrak{R}_n &\leq 4\sqrt{kn} + \eta_n^{-1}\log(k) + \sum_{t=1}^{n} \min\{1, \eta_t \mathfrak{d}_t^{\rho}\} \\
&\leq 4\sqrt{kn} + \eta_n^{-1}\log(k) + |S| + \sum_{t\in\bar{S}} \eta_t \tilde{\mathfrak{d}}_t \\
&\leq 4\sqrt{kn} + 5\sqrt{\tilde{\mathfrak{D}}_n \log(k)}\,.
\end{aligned}$$

Now we need to control the term $\sqrt{\tilde{\mathfrak{D}}_n \log(k)}$. Let's consider the case $\tilde{\mathfrak{D}}_n \leq 4\sqrt{\tilde{\mathfrak{D}}_n \log(k)}$, then $\sqrt{\tilde{\mathfrak{D}}_n \log(k)} \leq 4\log(k)$ and we are done. Otherwise, define $\tilde{d}_t = \sum_{s=t+1}^{t+d_t} a_t^s$, i.e., the contribution of round t to the sum $\tilde{\mathfrak{D}}_n$. Then we can decompose

$$\begin{aligned}
\tilde{\mathfrak{D}}_n &= \sum_{s=1}^{n}\sum_{t<s} \mathbb{I}\{t + d_t > s\} a_t^s \\
&= \sum_{t=1}^{n}\sum_{s>t} \mathbb{I}\{t + d_t > s\} a_t^s \\
&= \sum_{t=1}^{n}\sum_{s=t+1}^{t+d_t} a_t^s = \sum_{t=1}^{n} \tilde{d}_t\,.
\end{aligned}$$

Any element $t \in \bar{S}$ satisfies

$$\tilde{d}_t \leq \sqrt{\tilde{\mathfrak{D}}_t / \log(k)} \leq \sqrt{\tilde{\mathfrak{D}}_n / \log(k)}\,,$$

while any element $t \in S$ satisfies

$$\begin{aligned}
\tilde{d}_t &\leq \left\lceil \sqrt{\tilde{\mathfrak{D}}_t / \log(k)} \right\rceil \leq \left\lceil \sqrt{\tilde{\mathfrak{D}}_n / \log(k)} \right\rceil \\
&\leq \sqrt{\tilde{\mathfrak{D}}_n / \log(k)} + 1\,.
\end{aligned}$$

Therefore, we can bound for any $R \subset [n]$:

$$\begin{aligned}\sum_{t\in\bar{R}} d_t \geq \sum_{t\in\bar{R}} \tilde{d}_t &\geq \tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)} - |S| \\ &\geq \tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)} - 2\sqrt{\tilde{\mathfrak{D}}_n\log(k)} \\ &\geq \frac{1}{2}\tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)}\,.\end{aligned}$$

This implies that

$$\begin{aligned}&\min_{R\subset[n]} |R| + \sqrt{\sum_{t\in\bar{R}} d_t\log(k)} \\ &\geq \min_{r\in[0,\frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}]} r + \sqrt{\frac{1}{2}\tilde{\mathfrak{D}}_n\log(k) - r\sqrt{\tilde{\mathfrak{D}}_n\log(k)}}\,.\end{aligned}$$

The function is concave in r so the minimum is achieved at one of the endpoints of the interval, which happens to be $r = \frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}$ for which the function equals $\frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}$. Hence, we have shown

$$\sqrt{\tilde{\mathfrak{D}}_n\log(k)} \leq 2\min_{R\subset[n]}\left(|R| + \sqrt{\sum_{s\in\bar{R}} d_s\log(k)}\right),$$

which concludes the proof. □

5.6 Discussion

We confirmed an open conjecture from Cesa-Bianchi et al. [34] by presenting a simple FTRL algorithm for adversarial bandits with arbitrary delays and proving regret upper bound that matches the lower bound within constants. Furthermore, we proposed a refined tuning of the learning rate that achieves even tighter regret bound for highly unbalanced delays. We strictly improve on the state-of-the-art bounds and present the first anytime result requiring no doubling, skipping, or advance information about the delays.

If the delays are all 0, then our algorithm reduces to the Tsallis-INF algorithm of Zimmert and Seldin [111], which has been proven to be simultaneously optimal in both the stochastic and the adversarial setting. We conjecture that the algorithm presented in this paper is capable of obtaining logarithmic regret in the stochastic setting, but leave the analysis for future work.

Another open question is the tightness of our adaptive bound $\mathcal{O}(\sqrt{kn} + \min_{S\subset[n]}(|S| + \sqrt{D_S\log(k)}))$. We conjecture that for a fixed set of delays $\{d_1, \ldots, d_n\}$ which the adversary is allowed to permute without changing the magnitudes, the upper bound is actually tight.

Acknowledgements

We would like to thank András György and Tobias Sommer Thuner for fruitful discussions. The work was partly supported by the Independent Research Fund Denmark, grant number 9040-00361B.

Appendix

5.6.1 Properties of dependency preserving permutations

Lemma 5.6. *For any dependency preserving ρ, the sum of ρ-number of outstanding observations is identical to the total sum of delays:*

$$\sum_{t=1}^{n} \mathfrak{d}_t^{\rho} = \sum_{t=1}^{n} d_t = D\,.$$

Proof of Lemma 5.6.

$$\begin{aligned}\sum_{t=1}^{n} \mathfrak{d}_t^{\rho} &= \sum_{t=1}^{n} \sum_{s:\rho(s)<\rho(t)} \mathbb{I}\{s+d_s \geq t\}\\ &= \sum_{t=1}^{n} \rho(t) - 1 - \sum_{s:\rho(s)<\rho(t)} \mathbb{I}\{s+d_s < t\}\\ &= \sum_{t=1}^{n} t - 1 - \sum_{s} \mathbb{I}\{s+d_s < t\}\\ &= \sum_{t=1}^{n} \sum_{s:s<t} \mathbb{I}\{s+d_s \geq t\}\\ &= \sum_{s=1}^{n} \sum_{t:t>s} \mathbb{I}\{s+d_s \geq t\}\\ &= \sum_{s=1}^{n} d_s = D\,.\end{aligned}$$

□

Proof of Lemma 5.4. We define the permutation ρ iteratively. Let $\rho_0 = \text{id}$ be the identity permutation and let $(t_1, t_2, \dots, t_{|S|})$ be an increasing indexing of the set S. We iteratively define

$$\rho_m(s) := \begin{cases} \rho_{m-1}(t_m + d_{t_m}), & \text{if } s = t_m,\\ \rho_{m-1}(s), & \text{for } \rho_{m-1}(s) < \rho_{m-1}(t_m)\,,\\ \rho_{m-1}(s), & \text{for } \rho_{m-1}(s) > \rho_{m-1}(t_m + d_{t_m})\,,\\ \rho_{m-1}(s) - 1, & \text{otherwise.}\end{cases}$$

To get from ρ_{m-1} to ρ_m, we only move the element t_m directly behind the the element $t_m + d_{t_m}$. The final permutation is $\rho = \rho_{|S|}$.

Proof that ρ is dependency preserving Since we start with a dependency preserving permutation ρ_0, we only need to prove the induction step that ρ_m is dependency preserving under the condition that ρ_{m-1} is. In the step from ρ_{m-1} to ρ_m, we move the point t_m to the right, so for any $t \neq t_m$ and any s, we have that $\rho_{m-1}(t) < \rho_{m-1}(s) \Rightarrow \rho_m(t) < \rho_m(s)$. Hence, by the induction condition, for any $t \neq t_m$ we have: $t+d_t < s \Rightarrow \rho_{m-1}(t) < \rho_{m-1}(s) \Rightarrow \rho_m(t) < \rho_m(s)$. We only need to verify that $t_m+d_{t_m} < s \Rightarrow \rho_m(t_m) < \rho_m(s)$. In the permutation ρ_{m-1}, we know that $t_m + d_{t_m} < s \Rightarrow \rho_{m-1}(t_m + d_{t_m}) < \rho_{m-1}(s)$, because only elements smaller than t_m have been moved. The construction of ρ_m defines $\rho_m(t_m) = \rho_{m-1}(t_m + d_{t_m})$ and $\rho_m(s) = \rho_{m-1}(s)$ for all $\rho_{m-1}(s) > \rho_{m-1}(t_m + d_{t_m})$. Hence we have $t_m + d_{t_m} < s \Rightarrow \rho_m(t_m) = \rho_{m-1}(t_m + d_{t_m}) < \rho_{m-1}(s) = \rho_m(s)$, which concludes the first part of the lemma.

ρ-number of outstanding delays By definition we have

$$
\begin{aligned}
\mathfrak{d}_t^\rho := \sum_{s:\rho(s)<\rho(t)} \{s+d_s \geq t\} &= \sum_{s\in[n]\setminus S:\rho(s)<\rho(t)} \{s+d_s \geq t\} + \sum_{s\in S:\rho(s)<\rho(t)} \mathbb{I}\{s+d_s \geq t\} \\
&\overset{(a)}{=} \sum_{s\in[n]\setminus S:s<t} \mathbb{I}\{s+d_s \geq t\} + \sum_{s\in S:\rho(s)<\rho(t)} \mathbb{I}\{s+d_s \geq t\} \\
&\overset{(b)}{=} \sum_{s\in[n]\setminus S:s<t} \{s+d_s \geq t\} = \sum_{s:s<t} \mathbb{I}\{s \in [n]\setminus S\}\mathbb{I}\{s+d_s \geq t\}\,.
\end{aligned}
$$

(a) holds because the ordering does not change for $s, t \in [n] \setminus S$. (b) follows because any $s \in S$ has been moved behind $t' = s + d_s$, so $s + d_s \geq t$ implies $\rho(s) > \rho(t)$.

Bounded sum Using the above property, we have

$$
\begin{aligned}
\sum_{t\in[n]\setminus S} \mathfrak{d}_t^\rho &= \sum_{t\in[n]\setminus S} \sum_{s=1}^{t-1} \mathbb{I}\{s \in [n]\setminus S\}\mathbb{I}\{s+d_s \geq t\} \\
&= \sum_{s\in[n]\setminus S} \sum_{t=s+1}^{n} \mathbb{I}\{s \in [n]\setminus S\}\mathbb{I}\{s+d_s \geq t\} \\
&= \sum_{s\in[n]\setminus S} \sum_{t=s+1}^{s+d_s} \mathbb{I}\{s \in [n]\setminus S\} \leq \sum_{s\in[n]\setminus S} d_s\,.
\end{aligned}
$$

□

5.6.2 Auxiliary lemmas for Algorithm 8

Lemma 5.7. *Algorithm 8 will not deactivate more than 1 point at a time.*

By *deactivating* we mean setting $a_t^n = 0$.

Proof. We prove the lemma by contradiction. Assume that s_1, s_2 are both deactivated at time t. W.l.o.g. let $s_2 \leq s_1 - 1$. Deactivation of s_1 at time t means $t - s_1 \geq \sqrt{\mathfrak{D}_t/\log(k)} \geq \sqrt{\mathfrak{D}_{t-1}/\log(k)}$. At the same time we assumed $t - 1 - s_2 \geq t - s_1$, which means that s_2 would have been deactivated at round $t-1$ or earlier. □

Proof of Lemma 5.5. Recall that $\tilde{d}_t = \sum_{s=t+1}^{t+d_t} a_t^s$ is the contribution of a timestep t to the sum $\tilde{\mathfrak{D}}_n$.

Let $(t_1, \ldots, t_{|S|})$ be an indexing of S. By Lemma 5.7 we deactivate at most one $a_{t_m}^n$ per round. Thus, we have that

$$
\tilde{d}_{t_m} > \sqrt{\tilde{\mathfrak{D}}_{t_m+d_{t_m}}/\log(k)} \geq \sqrt{\sum_{i=1}^{m} \tilde{d}_{t_i}/\log(k)} = \frac{\sqrt{\tilde{d}_m + \sum_{i=1}^{m-1} \tilde{d}_{t_i}}}{\sqrt{\log(k)}}\,.
$$

By solving the quadratic inequality in d_{t_m} we obtain

$$
\tilde{d}_{t_m} > \frac{1 + \sqrt{1 + 4\log(k)\sum_{i=1}^{m-1} \tilde{d}_{t_i}}}{2\log(k)}\,.
$$

Now we prove by induction that $\tilde{d}_{t_m} > \frac{m}{2\log(k)}$. The induction base holds since $\tilde{d}_{t_1} = 1$. For the inductive step we have

$$\tilde{d}_{t_m} > \frac{1 + \sqrt{1 + 4\log(k)\sum_{i=1}^{m-1}\tilde{d}_{t_i}}}{2\log(k)} > \frac{1 + \sqrt{1 + m(m-1)}}{2\log(k)} > \frac{m}{2\log(k)}\,.$$

Finally, we have

$$\sqrt{\tilde{\mathfrak{D}}_n \log(k)} \geq \sqrt{\sum_{m=1}^{|S|} \tilde{d}_{t_m}\log(k)} > \sqrt{\frac{|S|(|S|+1)}{4}} > \frac{1}{2}|S|\,.$$

□

5.6.3 Standard properties of FTRL analysis

First we list some standard properties of FTRL that we use in the proofs of the remaining lemmas. We recall that $f_t(x) = -2\sqrt{t}\sqrt{x} + \eta_t^{-1}x\log(x)$.

Fact 5.1. *$f_t''(x) : \mathbb{R}_+ \to \mathbb{R}_+$ are monotonically decreasing functions and ${f^*}_t' : \mathbb{R} \to \mathbb{R}_+$ are convex and monotonically increasing.*

Proof. By definition $f_t''(x) = \frac{1}{2}\sqrt{t}x^{-3/2} + \eta_t^{-1}x^{-1}$, which concludes the first statement. Since f_t are Legendre functions, we have $f_t^{*\prime\prime}(y) = f_t''(f_t^{*\prime}(y))^{-1} > 0$. Therefore the function is monotonically increasing. Since both $f_t''(x)^{-1}$, as well as $f_t^{*\prime}(y)$ are increasing, the composition is as well and $f_t^{*\prime\prime\prime} > 0$. □

Fact 5.2. *For any convex F, for $L \in \mathbb{R}^k$ and $c \in \mathbb{R}$:*

$$\overline{F}^*(L + c\mathbf{1}_k) = \overline{F}^*(L) + c\,.$$

Proof. By definition $\overline{F}^*(L + c\mathbf{1}_k) = \max_{x\in\Delta([k])}\langle x, L + c\mathbf{1}_k\rangle - F(x) = \max_{x\in\Delta([k])}\langle x, L\rangle - F(x) + c = \overline{F}^*(L) + c$. □

Fact 5.3. *For any x_t there exists $c \in \mathbb{R}$, such that:*

$$x_t = \nabla\overline{F}_t^*(-\hat{L}_t^{obs}) = \nabla F_t^*(-\hat{L}_t^{obs} + c\mathbf{1}_k) = \nabla F_t^*(\nabla F_t(x_t))\,.$$

Proof. By the KKT conditions, there exists $c \in \mathbb{R}$, such that $x_t = \arg\max_{x\in\Delta([k])}\langle x, -\hat{L}_t^{obs}\rangle + F_t(x)$ satisfies $\nabla F_t(x_t) = -\hat{L}_t^{obs} + c\mathbf{1}_k$. The rest follows by the standard property $\nabla F = (\nabla F^*)^{-1}$ of Legendre F. □

Fact 5.4. *For any Legendre function F and $L \in \mathbb{R}^k$ it holds that*

$$\overline{F}^*(L) \leq F^*(L)$$

with equality iff there exists $x \in \Delta([k])$, such that $L = \nabla F(x)$.

Proof. The first statement follows from the definition, since for any $A \subset B$: $\max_{x\in A} f(x) \leq \max_{x\in B} f(x)$. The second part follows because equality means that $\arg\max_x\langle x, L\rangle - F(x) = \nabla F^*(L) \in \Delta([k])$, which is equivalent to the statement. □

Fact 5.5. *For any $x \in \Delta([k])$, $L \geq 0$ and $i \in [k]$:*

$$\nabla\overline{F}_t^*(\nabla F_t(x) - L)_i \geq \nabla F_t^*(\nabla F_t(x) - L)_i\,.$$

Proof. By fact 5.3, there exists $c \in \mathbb{R} : \nabla\overline{F}_t^*(\nabla F_t(x) - L) = \nabla F_t^*(\nabla F_t(x) - L + c\mathbf{1}_k)$. The statement is equivalent to c being non-negative, since $f^{*\prime}$ are monotonically increasing. If $c < 0$, then

$$1 = \sum_{i=1}^k (\nabla\overline{F}_t^*(\nabla F_t(x) - L))_i = \sum_{i=1}^k (\nabla F_t^*(\nabla F_t(x) - L + c\mathbf{1}_k))_i = \sum_{i=1}^k f_t^{*\prime}(f_t'(x_i) - L_i + c) < \sum_{i=1}^k f_t^{*\prime}(f_t'(x_i)) = 1\,,$$

which is a contradiction and completes the proof. □

Fact 5.6. *Let $D_F(x, y) = F(x) - F(y) - \langle x - y, \nabla F(y)\rangle$ be the Bregman divergence of a function F. For any Legendre function f with monotonically decreasing second derivative, $x \in \operatorname{dom}(f)$, and $\ell \geq 0$, such that $f'(x) - \ell \in \operatorname{dom}(f^*)$:*

$$D_{f^*}(f'(x) - \ell, f'(x)) \leq \frac{\ell^2}{2f''(x)}\,.$$

Proof. By Taylor's theorem, there exists $\tilde{x} \in [f^{*\prime}(f'(x) - \ell), x]$, such that $D_{f^*}(f'(x) - \ell, f'(x)) = \frac{\ell^2}{2f''(\tilde{x})}$. $\tilde{x}$ is smaller than x, since $f^{*\prime}$ is monotonically increasing. Finally, using the fact that the second derivative is decreasing allows to bound $f''(\tilde{x})^{-1} \leq f''(x)^{-1}$. □

5.6.4 Proofs of the Main Lemmas

Proof of Lemma 5.1.

$$\begin{aligned}
\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t\rangle &\overset{\text{(a)}}{=} \overline{F}_t^*(\nabla F_t(x_t) - \hat{\ell}_t) - \overline{F}_t^*(\nabla F_t(x_t)) + \langle x_t, \hat{\ell}_t\rangle \\
&\overset{\text{(b)}}{\leq} F_t^*(\nabla F_t(x_t) - \hat{\ell}_t) - F_t^*(\nabla F_t(x_t)) + \langle x_t, \hat{\ell}_t\rangle \\
&= \sum_{i=1}^k D_{f_t^*}(f_t'(x_{t,i}) - \hat{\ell}_{t,i}, f_t'(x_{t,i})) \\
&= D_{f_t^*}(f_t'(x_{t,A_t}) - \ell_{t,A_t} x_{t,A_t}^{-1}, f_t'(x_{t,A_t})) \\
&\overset{\text{(c)}}{\leq} \frac{1}{2}\ell_{t,A_t}^2 x_{t,A_t}^{-2} f_t''(x_{t,A_t})^{-1} \\
&\leq \frac{1}{2}\ell_{t,A_t}^2 x_{t,A_t}^{-2} f_{t1}''(x_{t,A_t})^{-1} \\
&= \frac{1}{2}\ell_{t,A_t}^2 x_{t,A_t}^{-2} \frac{2x_{t,A_t}^{\frac{3}{2}}}{\sqrt{t}} \\
&\leq \frac{x_{t,A_t}^{-\frac{1}{2}}}{\sqrt{t}}\,.
\end{aligned}$$

(a) Applies facts 5.2 and 5.3. (b) Follows from both parts of fact 5.4. (c) Uses fact 5.6. In expectation we get

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t\rangle\right] \leq \sum_{i=1}^k \frac{\sqrt{x_{t,i}}}{\sqrt{t}} \leq \frac{\sqrt{k}}{\sqrt{t}}\,.$$

□

Proof of Lemma 5.2. Let $\tilde{x}_t = \arg\max_{x \in \Delta([k])}\langle x, -\hat{L}_t^{\rho}\rangle - F_t(x)$, then

$$\overline{F}_t^*(-\hat{L}_t^{\rho}) = \langle \tilde{x}_t, \hat{L}_t^{\rho}\rangle - F_t(\tilde{x}_t).$$

Furthermore, since $\overline{F}^*(-\hat{L}^\rho_t) = \max_{x\in\Delta([k])}\langle x, -\hat{L}^\rho_t\rangle - F(x)$, we have

$$\begin{aligned} -\overline{F}^*_{t-1}(-\hat{L}^\rho_t) &\leq \langle \tilde{x}_t, -\hat{L}^\rho_t\rangle - F_{t-1}(\tilde{x}_t)\,, \\ -\overline{F}^*_n(-\hat{L}^\rho_{n+1}) &\leq \langle \mathbf{e}_{i^*}, -\hat{L}^\rho_{n+1}\rangle F_n(\mathbf{e}_{i^*}) = \sum_{t=1}^n \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\,. \end{aligned}$$

Plugging these inequalities into the LHS leads to

$$\begin{aligned} \sum_{t=1}^n \left(\overline{F}^*_t(-\hat{L}^\rho_t) - \overline{F}^*_t(-\hat{L}^\rho_{t+1}) - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right) &\leq \sum_{t=1}^n F_{t-1}(\tilde{x}_t) - F_t(\tilde{x}_t) \\ &\leq \sum_{t=1}^n \max_{x\in\Delta([k])} F_{t-1}(x) - F_t(x) \\ &= -F_n(\mathbf{1}_k/k) = 2\sqrt{kn} + \eta_n^{-1}\log(k)\,. \end{aligned}$$

□

Proof of Lemma 5.3. First we prove that the term is upper bounded by 1. We have

$$\begin{aligned} &\overline{F}^*_t(-\hat{L}^{obs}_t) - \overline{F}^*_t(-\hat{L}^{obs}_t - \hat{\ell}_t) - \overline{F}^*_t(-\hat{L}^\rho_t) + \overline{F}^*_t(-\hat{L}^\rho_{t+1}) \\ &\qquad = -D_{\overline{F}^*_t}(-\hat{L}^{obs}_t - \hat{\ell}_t, -\hat{L}^{obs}_t) - D_{\overline{F}^*_t}(-\hat{L}^\rho_t, -\hat{L}^\rho_{t+1}) + \langle x_t - \nabla\overline{F}^*(-\hat{L}^\rho_{t+1}), \hat{\ell}_t\rangle \leq 1\,. \end{aligned}$$

For the second part, we define $\hat{L}^{miss}_t = \hat{L}^\rho_t - \hat{L}^{obs}_t$. Then we have

$$\begin{aligned} -\overline{F}^*_t(-\hat{L}^\rho_t) + \overline{F}^*_t(-\hat{L}^\rho_{t+1}) &\overset{(a)}{=} -\int_0^1 \langle \hat{\ell}_t, \nabla\overline{F}^*_t(-\hat{L}^\rho_t - x\hat{\ell}_t)\rangle\,dx \\ &= -\int_0^1 \langle \hat{\ell}_t, \nabla\overline{F}^*_t(-\hat{L}^{obs}_t - \hat{L}^{miss}_t - x\hat{\ell}_t)\rangle\,dx \\ &\overset{(b)}{\leq} -\int_0^1 \langle \hat{\ell}_t, \nabla\overline{F}^*_t(-\hat{L}^{obs}_t - \hat{L}^{miss}_{t,A_t}\mathbf{e}_{A_t} - x\hat{\ell}_t)\rangle\,dx\,. \end{aligned}$$

(a) Uses the fundamental theorem of calculus. (b) Follows from the fact that $\nabla\overline{F}^*_t(-L)_{A_t}$ decreases if the loss in coordinates other than A_t is reduced. Therefore, we have

$$\begin{aligned} &\overline{F}^*_t(-\hat{L}^{obs}_t) - \overline{F}^*_t(-\hat{L}^{obs}_t - \hat{\ell}_t) - \overline{F}^*_t(-\hat{L}^\rho_t) + \overline{F}^*_t(-\hat{L}^\rho_{t+1}) \\ &\overset{(c)}{\leq} \int_0^1 \langle \hat{\ell}_t, \nabla\overline{F}^*_t(-\hat{L}^{obs}_t - x\hat{\ell}_t)\rangle\,dx - \int_0^1 \langle \hat{\ell}_t, \nabla\overline{F}^*_t(-\hat{L}^{obs}_t - \hat{L}^{miss}_{t,A_t}\mathbf{e}_{A_t} - x\hat{\ell}_t)\rangle\,dx \\ &\overset{(d)}{=} \int_0^1 \langle \hat{\ell}_t, \tilde{z}(x) - \nabla\overline{F}^*_t(\nabla F_t(\tilde{z}(x)) - \hat{L}^{miss}_{t,A_t}\mathbf{e}_{A_t})\rangle\,dx \\ &\overset{(e)}{\leq} \int_0^1 \langle \hat{\ell}_t, \tilde{z}(x) - \nabla F^*_t(\nabla F_t(\tilde{z}(x)) - \hat{L}^{miss}_{t,A_t}\mathbf{e}_{A_t})\rangle\,dx \\ &= \int_0^1 \hat{\ell}_{t,A_t}(\tilde{z}_{A_t}(x) - {f^*}'_t(f'_t(\tilde{z}_{A_t}(x)) - \hat{L}^{miss}_{t,A_t})\,dx \\ &\overset{(f)}{\leq} \int_0^1 \hat{\ell}_{t,A_t}({f^*}''_t(f'_t(\tilde{z}_{A_t}(x)))\hat{L}^{miss}_{t,A_t}\,dx \\ &= \int_0^1 \ell_{t,A_t} x^{-1}_{t,A_t} f''_t(\tilde{z}_{A_t}(x))^{-1}\hat{L}^{miss}_{t,A_t}\,dx \\ &\overset{(g)}{\leq} \int_0^1 \ell_{t,A_t} x^{-1}_{t,A_t} f''_t(x_{t,A_t})^{-1}\hat{L}^{miss}_{t,A_t}\,dx \\ &\leq x^{-1}_{A_t} f''_{t2}(x_{A_t})^{-1}\hat{L}^{miss}_{t,A_t} \\ &= \eta_t\hat{L}^{miss}_{t,A_t}\,. \end{aligned}$$

(c) uses the Fundamental theorem of calculus together with the inequality above. (d) substitutes $\tilde{z}(x) = \nabla \overline{F}_t^*(-\hat{L}_t^{obs} - x\hat{\ell}_t)$ and applies fact 5.3. (e) applies fact 5.5. (f) $f^{*\prime}(t)$ is convex, so $-f^{*\prime}(f'(\tilde{z}_{A_t}) - \ell) \leq -\tilde{z}_{A_t} + f^{*\prime\prime}(f'(\tilde{z}_{A_t}))$. (g) follows because $\tilde{z}_{A_t} \leq x_{t,A_t}$ and $f_t''(x)^{-1}$ is monotonically increasing. Finally, due to the unbiasedness of the loss estimators we have in expectation

$$\mathbb{E}[\hat{L}_{t,A_t}^{miss}] = \sum_{s:\rho(s)<\rho(t)} \mathbb{I}\{s + d_s \geq t\}\ell_{s,A_t} \leq \mathfrak{d}_t^{\rho}\,.$$

□

Chapter 6

Connections to Bayesian Bandits

The work presented in this chapter is based on a paper that has been accepted as [108].

[108] Zimmert, J. and Lattimore, T. (2019). Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems (NeurIPS)*

Abstract

The information-theoretic analysis by Russo and Van Roy [86] in combination with minimax duality has proved a powerful tool for the analysis of online learning algorithms in full and partial information settings. In most applications there is a tantalising similarity to the classical analysis based on mirror descent. We make a formal connection, showing that the information-theoretic bounds in most applications can be derived from existing techniques for online convex optimisation. Besides this, for k-armed adversarial bandits we provide an efficient algorithm with regret that matches the best information-theoretic upper bound and improve best known regret guarantees for online linear optimisation on ℓ_p-balls and bandits with graph feedback.

6.1 Introduction

The combination of minimax duality and the information-theoretic machinery by Russo and Van Roy [86] has proved a powerful tool in the analysis of online learning algorithms. This has led to short and insightful analysis for k-armed bandits, linear bandits, convex bandits and partial monitoring, all improving on prior best known results. The downside is that the approach is non-constructive. The application of minimax duality demonstrates the existence of an algorithm with a given bound in the adversarial setting, but provides no way of constructing that algorithm.

The fundamental quantity in the information-theoretic analysis is the 'information ratio' in round t, which informally is

$$\text{information ratio}_t = \frac{(\text{expected regret in round } t)^2}{\text{expected information gain in round } t}\,,$$

where the information gain is either measured using the mutual information [86] or a generalisation based on a Bregman divergence [72]. Proving the information ratio is small corresponds to showing that either the learner is suffering small regret in round t or gaining information, which ultimately leads to a bound on the cumulative regret. The aforementioned generalisation by Lattimore and Szepesvári [72] (restated in the supplementary) lead to a short analysis for k-armed adversarial bandits that is minimax optimal except for small constant factors. The authors speculated that the new idea should lead to improved bounds for a range of online learning problems and suggested a number of applications, including bandits with graph feedback [9] and linear bandits on ℓ_p-balls [27].

We started to follow this plan, successfully improving existing minimax bounds for bandits with graph feedback and online linear optimisation for ℓ_p-balls with full information (the bandit setting remains a mystery). Along the way, however, we noticed a striking connection between the analysis techniques for bounding the information ratio and controlling the stability of online stochastic mirror descent (OSMD), which is a classical algorithm for online convex optimisation. A connection was already hypothesised by Lattimore and Szepesvári [72], who noticed a similarity between the bounds obtained. Notably, why does using the negentropy potential in the information-theoretic analysis lead to almost identical bounds for k-armed bandits as Exp3? Why does this continue to hold with the Tsallis entropy and the INF strategy [14]?

Contribution Our main contribution is a formal connection between the information-theoretic analysis and OSMD. Specifically, we show how tools for analysing OSMD can be applied to a modified version of Thompson sampling that uses the same sampling strategy as OSMD, but replaces the mirror descent update with a Bayesian update. This contribution is valuable for several reasons: (a) it explains the similarity between the information-theoretic and OSMD style analysis, (b) it allows for the transfer of techniques for OSMD to Bayesian regret analysis and (c) it opens the possibility of a constructive transfer of ideas from Bayesian regret analysis to the adversarial framework, as we illustrate in the next contribution.

A curiosity in the Bayesian analysis of adversarial k-armed bandits is that the resulting bound was always a factor of 2 smaller than the corresponding bound for OSMD. This was true in the original analysis [86] and its generalisation [72]. Our new theorem entirely explains the difference, and indeed, allows us to improve the bounds for OSMD. This leads to an efficient algorithm for adversarial k-armed bandits with regret $\mathfrak{R}_n \leq \sqrt{2kn} + O(k)$, matching the information-theoretic upper bound except for small lower-order terms.

Finally, we improve the regret guarantees for two online learning problems. First, for bandits with graph feedback we improve the minimax regret in the 'easy' setting by a $\log(n)$ factor, matching the lower bound up to a factor of $\log^{3/2}(k)$. Second, for online linear optimisation over the ℓ_p-balls we improve existing bounds by arbitrarily large constant factors. At first we had proved these results using the information-theoretic tools and minimax duality, but here we present the unified view and consequentially the analysis also applies to OSMD for which we have efficient algorithms.

Related work The information-theoretic Bayesian regret analysis was introduced by [85, 86, 87]. The focus in these papers is on the analysis of Bayesian algorithms in the stochastic setting, a line of work continued recently by [46]. [28] noticed that the stochastic assumption is not required and that the results continued to hold in a Bayesian adversarial setting where the prior is over arbitrary sequences of losses, rather than over (parametric) distributions as is usual in Bayesian statistics. The idea to use minimax duality to derive minimax regret bounds is due to [3] and has been applied and generalised by a number of authors [28, 30, 54, 72]. Mirror descent was developed by [77] and [78] for optimisation. As far as we know its first application to bandits was by [4], which precipitated a flood of papers as summarised in the books by [26, 70]. We work in the partial monitoring framework, which goes back to [88]. Most of the focus since then has been on classifying the growth of the regret on the horizon for finite partial monitoring games [12, 21, 37, 48, 71]. Bandits with graph feedback are a special kind of partial monitoring problem and have been studied extensively [9, 10, 41, and others], with a monograph on the subject by [101]. Online linear optimisation is an enormous subject by itself. We refer the reader to the books by [35, 56].

Notation The reader will find omitted proofs in the appendix. Let $[n] = \{1, 2, \ldots, n\}$ and $B_p^d = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$ be the standard ℓ_p-ball. For positive definite A we write $\|x\|_A^2 = x^\top Ax$. Given a topological space X, let $\operatorname{int}(X)$ be its interior and $\Delta(X)$ be the space of probability measures on X with the Borel σ-algebra. We write $X^\circ = \{y \in \mathbb{R}^d : \sup_{x\in X} |\langle x, y\rangle| \leq 1\}$ for the functional analysts polar and $\operatorname{co}(X)$ for the convex hull of X. The domain of a convex function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is $\operatorname{dom}(F) = \{x : F(x) < \infty\}$. For $x, y \in \operatorname{dom}(F)$ the Bregman divergence between x and y with respect to F is $\mathrm{D}_F(x, y) = F(x) - F(y) - \nabla F_{x-y}(y)$ where $\nabla_v F(x)$ is the directional derivative of F at x in the direction v. The diameter of X with respect to F is $\operatorname{diam}_F(X) = \sup_{x,y\in X} F(x) - F(y)$. We abuse notation by writing $\nabla^{-2}F(x) = (\nabla^2 F(x))^{-1}$. For $x, y \in \mathbb{R}^d$ we let $[x, y] = \operatorname{co}(\{x, y\})$ be the convex hull of x and y, which is the set of points on the chord between x and y.

Linear partial monitoring Our results are most easily expressed in a linear version of the partial monitoring framework, which extends the standard adversarial linear bandit framework to general feedback structures. Let $\mathcal{A}$ be the action space and $\mathcal{L}$ the loss space, which are subsets of $\mathbb{R}^d$ with $\mathcal{A}$ compact. The convex hull of $\mathcal{A}$ is $\mathcal{X} = \operatorname{co}(\mathcal{A})$. When $\mathcal{A}$ is finite we let $k = |\mathcal{A}|$. The signal function is a known function $\Phi : \mathcal{A} \times \mathcal{L} \to \Sigma$ for some observation space Σ.

An adversary and learner interact over n rounds. First the adversary secretly chooses $(\ell_t)_{t=1}^n$ with $\ell_t \in \mathcal{L}$ for all t. In each round t the learner samples an action $A_t \in \mathcal{A}$ from a distribution depending on observations $A_1, \Phi_1, \dots, A_{t-1}, \Phi_{t-1}$ where $\Phi_s = \Phi(A_s, \ell_s)$ is the observation in round s. The regret of policy π in environment $(\ell_t)_{t=1}^n$ is

$$\mathfrak{R}_n(\pi, (\ell_t)_{t=1}^n) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a, \ell_t \rangle\right] ,$$

where the expectation is with respect to the randomness in the actions. The regret depends on a policy and the losses. The minimax regret is

$$\mathfrak{R}_n^* = \inf_{\pi} \sup_{(\ell_t)_{t=1}^n} \mathfrak{R}_n(\pi, (\ell_t)_{t=1}^n) ,$$

where the infimum is over all policies and the supremum over all loss sequences in $\mathcal{L}^n$. From here on the dependence of $\mathfrak{R}_n$ on the policy and loss sequence is omitted.

Examples The standard k-armed bandit is recovered when $\mathcal{A} = \{e_1, \dots, e_k\}$, $\mathcal{L} = [0,1]^k$ and $\Phi(a, \ell) = \langle a, \ell \rangle \in \Sigma = [0, 1]$. For linear bandits the set $\mathcal{A}$ is an arbitrary compact set and $\mathcal{L}$ is typically $\mathcal{A}^\circ$. Bandits with graph feedback have a richer signal function as we explain in Section 6.4.

Bayesian setting In the Bayesian setting the sequence of losses $(\ell_t)_{t=1}^n$ are sampled from a known prior probability measure ν on $\mathcal{L}^n$ and subsequently the learner interacts with the sampled losses as normal. The optimal action is now a random variable $A^* = \arg\min_{a \in \mathcal{A}} \sum_{t=1}^n \langle a, \ell_t \rangle$ and the Bayesian regret is

$$\mathfrak{BR}_n = \mathbb{E}\left[\sum_{t=1}^n \langle A_t - A^*, \ell_t \rangle\right] .$$

Finally, define $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \,|\, \mathcal{F}_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\,\cdot \,|\, \mathcal{F}_t]$ with $\mathcal{F}_t = \sigma(A_1, \Phi_1, \dots, A_t, \Phi_t)$, $\Delta_t = \langle A_t - A^*, \ell_t \rangle$. A crucial piece of notation is $X_t = \mathbb{E}_{t-1}[A_t] \in \mathcal{X}$, which is the conditional expected action played in round t.

6.2 Mirror descent, Thompson sampling and the information ratio

We now develop the connection between OSMD and the information-theoretic Bayesian regret analysis. Specifically we show that instances of OSMD can be transformed into an algorithm similar to Thompson sampling (TS) for which the Bayesian regret can be bounded in the same way as the regret of the original algorithm. The similarity to TS is important. Any instance of OSMD with a uniform bound on the adversarial regret enjoys the same bound on the Bayesian regret for any prior without modification. Our result has a different flavour because we prove a bound for a variant of OSMD that replaces the mirror descent update with a Bayesian update.

```
Algorithm 9: OSMD
  Input: $\mathscr{A} = (P, E, F)$ and $\eta$
  Initialize $X_1 = \arg\min_{a \in \mathcal{X}} F(a)$
  for $t = 1, \dots, n$ do
      Sample $A_t \sim P_{X_t}$ and observe $\Phi_t$
      Construct: $\hat{\ell}_t = E(X_t, A_t, \Phi_t)$
      Update: $X_{t+1} = f_t(X_t, A_t)$
```

OSMD is a modular algorithm that depends on defining three components: (1) A sampling scheme that determines how the algorithm explores, (2) a method for estimating the unobserved loss vectors, and (3) a convex ‘potential’ and learning rate that determines how the algorithm updates its iterates. The following definition makes this more precise.

Definition 6.1. *An instance of OSMD is determined by a tuple $\mathscr{A} = (P, F, E)$ and learning rate $\eta > 0$ such that*

(a) The sampling scheme is a collection $P = \{P_x : x \in \mathcal{X}\}$ of probability measures in $\Delta(\mathcal{A})$ such that $\mathbb{E}_{A \sim P_x}[A] = x$ for all $x \in \mathcal{X}$.

(b) The potential is a Legendre function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with $\operatorname{dom}(F) \cap \mathcal{X} \neq \emptyset$ and $\eta > 0$ is the learning rate.

(c) The estimation function is $E : \mathcal{X} \times \mathcal{A} \times \Sigma \to \mathbb{R}^d$, which we assume satisfies $\mathbb{E}_{A \sim P_x}[E(x, A, \Phi(A, \ell))] = \ell$ for all $\ell \in \mathcal{L}$ and $x \in \mathcal{X}$.

The assumptions on the mean of P_x and that E is unbiased are often relaxed in minor ways, but for simplicity we maintain the strict definition. For the remainder we fix $\mathscr{A} = (P, F, E)$ and $\eta > 0$ and abbreviate

$$E_t(x, a) = E(x, a, \Phi(a, \ell_t)) \qquad \text{and} \qquad \hat{\ell}_t = E(X_t, A_t, \Phi_t)\,.$$

You should think of $E_t(x, a)$ as the estimated loss vector when the learner plays action a while sampling from P_x and $\hat{\ell}_t$ as the realisation of this estimate in round t. OSMD starts by initialising X_1 as the minimiser of F constrained to $\mathcal{X}$. Subsequently it samples $A_t \sim P_{X_t}$ and updates

$$X_{t+1} = \operatorname*{arg\,min}_{y \in \mathcal{X}} \eta\langle y, \hat{\ell}_t\rangle + \mathrm{D}_F(y, X_t)\,.$$

A useful notation is to let $(f_t)_{t=1}^n$ and $(g_t)_{t=1}^n$ be sequences of functions from $\mathcal{X} \times \mathcal{A}$ to $\mathbb{R}^d$ with

$$\begin{aligned} f_t(x, a) &= \operatorname*{arg\,min}_{y \in \mathcal{X}} \left(\eta\langle y, E_t(x, a)\rangle + \mathrm{D}_F(y, x)\right) \qquad \text{and} \\ g_t(x, a) &= \operatorname*{arg\,min}_{y \in \operatorname{int}(\operatorname{dom}(F))} \left(\eta\langle y, E_t(x, a)\rangle + \mathrm{D}_F(y, x)\right)\,, \end{aligned}$$

which means that $X_{t+1} = f_t(X_t, A_t)$, while g_t is the same as f_t, but without the constraint to $\mathcal{X}$. The complete algorithm is summarised in Algorithm 9. The next theorem is well known [70, §28].

Theorem 6.1 (OSMD REGRET BOUND)**.** *The regret of OSMD satisfies*

$$\mathfrak{R}_n \leq \frac{\operatorname{diam}_F(\mathcal{X})}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}_t(X_t; \eta)\right]\,,$$

where $\operatorname{stab}_t(x; \eta) = \dfrac{2}{\eta}\mathbb{E}_{A \sim P_x}\left[\langle x - f_t(x, A), E_t(x, A)\rangle - \dfrac{\mathrm{D}_F(f_t(x, A), x)}{\eta}\right].$

The random variable $\operatorname{stab}_t(X_t; \eta)$ measures the stability of the algorithm relative to the learning rate and is usually almost surely bounded. The diameter term depends on how fast the algorithm can move from the starting point to optimal, which is large when the learning rate is small. In this sense the learning rate is tuned to balance the stability of the algorithm and the requirement that (X_t) can tend towards an optimal point. Note that $\operatorname{stab}_t(x)$ depends on P, E, F, η and the loss vector ℓ_t, which means that in the Bayesian setting the stability function is random. The next lemma is also known and is often useful for bounding the stability function.

Lemma 6.1. *Suppose that F is twice differentiable on* $\operatorname{int}(\operatorname{dom}(F))$*, then*

$$\operatorname{stab}_t(x; \eta) \leq \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, f_t(x, A)]} \|E_t(x, A)\|^2_{\nabla^{-2}F(z)}\right]\,.$$

Furthermore, provided that $g_t(x, a)$ exists for all a in the support of P_x, then

$$\operatorname{stab}_t(x; \eta) \leq \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, g_t(x, A)]} \|E_t(x, A)\|^2_{\nabla^{-2}F(z)}\right]\,.$$

Bayesian analysis Modified Thompson sampling (MTS) is a variant of TS summarised in Algorithm 10 that depends on a prior distribution ν and a sampling scheme P. The algorithm differs from Algorithm 9 in the computation of X_t. Rather than using the mirror descent update, it uses the Bayesian expected optimal action conditioned on the observations. Expectations in this subsection are with respect to both the prior and the actions, which means that $(\ell_t)_{t=1}^n$ are randomly distributed according to ν and consequently the functions f_t, g_t and stab_t are random. Our main theorem is the following bound on the Bayesian regret of MTS.

Algorithm 10: MTS

```
Input: Prior $\nu$ and $P$
Initialize $X_1 = \mathbb{E}[A^*]$
for $t = 1, \dots, n$ do
    Sample $A_t \sim P_{X_t}$ and observe $\Phi_t$
    Update: $X_{t+1} = \mathbb{E}_{t-1}[A^*]$
```

Theorem 6.2. *MTS satisfies* $\mathfrak{BR}_n \leq \dfrac{\mathrm{diam}_F(\mathcal{X})}{\eta} + \dfrac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \mathrm{stab}_t(X_t;\eta)\right]$.

Remark 6.1. *The stability function depends on $\mathscr{A} = (P, F, E)$ and η while Algorithm 10 only uses P. In this sense Theorem 6.2 shows that MTS satisfies the given bound for all E, F and η. MTS is the same as TS when sampling from the posterior is the same as sampling from P_{X_t}. A fundamental case where this always holds is when $\mathcal{A} = \{e_1, \dots, e_d\}$ because each $x \in \mathcal{X}$ is uniquely represented as a linear combination of elements in $\mathcal{A}$ and hence P_x is unique.*

Proof of Theorem 6.2. Beginning with the definition of the per-step regret,

$$
\begin{aligned}
\mathbb{E}_{t-1}[\Delta_t] &= \langle X_t, \mathbb{E}_{t-1}[\ell_t]\rangle - \mathbb{E}_{t-1}[\langle A^*, \ell_t\rangle] \\
&= \langle X_t, \mathbb{E}_{t-1}[\hat{\ell}_t]\rangle - \mathbb{E}_{t-1}\left[\langle A^*, \hat{\ell}_t\rangle\right] && (6.1)\\
&= \langle X_t, \mathbb{E}_{t-1}[\hat{\ell}_t]\rangle - \mathbb{E}_{t-1}\left[\langle \mathbb{E}_{t-1}[A^* \mid A_t, \Phi_t], \hat{\ell}_t\rangle\right] && (6.2)\\
&= \mathbb{E}_{t-1}\left[\langle X_t - X_{t+1}, \hat{\ell}_t\rangle\right] && (6.3)\\
&\leq \mathbb{E}_{t-1}\left[\langle X_t - f_t(X_t, A_t), \hat{\ell}_t\rangle - \frac{1}{\eta}\mathrm{D}_F(f_t(X_t, A_t), X_t) + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t)\right] && (6.4)\\
&\leq \mathbb{E}_{t-1}\left[\frac{\eta}{2}\mathrm{stab}_t(X_t;\eta) + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t)\right]. && (6.5)
\end{aligned}
$$

Eq. (6.1) uses that the loss estimators are unbiased. Eq. (6.2) follows using the tower rule for conditional expectations and the fact that $\hat{\ell}_t$ is a measurable function of X_t, A_t and Φ_t so that

$$\mathbb{E}_{t-1}[\langle A^*, \hat{\ell}_t\rangle] = \mathbb{E}_{t-1}[\mathbb{E}_{t-1}[\langle A^*, \hat{\ell}_t\rangle \mid A_t, \Phi_t]] = \mathbb{E}_{t-1}[\langle \mathbb{E}_{t-1}[A^* \mid A_t, \Phi_t], \hat{\ell}_t\rangle] = \mathbb{E}_{t-1}[\langle X_{t+1}, \hat{\ell}_t\rangle].$$

Eq. (6.3) uses the definitions of X_{t+1}. Eq. (6.4) follows from the definition of f_t, which implies that

$$\langle f_t(X_t, A_t), \hat{\ell}_t\rangle + \frac{1}{\eta}\mathrm{D}_F(f_t(X_t, A_t), X_t) \leq \langle X_{t+1}, \hat{\ell}_t\rangle + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t).$$

Finally, Eq. (6.5) follows from the definition of stab_t. The proof is completed by summing over the per-step regret, noting that $(X_t)_{t=1}^n$ is a $(\mathcal{F}_t)_t$-adapted martingale and by [72, Theorem 3],

$$\mathbb{E}\left[\sum_{t=1}^n \mathrm{D}_F(X_{t+1}, X_t)\right] \leq \mathbb{E}[F(X_{n+1})] - F(X_1) \leq \mathrm{diam}_F(\mathcal{X}). \qquad \square$$

The stability coefficient The only difference between Theorems 6.1 and 6.2 is the trajectory of $(X_t)_{t=1}^n$ and the randomness of the stability function. In most analyses of OSMD the final bound is obtained via a uniform bound on $\operatorname{stab}_t(x;\eta)$ that holds regardless of the losses and in this case the trajectory X_t is irrelevant. This is formalised in the following definition and corollary. Define the stability coefficients by

$$\operatorname{stab}(\mathscr{A};\eta) = \sup_{x\in\mathcal{X}} \max_{t\in[n]} \operatorname{stab}_t(x;\eta) \qquad \text{and} \qquad \operatorname{stab}(\mathscr{A}) = \sup_{\eta>0} \operatorname{stab}(\mathscr{A};\eta)\,.$$

Corollary 6.1. *The regret of Algorithm 9 for an appropriately tuned learning rate is bounded by*

$$\mathfrak{R}_n \le \sqrt{2\operatorname{diam}_F(\mathcal{X})\operatorname{stab}(\mathscr{A})n}\,.$$

The Bayesian regret of Algorithm 10 is bounded by $\mathfrak{BR}_n \le \sqrt{2\operatorname{diam}_F(\mathcal{X})\operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A}))n}$*.*

The essential supremum is needed because the stability coefficient depends on the losses $(\ell_t)_{t=1}^n$, which are random in the Bayesian setting. Generally speaking, however, bounds on the stability coefficient are proven in a manner that is independent of the losses.

Remark 6.2. *Often* $\operatorname{stab}(\mathscr{A};\eta) \le a+b\eta$ *for constants* $a, b \ge 0$ *and* $\operatorname{stab}(\mathscr{A}) = \infty$*. Nevertheless, the same argument shows that the regret of Algorithm 9 is bounded by*

$$\mathfrak{R}_n \le \sqrt{2a\operatorname{diam}_F(\mathcal{X})n} + \frac{b\operatorname{diam}_F(\mathcal{X})}{a}\,,$$

and similarly for the Bayesian regret of Algorithm 10.

Stability and the information ratio The generalised information-theoretic analysis by [72] starts by assuming there exists a constant $\alpha > 0$ such that the following bound on the information ratio holds almost surely:

$$\text{information ratio}_t = \mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \le \alpha\,. \tag{6.6}$$

Then [72, Theorem 3] shows that

$$\mathfrak{BR}_n \le \sqrt{\alpha n \operatorname{diam}_F(\mathcal{X})}\,. \tag{6.7}$$

The proof of Theorem 6.2 directly provides a bound on the information ratio in terms of the stability coefficient. To see this, notice that Eq. (6.5) holds for all measurable η and let

$$\eta = \sqrt{2\mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)]/\operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A}))}\,. \tag{6.8}$$

Then by Eq. (6.5) and the definition of $\operatorname{stab}(\mathscr{A})$ it follows that

$$\mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \le 2\operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A})) \;\; a.s.\,.$$

In other words, the usual methods for bounding the stability coefficient in the analysis of OSMD can be used to bound the information ratio in the information-theoretic analysis.

Example 6.1. *To make the abstraction more concrete, consider the k-armed bandit problem where* $\mathcal{L} = [0,1]^k$ *and* $\mathcal{A} = \{e_1, \ldots, e_k\}$*. In this case there is a unique sampling scheme defined* *by $P_x(a) = \langle x, a\rangle$. The standard loss estimation function is to use importance-weighting, which leads to*

$$E_t(x,a)_i = \ell_{ti}\mathbb{1}(a = e_i)/x_i\,. \tag{6.9}$$

A commonly used potential is the unnormalised negentropy $F(x) = \sum_{i=1}^k x_i \log(x_i) - x_i$ that satisfies $\nabla^{-2}F(x) = \operatorname{diag}(x)$. The instance of OSMD resulting from these choices is called Exp3 for which an explicit form for X_t is well known:

$$X_{ti} = \exp\left(-\eta \textstyle\sum_{s=1}^{t-1} \hat\ell_{si}\right) \Big/ \left(\textstyle\sum_{j=1}^k \exp\left(-\eta \textstyle\sum_{s=1}^{t-1} \hat\ell_{sj}\right)\right)\,.$$

A short calculation shows that $g_t(x,a)_i = x_i \exp(-\eta\hat\ell_{ti}) \le x_i$. The stability function is bounded using the second part of Lemma 6.1 by

$$\begin{aligned}\operatorname{stab}_t(x;\eta) &\le \mathbb{E}_{A\sim P_x}\left[\sup_{z\in[x,g_t(x,A)]} \|E_t(x,A)\|^2_{\nabla^{-2}F(z)}\right] \\ &= \mathbb{E}_{A\sim P_x}\left[\sup_{z\in[x,g_t(x,A)]} \sum_{i=1}^k z_{ti}\frac{\mathbb{1}(A=e_i)\ell_{ti}^2}{x_{ti}^2}\right] = \mathbb{E}_{A\sim P_x}\left[\frac{\mathbb{1}(A=e_i)\ell_{ti}^2}{x_{ti}}\right] \le \sum_{i=1}^k \ell_{ti}^2 \le k\,.\end{aligned}$$

Finally, the diameter of the probability simplex $\mathcal{X}$ with respect to the unnormalised negentropy is $\operatorname{diam}_F(\mathcal{X}) = \log(k)$. Applying Theorem 6.1 shows that the regret of OSMD and Bayesian regret of MTS satisfy

$$\mathfrak{R}_n \le \sqrt{2nk\log(k)} \quad \textit{(OSMD)} \qquad \textit{and} \qquad \mathfrak{BR}_n \le \sqrt{2nk\log(k)} \quad \textit{(MTS)}\,.$$

Remark 6.3. *Theorems 6.1 and 6.2 are vacuous when $\operatorname{diam}_F(\mathcal{X}) = \infty$. The most straightforward resolution is to restrict X_t to a subset of $\mathcal{X}$ on which the diameter is bounded and then control the additive error. This idea also works in the Bayesian setting as described by [72]. We omit a detailed discussion to avoid technicalities.*

6.3 Bandits

The best known bound on the minimax regret for k-armed bandits is $\mathfrak{R}_n \le \sqrt{2kn}$ by [72]. They let $F(x) = -2\sum_{i=1}^k \sqrt{x_i}$ be the 1/2-Tsallis entropy and prove that

$$\mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \le \sqrt{k}\,.$$

By Cauchy-Schwarz $\operatorname{diam}_F(\mathcal{X}) \le 2\sqrt{k}$ and then Eq. (6.7) shows that $\mathfrak{BR}_n \le \sqrt{2nk}$ for all priors ν. Minimax duality is used to conclude that $\mathfrak{R}_n^* \le \sqrt{2kn}$. Meanwhile, using the importance-weighted estimator in Eq. (6.9) leads to a bound on the stability coefficient of $\operatorname{stab}(\mathscr{A}) \le 2\sqrt{k}$ and then Theorem 6.1 yields a bound of $\mathfrak{R}_n \le \sqrt{8nk}$. The discrepancy between these methods is entirely explained by the naive choice of importance-weighted estimator. The approach based on bounding the information ratio is effectively shifting the losses, which can be achieved in the OSMD framework by shifting the importance-weighted estimators (see Fig. 6.1). This idea reduces the worst-case variance of the importance weighted estimators by a factor of 4.

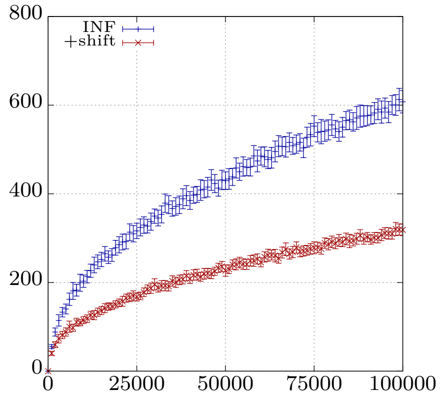


Figure 6.1: Comparison of INF with

Lemma 6.2. *If the loss estimator in Example 6.1 with $F(s) = -2\sum_{i=1}^k \sqrt{x_i}$ is replaced by*

$$E_t(x,a)_i = \frac{(\ell_{ti} - c_{ti})\mathbb{1}(a = e_i)}{x_i} + c_{ti}\,,$$
$$\text{where } c_{ti} = \frac{1}{2}(1 - \mathbb{1}(X_{ti} < \eta^2))\,,$$

then the stability coefficient for any $\eta \leq 1/2$ is bounded by $\operatorname{stab}(\mathscr{A};\eta) \leq k^{1/2}/2 + 12k\eta$.

Theorem 6.3. *The regret of OSMD with the loss estimator of Lemma 6.2 and appropriate learning rate satisfies: $\mathfrak{R}_n \leq \sqrt{2kn} + 48k$.*

6.4 Bandits with graph feedback

In bandits with graph feedback the action set is $\mathcal{A} = \{e_1, \ldots, e_k\}$ and $\mathcal{L} = [0,1]^k$. Let $E \subseteq [k] \times [k]$ be a set of directed edges over vertex set $[k]$ so that $\mathcal{G} = ([k], E)$ is a directed graph. The signal function is $\Phi(e_i, \ell) = \{(j, \ell_j) : j \in \mathcal{N}(i)\}$. The standard bandit framework is recovered when $E = \{(i,i) : i \in [k]\}$ while the full information setup corresponds to $E = [k] \times [k]$. Of course there are settings between and beyond these extremes. The difficulty of the graph feedback problem is determined by the connectivity of the graph. For example, when $E = \emptyset$, the learner has no way to estimate the losses and the regret is linear in the worst case. Like finite partial monitoring, graph feedback problems can be classified into one of four regimes for which:

$$\mathfrak{R}_n^* \in \left\{\mathcal{O}(1),\, \tilde{\Theta}(n^{1/2}),\, \Theta(n^{2/3}),\, \Omega(n)\right\}\,.$$

Our focus is on graph feedback problems that fit in the second category, which is the most challenging to analyse.

Definition 6.2. *$\mathcal{G}$ is called strongly observable if for every vertex $i \in [k]$ at least one of the following holds: (a) $a \in \mathcal{N}(b)$ for all $b \neq a$ or (b) $a \in \mathcal{N}(a)$.*

Alon et al. [9] prove the minimax regret for bandits with graph feedback is $\tilde{\Theta}(n^{1/2})$ if and only if $k > 1$ and $\mathcal{G}$ is strongly observable. They also prove the following theorem upper and lower bounding the dependence of the minimax regret on the horizon, the number of actions and a graph functional called the independence number.

Theorem 6.4 ([9])**.** *Let $\mathcal{G}_{ind}$ be the independence number of $\mathcal{G}$, which is the cardinality of the largest subset of vertices such that no tow distinct vertices are connected by an edge. Suppose $k > 1$ and $\mathcal{G}$ is strongly observable. Then $\mathfrak{R}_n^* = \mathcal{O}(\sqrt{\mathcal{G}_{ind} n}\log(kn))$ and $\mathfrak{R}_n^* = \Omega(\sqrt{\mathcal{G}_{ind} n})$.*

The logarithmic dependence on n in the proof of Theorem 6.4 appears quite naturally, which raises the question of whether or not the upper or lower bound is tight. In fact, as n tends to infinity the upper bound in Theorem 6.4 could be improved to $\mathcal{O}(\sqrt{nk})$ by using a finite-armed algorithm that ignores the feedback except for the played action. Perhaps the independence number is not as fundamental as first thought? The following theorem shows the upper bound can be improved.

Theorem 6.5. *Let $\mathscr{A} = (P, E, F)$ be a triple defining OSMD with $P_x(a) = \langle a, x \rangle$,*

$$F(x) = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^k x_i^\alpha \qquad \text{where} \quad \alpha = 1 - 1/\log(k)\,.$$

Finally, define the unbiased loss estimation function E by

$$E_t(x,a)_i = \frac{\ell_{ti}\mathbb{1}(a \in \mathcal{N}(i))}{\sum_{b \in \mathcal{N}(i)} x_b} \text{ for } i \notin I_t, \text{ and } E_t(x,a)_i = \frac{(\ell_{ti}-1)\mathbb{1}(a \neq i)}{1-x_i} + 1 \text{ otherwise},$$

where $I_t = \{i \in [k] : i \notin \mathcal{N}(i) \text{ and } X_{ti} > 1/2\}$. Then for any $k \geq 8$ and an appropriately tuned learning rate the regret of OSMD with $\mathscr{A}$ satisfies $\mathfrak{R}_n = \mathcal{O}(\sqrt{\mathcal{G}_{ind} n} \log(k)^3)$.

6.5 Online linear optimisation over ℓ_p-balls

We now consider full information online linear optimization on the ℓ_p balls with $p \in [1,2]$, which is modelled in our framework by choosing $\mathcal{A} = B_p^d$ and $\mathcal{L} = B_q^d$ with $1/p + 1/q = 1$ and $\Phi(a,\ell) = \ell$. Table 6.1 summarises the known results. When $p = 1$ the situation is unambiguous, with matching upper and lower bounds. For $p \in (1,2]$ there exist algorithms for which the regret is dimension free, but with constants that become arbitrarily large as p tends to 1. Known results for online gradient descent (OGD) prove the blowup in terms of p is avoidable, but with a price that is polynomial in the dimension.

p	Regret	Algorithm
$p = 1$	$\sqrt{n \log(d)}$	Hedge
$p > 1$	$\sqrt{n/(p-1)}$	[35, §11.5]
$p \geq 1$	$\sqrt{d^{2/p-1} n}$	OGD [56]

Table 6.1: Known results for ℓ_p-balls

Theorem 6.6. *For any $p \in [1,2]$, let h be the following convex and twice continuously differentiable function:*

$$h(x) = \begin{cases} \frac{d}{2}x^2 & \text{if } |x| \leq d^{\frac{1}{p-2}} \\ \frac{p-2}{p-1} d^{\frac{p-1}{p-2}} |x| + \frac{|x|^p}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} & \text{otherwise}. \end{cases}$$

Then for OSMD using potential $F(x) = \sum_{i=1}^d h(x_i)$, loss estimator $E(x,a,\sigma) = \sigma$, an arbitrary exploration scheme and appropriately tuned learning rate,

$$\mathfrak{R}_n = \mathcal{O}\left(\sqrt{\min\{1/(p-1), \log(d)\}\, n}\right).$$

Furthermore, the Bayesian regret of TS is bounded by the same quantity.

Remark 6.4. *In the full information setting the loss estimation is independent of the action, which explains the arbitrariness of the exploration scheme. The intuitive justification for the slightly cryptic potential function is provided in the appendix.*

6.6 Discussion

We demonstrated a connection between the information-theoretic analysis and OSMD. For k-armed bandits, we explained the factor of two difference between the regret analysis using information-theoretic and convex-analytic machinery and improved the bound for the latter. For graph bandits we improved the regret by a factor of $\log(n)$. Finally, we designed a new potential for which the regret for online linear optimisation over the ℓ_p-balls improves the previously best known bound by arbitrarily large constant factors.

Open problems The main open problem is whether or not we can 'close the circle' and use the information-theoretic analysis to directly construct OSMD algorithms. Another direction is to try and relax the assumption that the loss is linear. The leading constant in the new bandit analysis now matches the best known information-theoretic bound [72]. There is still a constant lower-order term, which presently seems challenging to eliminate. In bandits with graph feedback one can ask whether the $\log(k)$ dependency can be improved. Lower bounds are still needed for ℓ_p-balls and extending the idea to the bandit setting is an obvious followup. Finally, the best known algorithms for finite partial monitoring also use the information-theoretic machinery. Understanding how to borrow the ideas for OSMD remains a challenge.

Appendix

Theorem 3 of [72]

Theorem. *Let $(M_t)_{t=1}^{n+1}$ be an $\mathbb{R}^d$-valued martingale adapted to $(\mathcal{F}_t)_{t=1}^{n+1}$ and $M_t \in \mathcal{X} \subset \mathbb{R}^d$ almost surely for all t. Then let F be a convex function with $\operatorname{diam}_F(\mathcal{X}) < \infty$. Suppose there exist constants $\alpha, \beta \geq 0$ such that $\mathbb{E}_t[\Delta_t] \leq \alpha + \sqrt{\beta \mathbb{E}_t[D_F(M_{t+1}, M_t)]}$ almost surely for all t. Then $\mathfrak{BR}_n \leq \alpha n + \sqrt{n\beta \operatorname{diam}_F(\mathcal{X})}$.*

Proof of Lemma 6.1

The proof is rather standard. In fact, the first part is [70, Theorem 26.13]. For the second part, fix $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and define

$$\Psi(y) = \eta\langle y, E_t(x,a)\rangle + \mathrm{D}_F(y,x)\,.$$

By the assumption that $g_t(x,a) \in \operatorname{int}(\operatorname{dom}(F)) = \operatorname{int}(\operatorname{dom}(\Psi))$ and the definition of $g_t(x,a)$ as the minimizer of Ψ it follows that

$$0 = \nabla\Psi(g_t(x,a)) = \eta E_t(x,a) + \nabla F(g_t(x,a)) - \nabla F(x)\,.$$

Hence

$$\begin{aligned}
\operatorname{stab}_t(x) &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_t(x,A), E_t(x,A)\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\
&= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{1}{\eta}\langle x - f_t(x,A), \nabla F(x) - \nabla F(g_t(x,a))\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\
&= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{1}{\eta}\,\mathrm{D}_F(x, g_t(x,A)) - \frac{1}{\eta}\,\mathrm{D}_F(f_t(x,a), g_t(x,A))\right] \\
&\leq \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{\mathrm{D}_F(x, g_t(x,A))}{\eta}\right]\,. \qquad (6.10)
\end{aligned}$$

Let F^* be the Legendre dual of F. Since F is Legendre and twice differentiable on $\operatorname{int}(\operatorname{dom}(F))$ it follows from Taylor's theorem and duality that there exists a $z^* \in [\nabla F(x), \nabla F(x) - \eta E_t(x,a)]$ such that

$$\begin{aligned}
\mathrm{D}_F(x, g_t(x,a)) &= \mathrm{D}_{F^*}(\nabla F(g_t(x,a)), \nabla F(x)) \\
&= \mathrm{D}_{F^*}(\nabla F(x) - \eta E_t(x,a), \nabla F(x)) \\
&= \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^2 F^*(z^*)} \\
&= \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^{-2} F(\nabla F^*(z^*))} \\
&\leq \sup_{z\in[x, g_t(x,a)]} \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^{-2}F(z)}\,.
\end{aligned}$$

Substituting into Eq. (6.10) completes the result.

Refined bound for the probability simplex For the proofs in the next sections, we require a refined version of Lemma 6.1. Let 1_k denote the vector with all ones.

Lemma 6.3. *Assume that $\mathcal{A} = \{e_1, \dots, e_k\}$ and for $c \in \mathbb{R}$ define*

$$\begin{aligned} f_{tc}(x,a) &= \operatorname*{arg\,min}_{y \in \mathcal{X}} \left(\eta\langle y, E_t(x,a) + c1_k\rangle + \mathrm{D}_F(y,x)\right) , \\ g_{tc}(x,a) &= \operatorname*{arg\,min}_{y \in \operatorname{int}(\operatorname{dom}(F))} \left(\eta\langle y, E_t(x,a) + c1_k\rangle + \mathrm{D}_F(y,x)\right) . \end{aligned}$$

Provided that $g_{tc}(x,a)$ exists for all a in the support of P_x,

$$\operatorname{stab}_t(x;\eta) \le \frac{2}{\eta^2}\mathbb{E}_{A\sim P_x}\left[\mathrm{D}_F(x, g_{tc}(x,A))\right] \le \mathbb{E}_{A\sim P_x}\left[\sup_{z\in[x,g_{tc}(x,A)]} \|E_t(x,A) + c1_k\|^2_{\nabla^{-2}F(z)}\right] .$$

Proof. Since $\mathcal{X}$ is the probability simplex $\langle y, c1_k\rangle = c$ for all $y \in \mathcal{X}$. Therefore $f_{tc}(x,a) = f_t(x,a)$ and $\langle x - f_t(x,a), c1_k\rangle = 0$. Hence

$$\begin{aligned} \operatorname{stab}_t(x) &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_t(x,A), E_t(x,A)\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\ &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_{tc}(x,A), E_t(x,A) + c1_k\rangle - \frac{\mathrm{D}_F(f_{tc}(x,A),x)}{\eta}\right] . \end{aligned}$$

The remaining proof is analogous to the proof of Lemma 6.1 substituting f_t, g_t by f_{tc}, g_{tc} and the loss $E_t(x,a)$ by $E_t(x,a) + c1_k$. □

Proof of Corollary 6.1

Starting with the adversarial regret bound. By Theorem 6.1,

$$\mathfrak{R}_n \le \frac{\operatorname{diam}_F(\mathcal{X})}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}_t(X_t)\right] \le \frac{\operatorname{diam}_F(\mathcal{X})}{\eta} + \frac{\eta n \operatorname{stab}(\mathscr{A})}{2} .$$

The first part follows by choosing

$$\eta = \sqrt{\frac{2\operatorname{diam}_F(\mathcal{X})}{n \operatorname{stab}(\mathscr{A})}} .$$

The Bayesian case follows from an identical argument and Theorem 6.2 and the fact that

$$\mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}_t(X_t)\right] \le \mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}(\mathscr{A})\right] \le n \operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A})) .$$

The result claimed in Remark 6.2 follows similarly with the same choice of learning rate.

Proof of Theorem 6.3

Proof of Lemma 6.2. We use Lemma 6.3 with $c = -\frac{1}{2}$. As a reminder, we have

$$E_t(x,a)_i + c = \frac{(\ell_{ti} - c_{ti})\mathbb{1}(a = e_i)}{x_i} + c_{ti} + c\,, \text{ where } c_{ti} = \frac{1}{2}(1 - \mathbb{1}(X_{ti} < \eta^2).$$

Let $\tilde{\ell}_t = E_t(X_t, A_t) + c1_k$. We start by calculating the Hessian of F. Since $F(a) = -\sum_{i=1}^k 2\sqrt{a_i}$,

$$\nabla F(a) = -1/\sqrt{a} \qquad \text{and} \qquad \nabla^2 F(a) = \operatorname{diag}(a^{-3/2}/2) .$$

The next step is to bound $g_{tc}(X_t, A_t)_i^{\frac{3}{2}}$. By definition

$$g_{tc}(X_t, A_t) = \operatorname*{arg\,min}_{y \in \operatorname{int}(\operatorname{dom}(F))} \eta\langle y, \tilde{\ell}_t\rangle + F(y) - F(X_t) - \langle y - X_t, \nabla F(X_t)\rangle\,,$$

which implies that $\eta\tilde{\ell}_t + \nabla F(g_{tc}(X_t, A_t)) - \nabla F(X_t) = 0$. Substituting the gradient of the potential shows that

$$\eta\tilde{\ell}_{ti} - \frac{1}{\sqrt{g_{tc}(X_t, A_t)_i}} + \frac{1}{\sqrt{X_{ti}}} = 0\,.$$

Solving for $g_{tc}(X_t, A_t)_i$ yields

$$g_{tc}(X_t, A_t)_i^{\frac{3}{2}} = \frac{X_{ti}^{\frac{3}{2}}}{(1 + \tilde{\ell}_t\eta X_{ti}^{\frac{1}{2}})^3}\,. \tag{6.11}$$

For $\tilde{\ell}_{ti} \geq 0$, Eq. (6.11) directly implies $g_{tc}(X_t, A_t)_i^{\frac{3}{2}} \leq X_{ti}^{\frac{3}{2}}$. Let $\tilde{\ell}_{ti} < 0$, then we get the following lower bound by definition of $\tilde{\ell}_t$:

$$\begin{aligned}
X_{ti} \geq \eta^2 :\ \tilde{\ell}_{ti} &= -\frac{(\ell_{ti} - 1)\mathbb{1}(A_t = e_i)}{2X_{ti}} \geq -\frac{1}{2X_{ti}} \geq -\frac{1}{2\eta X_{ti}^{1/2}},\\
X_{ti} < \eta^2 :\ \tilde{\ell}_{ti} &= \frac{\ell_{ti}\mathbb{1}(A_t = e_i)}{X_{ti}} - \frac{1}{2} \geq -\frac{1}{2\eta X_{ti}^{1/2}} \geq -\frac{1}{2X_{ti}}.
\end{aligned}$$

This directly implies $-\tilde{\ell}_{ti}\eta X_{ti}^{1/2} \leq \frac{1}{2}\eta X_{ti}^{-1/2}$ and $1 + \tilde{\eta} X_{ti}^{1/2} \geq \frac{1}{2}$. Going back to Eq. (6.11), the following bound on $f(x) = x^{-3}$ holds due to convexity for all $x > -1$: $f(1+x) \leq f(1) + xf'(1+x)$. Using all three inequalities provides the bound

$$X_{ti}^{\frac{3}{2}}(1 + \tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}})^{-3} \leq X_{ti}^{\frac{3}{2}}\left(1 - 3(1 + \tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}})^{-4}\tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}}\right) \leq X_{ti}^{\frac{3}{2}} + 24\eta X_{ti}\,.$$

Hence for any $z \in [X_t, g_{tc}(X_t, A_t)]$ we have

$$\nabla^{-2}F(z) \preceq \operatorname{diag}(2X_t^{\frac{3}{2}} + 48\eta X_t \circ \mathbb{1}(\tilde{\ell}_t < 0))\,,$$

where $\mathbb{1}(\tilde{\ell}_t > 0)$ is vector of element wise applied indicator function. Finally we are ready to bound the stability:

$$\begin{aligned}
&\mathbb{E}_{A\sim P_{X_t}}\left[\sup_{z\in[X_t, g_{tc}(X_t, A)]} \|E_t(X_t, A) + c1_k\|^2_{\nabla^{-2}F(z)}\right]\\
&\leq \sum_{i:X_{ti}\geq\eta^2} X_{ti}\frac{(\ell_{ti} - \frac{1}{2})^2}{X_{ti}^2}(2X_{ti}^{\frac{3}{2}} + 48\eta X_{ti}) + \sum_{i:X_{ti}<\eta^2}\frac{1}{2^2}(2X_{ti}^{\frac{3}{2}} + 48\eta X_{ti}) + X_{ti}\frac{\ell_{ti}^2}{X_{ti}^2}2X_{ti}^{\frac{3}{2}} &(6.12)\\
&\leq \sum_{i:X_{ti}\geq\eta^2}\frac{X_{ti}^{\frac{1}{2}}}{2} + 12\eta + \sum_{i:X_{ti}<\eta^2}\frac{25\eta^3}{2} + 2\eta \leq \frac{\sqrt{k}}{2} + 12\eta k\,. &(6.13)
\end{aligned}$$

Eq. (6.12) follows because for $X_{ti} \geq \eta^2$ the term $E_t(X_t, A)_i + c$ is non zero with probability X_{ti}, while for $X_{ti} < \eta^2$, $E_t(X_t, A)_i + c$ is either non positive and bounded by $-\frac{1}{2}$, or it is positive with probability lower or equal to X_{ti}. Eq. (6.13) uses the condition $X_{ti} \leq \eta$ in the second sum and the upper bound $\eta \leq 1/2$. □

Proof of Theorem 6.3. Combine Lemma 6.2 with Theorem 6.1, Corollary 6.1, and Remark 6.2. □

Proof of Theorem 6.5

We make use of the following lemma.

Lemma 6.4 (Alon et al. 9)**.** *Let* $p \in \Delta([k])$*. Then*

$$\sum_{i=1}^{k} \frac{p_i}{\sum_{j\in\mathcal{N}(i)} p_j} \le 4\mathcal{G}_{ind} \log\left(\frac{4k}{\mathcal{G}_{ind}\min_i p_i}\right)\,.$$

Proof of Theorem 6.5. Starting from Corollary 6.1 we need to bound the diameter and stability.

$$\operatorname{diam}_F(\mathcal{X}) \le \frac{k^{1-\alpha}}{\alpha(1-\alpha)} = \frac{k^{\frac{1}{\log(k)}}\log(k)}{1-\frac{1}{\log(k)}} = \frac{e\log(k)}{1-\frac{1}{\log(k)}} \le 2e\log(k)\,,$$

where in the last inequality we used the assumption that $k \ge 8 > e^2$. Moving to the stability term. As a reminder we have

$$E_t(X_t, A_t)_i = \frac{\ell_{ti}\mathbb{1}(A_t \in \mathcal{N}(i))}{\sum_{b\in\mathcal{N}(i)} X_{tb}} \text{ for } i \in I_t \text{ and } E_t(X_t, A_t)_i = \frac{(\ell_{ti}-1)\mathbb{1}(A_t \neq i)}{1-X_{ti}} + 1 \text{ otherwise}$$

where $I_t = \{i \in [k] : i \notin \mathcal{N}(i) \text{ and } X_{ti} > 1/2\}$. The set I_t is either empty or contains exactly one element, since the action set it the probability simplex. As a slight abuse of notation, I_t denotes either the (possible empty) set or the unique element within. We use Lemma 6.3 with

$$c = \mathbb{1}(I_t \neq \emptyset)\frac{(1-\ell_{tI_t})\mathbb{1}(a \in \mathcal{N}(I_t))}{1-X_{tI_t}} \ge 0\,.$$

The Hessian of F is $\nabla F^2(x) = \operatorname{diag}(x^{\alpha-2})$. The non-negativity of $E_t(X_t, A_t) + c1_k$ ensures that $g_t(X_t, A_t)_i \le X_{ti}$ almost surely and hence by the definition of the potential $\nabla^{-2}F(z) \preceq \nabla^{-2}F(X_t)$ for all $z \in [X_t, g_t(X_t, A_t)]$,

$$\begin{aligned}
&\mathbb{E}_{A\sim P_{X_t}}\left[\sup_{z\in[X_t, g_{tc}(X_t,A)]} \|E_t(X_t,A) + c1_k\|^2_{\nabla^{-2}F(z)}\right] \\
&= \mathbb{E}_{A\sim P_{X_t}}\left[\|E_t(X_t,A) + c1_k\|^2_{\nabla^{-2}F(X_t)}\right] \\
&= \sum_{i\notin I_t} \mathbb{E}_{A\sim P_{X_t}}\left[(E_t(X_t,A)_i + c)^2 \nabla^{-2}F(X_t)_{ii}\right] + \mathbb{1}(I_t \neq \emptyset)\mathbb{E}_{A\sim P_{X_t}}[\nabla^{-2}F(X_t)_{I_tI_t}] \\
&\le 2\sum_{i\notin I_t} \mathbb{E}_{A\sim P_{X_t}}\left[E_t(X_t,A)_i^2 X_{ti}^{2-\alpha}\right] + 2\mathbb{E}_{A\sim P_{X_t}}[c^2]\sum_{i\notin I_t} X_{ti}^{2-\alpha} + 1\,.
\end{aligned}$$

We first bound the c term

$$2\mathbb{E}_{A\sim P_{X_t}}[c^2]\sum_{i\notin I_t} X_{ti}^{2-\alpha} = 2\mathbb{1}(I_t \neq \emptyset)\sum_{i\notin I_t} X_{ti}\left(\frac{1-\ell_{tI_t}}{\sum_{i\notin I_t} X_{ti}}\right)^2 \sum_{i\notin I_t} X_{ti}^{2-\alpha} \le 2.$$

Then we bound the contribution of arms i with $i \notin \mathcal{N}(i)$ and $i \notin I_t$, which implies $X_{ti} \le 1/2$

$$2\mathbb{E}_{A\sim P_x}\left[\sum_{i:i\notin\mathcal{N}(i)\cup I_t} E_t(X_t,A)_i^2 X_{ti}^{2-\alpha}\right] = 2\sum_{i:i\notin\mathcal{N}(i)\cup I_t} \frac{\ell_{ti}^2 X_{ti}^{2-\alpha}}{1-X_{ti}} \le 4\,.$$

Finally we bound the remaining term

$$2\mathbb{E}_{A\sim P_x}\left[\sum_{i:i\in\mathcal{N}(i)} E_t(X_t,A)_i^2 X_{ti}^{2-\alpha}\right] \le 2\sum_{i:i\in\mathcal{N}(i)} \frac{\ell_{ti}^2 X_{ti}^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)} X_{tj}} \le 2\max_{a\in\Delta([k])}\sum_{i=1}^{k} \frac{a_i^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)} a_j}.$$

We bound the max using Lemma 6.4:

$$\begin{aligned}\max_{a\in\Delta([k])}\sum_{i=1}^{k}\frac{a_i^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j} &= \max_{a\in\Delta([k])}\sum_{i:a_i>\exp(-\log(k)^2)}\frac{a_{ti}^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j}+\sum_{i:a_i\le\exp(-\log(k)^2)}\frac{a_i^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j}\\ &\le 4\mathcal{G}_{ind}\log\left(\frac{4k\exp(\log(k)^2)}{\mathcal{G}_{ind}}\right)+k\exp(-\log(k)^{-1}\log(k)^2)\\ &= 4\mathcal{G}_{ind}\left(\log\left(\frac{4k}{\mathcal{G}_{ind}}\right)+\log(k)^2\right)+1\,,\end{aligned}$$

where in the final inequality we used Lemma 6.4 on the sub-graph $\{a : X_{ta} > \exp(-\log(k)^2)$ and noted the fact the independence number of a sub-graph of $\mathcal{G}$ cannot be larger than the independence number of $\mathcal{G}$. Combining everything, we have shown that

$$\operatorname{stab}(\mathscr{A})\le 8\mathcal{G}_{ind}\left(\log\left(\frac{4k}{\mathcal{G}_{ind}}\right)+\log(k)^2\right)+9.$$

The proof is completed by tuning the learning rate according to Corollary 6.1. □

Proof of Theorem 6.6

Remember that the potential is $F(x)=\sum_{i=1}^{d}h(x_i)$ where

$$h(x)=\begin{cases}\frac{d}{2}x^2 & \text{if } |x|\le d^{\frac{1}{p-2}}\\ \frac{p-2}{p-1}d^{\frac{p-1}{p-2}}|x|+\frac{|x|^p}{p(p-1)}+\frac{2-p}{2p}d^{\frac{p}{p-2}} & \text{otherwise}\,.\end{cases}$$

Before the proof we provide some intuition for this choice of the potential. By the problem setting for $q=\frac{p}{1-p}$, it holds that $\|\ell_t\|_q, \|X_t\|_p\le 1$. Assuming we have a 'separable' potential $F(x)=\sum_{i=1}^{d}\tilde{h}(x_i)$, we can write the stability term as

$$\|\ell_t\|^2_{\nabla^{-2}F(z)}=\langle\ell_t\circ\ell_t,(\tilde{h}''(z_i)^{-1})_{i=1,\dots,d}\rangle\le\|\ell_t\circ\ell_t\|_{q'}\|(\tilde{h}''(z_i)^{-1})_{i=1,\dots,d}\|_{p'}\,.$$

Choosing $q'=\frac{q}{2}, p'=\frac{q'}{q'-1}=\frac{p}{2-p}$, the first factor is bounded by 1 and setting $\tilde{h}''(z_i)=|z_i|^{p-2}$ ensures the second factor is bounded by 1. Unfortunately, this leads to the potential $\tilde{h}(x)=\frac{1}{p(p-1)}|x|^p$, whose diameter can be arbitrarily large. To prevent the potential from exploding, we clip $h''(x)$ at d, as shown in Fig. 6.2. Any upper bound on the second derivative will serve the purpose of decreasing the diameter, however the threshold must be chosen such that the stability doesn't suffer too much. The value d happens to be the lowest value that keeps the stability dimension independent.

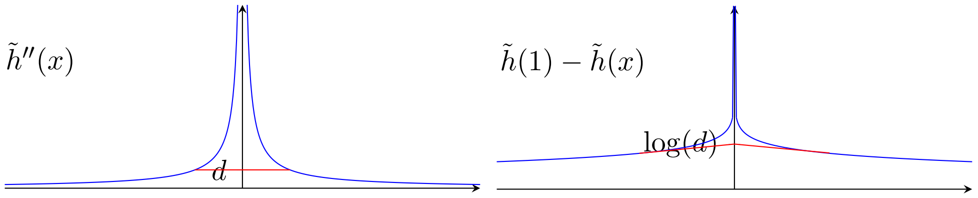


Figure 6.2: $p=1$: $\tilde{h}''(x)$ and $\tilde{h}(1)-\tilde{h}(x)$ for $p=1$. Red lines indicate h'' and h respectively.

Proof of Theorem 6.6. By the definition of the loss estimator $\hat{\ell}_t=\ell_t$. As usual, our plan is to bound the stability and diameter and then apply Corollary 6.1.

Bounding the stability By definition $h''(x) = \min\{|x|^{p-2}, d\}$. Then by Lemma 6.1 and the assumption that $E_t(x, a) = \ell_t$ for all x and a,

$$\begin{aligned}
\mathrm{stab}_t(x; \eta) &\leq \max_{z \in \mathcal{X}} ||\ell_t||^2_{\nabla F^{-2}(z)} \\
&\leq \max_{z \in \mathcal{X}} \left(\sum_{i : |z_i| \geq d^{\frac{1}{p-2}}} \ell_{ti}^2 |z_i|^{2-p} + \sum_{i : |z_i| < d^{\frac{1}{p-2}}} \frac{1}{d} \right) \\
&\leq \max_{z \in \mathcal{X}} \left(\sum_{i=1}^{d} \ell_{ti}^2 |z_i|^{2-p} + 1 \right) \\
&\leq \max_{z \in \mathcal{X}} \left(\left(\sum_{i=1}^{d} (\ell_{ti}^2)^{\frac{p}{2p-2}} \right)^{\frac{2p-2}{p}} \left(\sum_{i=1}^{d} (|z_i|^{2-p})^{\frac{p}{2-p}} \right)^{\frac{2-p}{p}} + 1 \right) \qquad (6.14) \\
&= \left(\max_{z \in \mathcal{X}} \|\ell_t\|_q^2 \|z\|_p^{2-p} + 1 \right) \leq 2 \,,
\end{aligned}$$

where Eq. (6.14) follows from Cauchy-Schwarz.

Bounding the diameter First notice that $F(x) \geq 0$ for all $x \in \mathcal{X}$ and $F(0) = 0$. Hence

$$\mathrm{diam}_F(\mathcal{X}) = \max_{x \in \mathcal{X}} F(x) \,.$$

For arbitrary $x \in \mathcal{X}$ define $J = \{i \in [d] | x_i \geq d^{\frac{1}{p-2}}\}$, $I = [d] \setminus J$ and for any $S \subset [d]$ define the vector x_S as the $|S|$-dimensional vector consisting of entries $(x_i)_{i \in S}$. Then it holds

$$F(x) = \frac{d}{2}\|x_I\|_2^2 - \frac{2-p}{p-1} d^{\frac{p-1}{p-2}} \|x_J\|_1 + \frac{\|x_J\|_p^p}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} |J|.$$

Maximizing this expression over x_J under the constraints of keeping both the set J and $\|x_J\|_p$ constant is setting all but 1 coordinate in x_J to $d^{\frac{1}{p-2}}$ and shifting all other weight towards a single entry. This follows directly from the fact that $\|x\|_p$ is convex, so the minimum of $\|x\|_1$ under constant $\|x\|_p$ is on the boundary. The optimal $y \in \arg\max_{x \in \mathcal{X}} F(x)$ can therefore only have a single coordinate i such that $|y_i| > d^{\frac{1}{p-2}}$, which we assume without loss of generality is $i = 1$.

$$F(y) = h(y_1) + \frac{d}{2} \sum_{i=2}^{d} y_i^2 \leq h(y_1) + \frac{d^2}{2} d^{\frac{2}{p-2}} \leq h(1) + \frac{1}{2} \,.$$

It follows that

$$\begin{aligned}
\mathrm{diam}_F(\mathcal{X}) \leq h(1) + \frac{1}{2} &= \frac{p-2}{p-1} d^{\frac{p-1}{p-2}} + \frac{1}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} + \frac{1}{2} \\
&= \frac{1 - d^{\frac{p-1}{p-2}}}{p-1} + d^{\frac{p-1}{p-2}} - \frac{1}{p} + \frac{2-p}{2p} d^{\frac{p}{p-2}} + \frac{1}{2} \leq \frac{1 - d^{\frac{p-1}{p-2}}}{p-1} + 1.
\end{aligned}$$

We immediately get the bound $\mathrm{diam}_T(\mathcal{X}) \leq \frac{2}{p-1}$. Let $p \leq \frac{3}{2}$, we substitute $z = \frac{p-1}{2-p}$ and get

$$\mathrm{diam}_F(\mathcal{X}) \leq \frac{1 - d^{-z}}{(2-p)z} + 1 \leq 2\frac{1 - d^{-z}}{z} + 1 \leq 2\log(d) + 1,$$

where we use the fact that for $z \geq 0$ the term $\frac{1-d^{-z}}{z}$ is monotonically decreasing in z with limit $\log(d)$ for $z \to 0$.

We have shown that $\mathrm{diam}_F(\mathcal{X}) \leq \mathcal{O}(\min\{\frac{1}{p-1}, \log(d)\})$ and $\mathrm{stab}(\mathscr{A}) \leq \mathcal{O}(1)$. The proof is completed by tuning the learning rate according to Corollary 6.1. □

Chapter 7

Summary and Discussion

In Chapter 2, we introduced the factored bandits model, a framework that generalises multiple problems from the bandit literature such as rank-1 bandits and utility-based dueling bandits. We presented an algorithm that is computationally and stochastically efficient and provided matching upper and lower bounds for the problem up to constant factors. The problem served as a motivating example for Chapters 3 and 4 since a crucial difficulty lies in obtain unbiased estimates of the relative quality of arms under non-stationary means.

In Chapter 3, we solved a problem that remained open for almost a decade. We provided an algorithm that enjoys optimal regret guarantees in both the stochastic and the adversarial environment while being oblivious to the environment at hand. We introduced a novel proof technique based on the self-bounding property of the regret, circumventing the need of controlling the variance of loss estimates. We also provided empirical evidence that our algorithm outperforms Ucb1 in stochastic environments and is significantly more robust than Ucb1 and Thompson Sampling in non-i.i.d. settings. Finally, we showed that our results extend to two intermediate settings from prior literature, namely stochastically constrained adversaries and adversarially corrupted stochastic bandits, and a combination of the two. We also showed that Tsallis-Inf can be applied to achieve stochastic and adversarial optimality in utility-based dueling bandits.

In Chapter 4, we extended our best-of-both-worlds results to combinatorial semi-bandits, via an Ftrl-based algorithm with a novel hybrid regulariser. Our bounds are worst-case optimal and also optimal in two particular instances of the problem.

In Chapter 5, we confirmed an open conjecture from Cesa-Bianchi et al. [34] by presenting a simple generalisation of our algorithm from Chapter 3 for adversarial bandits with arbitrary delays and proved a regret upper bound that matches the lower bound within constants. Furthermore, we proposed a refined tuning of the learning rate that achieves an even tighter regret bound when the delays are highly unbalanced. We are strictly improving on the state-of-the-art of bounds and also presenting the first anytime result requiring no doubling, skipping or advance information about the delays.

In Chapter 6, we demonstrated a connection between the information-theoretic analysis and Omd. For k-armed bandits, we explained the factor of two difference between the regret analysis using information-theoretic and convex-analytic machinery and improved the bound for the latter. For graph bandits, we improved the regret by a factor of $\log(n)$. Finally, we designed a new potential for which the regret for online linear optimisation over the ℓ_p-balls improves the previously best known bound by arbitrarily large constant factors.

7.1 Open questions

With regard to multi-armed bandits, the major open problems are as follows:

1. Our analysis requires a unique best arm. Based on our experiments, we conjecture that this is merely an artifact of the analysis, but extending the proof remains an important open problem.

2. Can the factor 2 gap between the asymptotical upper and lower bound be closed? It is unclear if that is a universal trade-off for obtaining best-of-both-worlds guarantees, if the algorithm can be further improved or if the lower bound is not tight for the stochastically constrained adversarial setting.

3. Is logarithmic regret achievable by Tsallis-Inf in the intermediate regimes defined by Seldin and Slivkins [91]? We presented this question in more detail in Section 3.5.2.

4. Can we obtain refined KL bounds, or at least second order approximations with a modification of Tsallis-Inf?

If we move beyond multi-armed bandits, the main question is whether any kind of best-of-both-worlds result is achievable in the linear bandit problem. We obtained preliminary results in Chapter 4, but it is unclear how the algorithm can be adapted for non-trivial action sets.

For bandits with delay, we conjecture that our algorithm obtains logarithmic regret bounds in the stochastic setting, but verifying this conjecture remains open. Furthermore, due to the lack of a matching lower bound, it is unclear if our refined upper bound for unbalanced delays is tight.

Lastly, regarding the connection between Omd, Thompson Sampling and the Information Ratio, the main open problem is whether or not we can ‘close the circle’ and use the information-theoretic analysis to directly construct Omd algorithms.

List of Publications

1. Zimmert, J. and Seldin, Y. (2018). Factored bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2835–2844
2. Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475
3. Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*
4. Zimmert, J. and Seldin, Y. (2020a). An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*
5. Zimmert, J. and Lattimore, T. (2019). Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems (NeurIPS)*

In preparation

6. Zimmert, J. and Seldin, Y. (2020b). Tsalls-INF: An optimal algorithm for stochastic and adversarial bandits. *arXiv preprint arXiv:1807.07623*

Bibliography

[1] Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., Malek, A., and Valko, M. (2018). Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the Conference on Learning Theory (COLT)*.

[2] Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2312–2320. Curran Associates, Inc.

[3] Abernethy, J., Agarwal, A., Bartlett, P. L., and Rakhlin, A. (2009). A stochastic view of optimal regret through minimax duality. In *Proceedings of the Conference on Learning Theory (COLT)*.

[4] Abernethy, J. D., Hazan, E., and Rakhlin, A. (2008). Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 263–274. Omnipress.

[5] Abernethy, J. D., Lee, C., Sinha, A., and Tewari, A. (2014). Online linear optimization via smoothing. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 35 of *Proceedings of Machine Learning Research*, pages 807–823, Barcelona, Spain. PMLR.

[6] Abernethy, J. D., Lee, C., and Tewari, A. (2015). Fighting bandits with a new kind of smoothness. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2197–2205. Curran Associates, Inc.

[7] Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. (2017). Corralling a band of bandit algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*.

[8] Ailon, N., Karnin, Z., and Joachims, T. (2014). Reducing dueling bandits to cardinal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, ICML'14, pages II–856–II–864. JMLR.org.

[9] Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35, Paris, France. PMLR.

[10] Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. (2017). Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826.

[11] Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976.

[12] Antos, A., Bartók, G., Pál, D., and Szepesvári, C. (2013). Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99.

[13] Audibert, J.-V. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836.

[14] Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 217–226.

[15] Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2013). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.

[16] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

[17] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

[18] Auer, P. and Chiang, C. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pages 116–120, Columbia University, New York, New York, USA. PMLR.

[19] Auer, P., Jaksch, T., and Ortner, R. (2009). Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 89–96.

[20] Bartók, G. (2013). A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 30, pages 696–710. PMLR.

[21] Bartók, G., Foster, D. P., Pál, D., Rakhlin, A., and Szepesvári, C. (2014). Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997.

[22] Bertsekas, D. P., Nedi, A., Ozdaglar, A. E., et al. (2003). *Convex analysis and optimization.* Athena Scientific.

[23] Besson, L. and Kaufmann, E. (2018). What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971.*

[24] Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. (2019). Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11345–11354.

[25] Bubeck, S. (2010). *Bandits games and clustering foundations.* PhD thesis, Université des Sciences et Technologie de Lille-Lille I.

[26] Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems.* Foundations and Trends in Machine Learning. Now Publishers Incorporated.

[27] Bubeck, S., Cohen, M. B., and Li, Y. (2018). Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*.

[28] Bubeck, S., Dekel, O., Koren, T., and Peres, Y. (2015). Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 266–278, Paris, France. PMLR.

[29] Bubeck, S., Perchet, V., and Rigollet, P. (2013). Bounded regret in stochastic multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 30 of *Proceedings of Machine Learning Research*, pages 122–134, Princeton, NJ, USA. PMLR.

[30] Bubeck, S. and Sellke, M. (2019). First-order regret analysis of Thompson sampling. Technical report, arXiv preprint arXiv:1902.00681.

[31] Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 42.1–42.23.

[32] Cappé, O., Garivier, A., Maillard, O., Munos, R., and Stoltz, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541.

[33] Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2018). Nonstochastic bandits with composite anonymous feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 750–773.

[34] Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. (2016). Delay and cooperation in nonstochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*.

[35] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[36] Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.

[37] Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. (2006). Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580.

[38] Chapelle, O., Manavoglu, E., and Rosales, R. (2014). Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–34.

[39] Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 151–159.

[40] Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33.

[41] Cohen, A., Hazan, T., and Koren, T. (2016). Online learning with feedback graphs without the graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 811–819.

[42] Combes, R., Magureanu, S., and Proutière, A. (2017). Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1761–1769.

[43] Combes, R., Shahi, M., Proutiere, A., and Lelarge, M. (2015). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2116–2124. Curran Associates, Inc.

[44] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edition.

[45] Dani, V., Kakade, S. M., and Hayes, T. P. (2008). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 345–352.

[46] Dong, S. and Van Roy, B. (2018). An information-theoretic analysis for thompson sampling with many actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4157–4165.

[47] Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 586–594. Curran Associates, Inc.

[48] Foster, D. and Rakhlin, A. (2012). No internal regret via neighborhood watch. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *Proceedings of Machine Learning Research*, pages 382–390, La Palma, Canary Islands. PMLR.

[49] Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. (2016). Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4734–4742.

[50] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1).

[51] Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478.

[52] Gaillard, P., Stoltz, G., and Van Erven, T. (2014). A second-order bound with excess losses. In *Proceedings of the Conference on Learning Theory (COLT)*.

[53] Gopalan, A., Mannor, S., and Mansour, Y. (2014). Thompson sampling for complex online problems. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[54] Gravin, N., Peres, Y., and Sivan, B. (2016). Towards optimal algorithms for prediction with expert advice. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on discrete algorithms*.

[55] Gupta, A., Koren, T., and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Conference on Learning Theory (COLT)*.

[56] Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

[57] Joulani, P., György, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *Proceedings of the International Conference on Machine Learning (ICML).*

[58] Joulani, P., Gyorgy, A., and Szepesvári, C. (2016). Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Thirtieth AAAI Conference on Artificial Intelligence.*

[59] Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017a). Bernoulli rank-1 bandits for click feedback. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence.*

[60] Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017b). Stochastic rank-1 bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS).*

[61] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pages 199–213. Springer.

[62] Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 1141–1154, Paris, France. PMLR.

[63] Koolen, W. M., Grünwald, P., and van Erven, T. (2016). Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems (NeurIPS).*

[64] Koolen, W. M., Warmuth, M. K., and Kivinen, J. (2010). Hedging structured concepts. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 93–105. Omnipress.

[65] Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. (2015a). Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA. PMLR.

[66] Kveton, B., Wen, Z., Ashkan, Z., and Szepesvári, C. (2015b). Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1450–1458. Curran Associates Inc.

[67] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

[68] Lattimore, T., Kveton, B., Li, S., and Szepesvári, C. (2018). Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3949–3958. Curran Associates, Inc.

[69] Lattimore, T. and Szepesvári, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In Singh, A. and Zhu, J., editors, *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737, Fort Lauderdale, FL, USA. PMLR.

[70] Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms.* Cambridge University Press (preprint).

[71] Lattimore, T. and Szepesvári, C. (2019a). Cleaning up the neighbourhood: A full classification for adversarial partial monitoring. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*.

[72] Lattimore, T. and Szepesvári, C. (2019b). An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the Conference on Learning Theory (COLT)*.

[73] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

[74] Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: Adanormalhedge. In *Proceedings of the Conference on Learning Theory (COLT)*.

[75] Luo, H., Wei, C.-Y., and Zheng, K. (2018). Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[76] Lykouris, T., Mirrokni, V., and Leme, R. P. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the Symposium on Theory of Computing (STOC)*.

[77] Nemirovsky, A. S. (1979). Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15.

[78] Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

[79] Neu, G. (2015). First-order regret bounds for combinatorial semi-bandits. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 1360–1375, Paris, France. PMLR.

[80] Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*.

[81] Orabona, F., Crammer, K., and Cesa-Bianchi, N. (2015). A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3).

[82] Pogodin, R. and Lattimore, T. (2019). On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

[83] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

[84] Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

[85] Russo, D. and Van Roy, B. (2014a). Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1583–1591. Curran Associates, Inc.

[86] Russo, D. and Van Roy, B. (2014b). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

[87] Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471.

[88] Rustichini, A. (1999). Minimizing regret: The general case. *Games and Economic Behavior*, 29(1):224–243.

[89] Seldin, Y., Bartlett, P. L., Crammer, K., and Abbasi-Yadkori, Y. (2014). Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML).*

[90] Seldin, Y. and Lugosi, G. (2017). An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pages 1743–1759, Amsterdam, Netherlands. PMLR.

[91] Seldin, Y. and Slivkins, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 1287–1295, Bejing, China. PMLR.

[92] Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2).

[93] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[94] Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

[95] Suehiro, D., Hatano, K., Kijima, S., Takimoto, E., and Nagano, K. (2012). Online prediction under submodular constraints. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pages 260–274. Springer.

[96] Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

[97] Thune, T. and Seldin, Y. (2018). Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2909–2918.

[98] Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. (2019). Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems (NeurIPS).*

[99] Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2).

[100] Urvoy, T., Clerot, F., Féraud, R., and Naamane, S. (2013). Generic exploration and k-armed voting bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 91–99.

[101] Valko, M. (2016). Bandits on graphs and structures.

[102] Warmuth, M. K. and Kuzmin, D. (2008). Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(Oct):2287–2320.

[103] Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT).*

[104] Wu, H. and Liu, X. (2016). Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 649–657. Curran Associates, Inc.

[105] Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2009). The k-armed dueling bandits problem. In *Proceedings of the Conference on Learning Theory (COLT)*.

[106] Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[107] Yue, Y. and Joachims, T. (2011). Beat the mean bandit. In *Proceedings of the International Conference on Machine Learning (ICML)*, ICML, pages 241–248, New York, NY, USA. ACM.

[108] Zimmert, J. and Lattimore, T. (2019). Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[109] Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[110] Zimmert, J. and Seldin, Y. (2018). Factored bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2835–2844.

[111] Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475.

[112] Zimmert, J. and Seldin, Y. (2020a). An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[113] Zimmert, J. and Seldin, Y. (2020b). Tsalls-INF: An optimal algorithm for stochastic and adversarial bandits. *arXiv preprint arXiv:1807.07623*.

[114] Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 10–18, Bejing, China. PMLR.