COMPETING AGAINST ADAPTIVE AGENTS BY MINIMIZING COUNTERFACTUAL NOTIONS OF REGRET

by
Teodor Vanislavov Marinov

A dissertation submitted to The Johns Hopkins University in conformity
with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
June, 2021

Abstract

Online learning or sequential decision making is formally defined as a repeated game between an adversary and a player. At every round of the game the player chooses an action from a fixed action set and the adversary reveals a reward/loss for the action played. The goal of the player is to maximize the cumulative reward of her actions. The rewards/losses could be sampled from an unknown distribution or other less restrictive assumptions can be made. The standard measure of performance is the cumulative ***regret***, that is the difference between the cumulative reward of the player and the best achievable reward by a fixed action, or more generally a fixed policy, on the observed reward sequence. For adversaries which are oblivious to the player's strategy, regret is a meaningful measure. However, the adversary is usually adaptive, e.g., in healthcare a patient will respond to given treatments, and for self-driving cars other traffic will react to the behavior of the autonomous agent. In such settings the notion of regret is hard to interpret as the best action in hindsight might not be the best action overall, given the behavior of the adversary. To resolve this problem a new notion called ***policy regret*** is introduced. Policy regret is fundamentally different from other forms of regret as it is counterfactual in nature, i.e., the player competes against all other policies whose reward is calculated by taking into account how the adversary would have behaved had the player chosen another policy. This thesis studies policy regret in a partial (bandit) feedback environment, beyond the worst case setting, by leveraging additional structure such as stochasticity/stability of the adversary or additional feedback.

Thesis Readers

Dr. Raman Arora (Primary Advisor)
Assistant Professor
Department of Computer Science
Johns Hopkins University

Dr. Michael Dinitz
Associate Professor
Department of Computer Science
Johns Hopkins University

Dr. Mehryar Mohri
Professor of Computer Science and Mathematics
Courant Institute of Mathematical Sciences
New York University and Google Research

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Raman Arora, for his guidance, unwavering support, and enthusiasm throughout my PhD. He gave me the opportunity to be a part of this world even though I had little research experience before I started this program. I am not only grateful for his ideas and discussions but also for teaching me how to best present my results and communicate with a broader audience. I would also like to thank him for introducing me to amazing researchers and opening doors for many fruitful collaborations.

Throughout this program, I have been extremely fortunate to work closely with Dr. Michael Dinitz and Dr. Mehryar Mohri. Their mentorship has been invaluable to my research career and has helped to shape and broaden my research interests. The quality of the problems studied in this thesis would not have been the same without their help. I would also like to thank Mehryar for giving me the opportunity to collaborate with the wonderful people at the Learning Theory group at Google Research, New York and for introducing me to many other researchers.

I would also like to thank all of my other co-authors and collaborators: Dr. Christoph Dann, Dr. Nikita Ivkin, Poorya Mianjy, Enayat Ullah, Dr. Jalaj Upadhyay, Yunjuan Wang, and Dr. Julian Zimmert. I have learned so much from our discussions and our collaborations have resulted in many great papers.

I have been very lucky to meet many of my closest friends during my stay at Johns Hopkins University. Their friendship, emotional support and all of our shared experiences truly made my PhD years fun and enjoyable. I would also like to thank all of my friends back home for still keeping in touch and sharing all the good times with me.

Finally, I would like to thank my family and especially my parents, Teresa and Vanislav, for all their sacrifices, love, kindness, patience, and guidance. Even though you might not understand my research completely, this thesis would not have been possible without you. Thank you!

Contents

Chapter 1

Introduction

Machine learning has exploded as a field in the past decade. Complex systems have been applied in impactful areas like healthcare, economically important tasks like market predictions, futuristic tasks like building self-driving cars and more mundane tasks like car navigation and ad placement/recommendations. While the practical success of such systems is undeniable, theoretical understanding of this success has been somewhat lacking. Indeed, classical theory of statistical learning, while well suited to problems in which there is an abundance of data all sampled from the same population, has struggled to explain the behavior of systems applied to any of the tasks above, in which data is plagued by adversarial corruptions, hidden confounding and its distribution may shift over time and adapt to the system's outputs. Moreover, in the above examples, successful systems are constantly evolving and learning is perpetual. The paradigm of online learning is arguably better suited to explain the performance of such systems.

1.1 Summary of contributions of this work

This thesis consists of four seemingly different problems spanning the topics of game theory, online learning and reinforcement learning. All the works, however, share a common theme, which is to study the online learning game in an adaptive setting.

Formally, we study ***policy regret minimization*** and its benefits in novel settings, beyond the worst case, provide meaningful regret guarantees for these settings and study how close to optimal these guarantees are.

We begin the study of policy regret in Chapter 2 by showing that it is indeed possible to improve on the bandit regret bound, if the player receives additional feedback about rewards of other actions together with the the reward of the played action. The feedback is modeled by a graph, where each vertex represents an action and playing an action also reveals the reward of all of its neighbors. While standard guarantees scale with the number of actions (K or $|\mathcal{A}|$), the novel bound only depends on the amount of information actions reveal about each other. In particular, if a few actions reveal information about all other actions, then the bound is constant in the number of actions. Formally the bound scales with the ***domination number*** of the feedback graph. While bounds scaling with domination number are known in the stochastic setting (Buccapatnam et al., 2014b), that is the losses are sampled from an unknown distribution, this is not the case in the adversarial setting. Most other work in the adversarial online learning with graph feedback literature scale with the ***independence number*** of the feedback graph (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014; Alon et al., 2015; Cohen et al., 2016; Valko, 2016; Lykouris et al., 2018; Lee et al., 2020b), which is always greater than the domination number. The upper bounds are supplemented by min-max lower bounds, showing that it is impossible to go beyond the $T^{2/3}$ regret bounds, unless all actions are observable at the same time. Further, we show a problem instance on which it is impossible to do better than the proposed regret upper bounds, including the dependence on domination number.

In many real world problems there is no single adversary but rather multiple self-interested agents competing in an environment, each with their own set of rewards. Our goal is to select the agent yielding the highest total reward. As a motivating example consider the online contractual display ads allocation problem: when users visit a website, say some page of the online site of a national newspaper, an ads allocation algorithm (the player) chooses an ad to display at each specific slot with the goal of achieving the largest value. To do so, the ads allocation algorithm chooses one out of a large set of advertisers (self-interested agents). Each advertiser has their own marketing strategy, usually following their own no-regret algorithm. The number of ads or arms can be very large. The number of advertisers can also be large in practice, depending on the domain. The number of times the ads allocation is run is in the order of millions or even billions per day, depending on the category of items. In the above example, at every round, the player will only be able to provide feedback to the agent whose ad was displayed. This in turn implies that only that agent will be able to update its internal state. Because the rewards we observe are dependent on the displayed ads and therefore dependent on the internal state of each agent, solving this ***corralling*** problem, requires counterfactual reasoning. Thus our goal is to minimize policy regret, which would guarantee that we would not discard the best agent should she perform poorly in initial rounds. The study of corralling adversarial bandit algorithms was initiated in Agarwal et al. (2016), who give sublinear regret guarantees. In Chapter 3, we extend their work by proposing a corralling strategy which also works for stochastic bandit algorithms and enjoys gap-dependent regret bounds, under the assumption that there is a positive gap between the reward of the best arm of the best base algorithm and the rewards of any other algorithm. These regret bounds are syntactically similar to instance dependent bounds discussed in Section 1.2.2.2. The algorithm is a version of FTRL with 1/2-Tsallis entropy regularization and a special step-size schedule inspired by Agarwal et al. (2016). As such it is able to corral bandit algorithms both when rewards are stochastic or adversarial. Finally, we supplement our theoretical results with experiments on a synthetic dataset, showing that our approach outperforms prior work.

In Chapter 4 we investigate policy regret in a game-theoretic setting. In this setting there is no central algorithm as in the corralling setting. Rather, each player's rewards are both a function of the environment and of other players actions. The goal of each player is again to maximise their own cumulative reward. Each player observes the reward of her actions at every round, unlike in the corralling setting. Two examples of games are network package routing and vehicle traffic routing. In both cases players are competing for shared resources and the utility of each player decreases as function of the number of other players using the same resource. Classical game theory studies how players can converge to equilibrium states in such games and what the social welfare is in such states. It is well known that if all parties play according to a no-regret rule, that is the regret of every player with respect to the observed rewards is $o(T)$, then their play would converge to a coarse correlated equilibrium (CCE) (see for example Nisan et al. (2007)). In general, different types of no-regret play have been shown to converge to different types of equilibria (Hazan and Kale, 2008). Further the social welfare of such equilibria has been investigated for certain classes of games (Roughgarden, 2015), which contain the above examples, showing that it is no worse than twice the best possible. Policy regret intuitively seems like a stronger notion than regret, so it is natural to ask what would happen in a game if players were to play strategies minimizing policy regret rather than simple regret. In (Arora et al., 2018), we show that in 2-player games, the players will converge to a new type of equilibrium called Policy Equillibrium (PE). Surprisingly, we also show that the class of PE contains the class of CCE as a strict subset and that policy regret is in fact not comparable to external regret. Finally, we show that as long as both agents play natural no-regret algorithms they will also achieve sublinear policy regret. This is in contrast with the adversarial setting, where different algorithms are needed for achieving strong policy regret guarantees.

In Chapter 5, we study the Reinforcement learning (RL) problem. The reinforcement learning problem is often modeled as a Markov Decision Process (MDP). The standard comparator in RL is the difference between the total reward of the player's policy and the best overall policy for the given MDP making policy regret the defacto measurement of success. There is a vast variety of settings under which RL problems are studied. Our work focuses on the episodic tabular setting, in which the interactions with the environment proceed in K episodes, each of length H. The MDP is assumed to have finite number of states and actions, however, no additional assumptions are made. Most prior work has focused on deriving regret bounds sub-linear in the number of interactions $T = HK$. Notable exceptions are the works of (Simchowitz and Jamieson, 2019; Lykouris et al., 2019; Jin and Luo, 2020), who all derive optimistic bounds, based on a definition of gap which captures the difference between total reward of the best policy and sub-optimal policies at every state-action pair. We investigate the smallest achievable optimistic regret by deriving information theoretic lower bounds for two classes of MDPs – MDPs with deterministic transitions and MDPs in which every state is visited by an optimal policy with non-zero probability. To derive such lower bounds we also need to assume that the value function of optimal policies and all reward functions are uniformly bounded by one. With the insight from these lower bounds we show that the regret of optimistic algorithms can be much smaller, depending on the structure of the MDP, than what was previously showed.

1.2 The online learning game

Online learning is concerned with studying the sequential decision making problem which is often modelled as a repeated game between an adversary or an unknown environment and a player, usually represented by an algorithm. While there are many ways to model this repeated game, we are going to focus on the following finite horizon scenario.

- The game will have $T \in \mathbb{N}$ rounds.
- During every round, $t \in [T]$[1], of the game the player must take an action a_t from a fixed action set $\mathcal{A}$.
- Further the adversary prepares a hidden loss ℓ_t (or reward r_t)[2] function which maps actions to real numbers.
- The player then observes the loss of her actions and possibly the loss over some or all remaining actions.

In general we will assume that the loss is bounded in $[0, 1]$, or if it is random, then has bounded sub-Gaussian norm. The goal of the player is to minimise her cumulative loss over the span of the game. The above game is very flexible and can represent multiple problems in Machine Learning, Statistics, Optimization and Game Theory. We now give two examples.

The experts problem: In the experts problem (Freund and Schapire, 1997) the set of actions, $\mathcal{A} = \{1, 2, \ldots, K\}$, consists of different expert advisors. At every round of the game, the player selects an expert $i_t \in \mathcal{A}$ and follows their advise. After selecting the expert, the player gets to observe the loss for the advise given by all the experts. The experts could be financial advisors managing the player's portfolio or in a gambling scenario could be the player's friends who bet on outcomes of sport games.

Convex Optimization: One could view the problem of optimizing a convex function, $f : \mathcal{A} \to \mathbb{R}$, through first order oracle queries as a version of the above game. At every round the player selects an iterate a_t as a guess about the optimum of the convex function and the adversary presents to the player the gradient (or more generally an element of the sub-differential) at a_t, given by $\nabla f(a_t)$. It is not

[1] $[T]$ denotes the set $\{1, \ldots, T\}$.

[2] Because our work deals with both stochastic and adversarial online learning we will use losses or rewards depending on what is more convenient for the discussion.

immediately clear why we are interested in minimizing the cumulative loss, however, using convexity it is possible to convert a regret bound on the sequence $(\nabla f(a_t))_{t=1}^T$ to a guarantee about approximately minimizing f. This is done using the so called online to batch conversion which uses $\widehat{a} = \frac{1}{T}\sum_{t=1}^T a_t$ as an approximate minimizer.

While in the convex optimization example above the set $\mathcal{A}$ consists of infinitely many actions, in our work we are primarily interested in sets of finite size $|\mathcal{A}| = K$.

1.2.1 Regret minimization

As already mentioned, the goal of the player is to minimize her cumulative loss. However, this is not sufficient to determine how well the player is doing after the T rounds have expired. For example, it is unreasonable to expect that the player will incur the smallest possible loss on every round of the game. Thus we would like to choose a suitable comparator against which the player's loss will be measured. In the experts problem described above one possible benchmark is to compare with the smallest cumulative loss obtained by an expert. In general we can compete with the smallest cumulative loss of any fixed action. The difference between the player's loss and the smallest cumulative loss is known as (external) ***regret*** (Foster and Vohra, 1993; Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi et al., 1997):

$$\mathscr{R}(T) = \sum_{t=1}^{T} \ell_t(a_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t(a). \tag{1.1}$$

In general, one is interested in algorithms which enjoy sublinear regret guarantees over all problem instances, that is for all adversaries the regret of the player's strategy is bounded by $o(T)$ with high probability.

There are several related quantities to 1.1. First, we can define the regret of the player with respect to a fixed action $a \in \mathcal{A}$ given as

$$\mathscr{R}_a(T) = \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(a). \tag{1.2}$$

Most algorithms in this work will have expected regret bounds of the form $\mathbb{E}[\mathcal{R}_a(T)] \leq o(T), \forall a \in \mathcal{A}$. Here the expectation is taken with respect to the possible randomization in the algorithm and any randomness coming from the strategy of the adversary. Note that the strategy of the adversary might depend on the player's strategy as well, which introduces additional randomness in the losses. One might be tempted to reason that if $\mathbb{E}[\mathcal{R}_a(T)] \leq o(T)$ for all actions then the expected regret, $\mathbb{E}[\mathcal{R}(T)]$, is also bounded. Unfortunately this is not always true, exactly because of the fact that the adversary might be ***adaptive***, that is they tailor their strategy based on what the player has done so far. Another quantity of interest which arises from bounds on $\mathbb{E}[\mathcal{R}_a(T)]$ is the ***pseudo regret***[3]:

$$\mathbb{E}[R(T)] = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] = \max_{a \in \mathcal{A}} \mathbb{E}[\mathscr{R}_a(T)]. \tag{1.3}$$

As already discussed, pseudo regret is a weaker notion than the expected regret and a simple application of Jensen's inequality shows that $\mathbb{E}[R(T)] \leq \mathbb{E}[\mathscr{R}(T)]$. In the case when the adversary is ***oblivious***, i.e., they have prepared the loss sequence $(\ell_t)_{t=1}^T$ before the start of the game then it holds that $\mathbb{E}[R(T)] = \mathbb{E}[\mathscr{R}(T)]$. One instance of oblivious adversaries is when each of the losses, $\ell_t(a)$, is sampled according to some unknown distribution.

1.2.1.1 Other notions of regret

External regret was introduced by comparing the cumulative loss of the player to the cumulative loss of the best fixed action in hindsight. Changing the comparator class gives rise to other different notions of regret. For example, one can compare against the best fixed function $\tau : \mathcal{A} \rightarrow \mathcal{A}$. The resulting regret is known as ***swap regret*** (Blum and Mansour, 2007). It turns out that one can build algorithms with at most $o(T)$ swap regret from algorithms with $o(T)$ external regret by an elegant reduction

[3] Most of this work focuses on bounding pseudo regret and hence we will refer to pseudo-regret simply as regret.

provided by Blum and Mansour (2007). One can take this notion of regret further by using functions as the comparator which might depend on the history of the player's actions (Lehrer, 2003). Mohri and Yang (2014) design efficient algorithms for the setting when $\tau : \mathcal{A}^s \to \mathcal{A}$ is a function defined for fixed s (independent of the time horizon T) and at time t exchanges the past s actions of the player for the comparator action $\tau(a_{t-s+1}, \ldots, a_t)$. The regret in this setting is called ***conditional swap regret***. Further, Mohri and Yang (2017) extend the comparator class to weighted finite-state automata and show how to minimize the respective ***transductive regret***.

If one takes $s = t$ at round t, that is to consider an unrestricted comparator function, they will arrive at what is known as ***dynamic regret*** (Zinkevich, 2003). Unsurprisingly, in the worst case it is impossible to design algorithms with $o(T)$ dynamic regret. However, it is possible to design strategies with dynamic regret bounded by meaningful quantities depending on how rapidly losses change, which in the worst case are linear in T.

1.2.2 Full information and Bandit games

In our description of the online learning game we did not specify exactly what feedback the player receives from the adversary. Intuitively, games in which the player only observes the loss of the action she played at round t should be harder than games in which she observes the loss of all possible actions in $\mathcal{A}$. Indeed, the player receives $|\mathcal{A}|$ times less information in the former scenario. We call feedback in which only the loss of the played action is revealed ***bandit feedback***. Feedback in which the player observes the full loss vector at every round is referred to as ***full information*** feedback.

The experts and convex optimization problems which we discussed above are both problems with full information feedback. The following simplified version of the experts problem is first studied by Littlestone (1988): at every step all the experts make a binary prediction. The observed loss is just the zero-one loss. Further, it is assumed that there exists at least one expert which is correct at every round of the game, that is they always have zero loss on their prediction. A halving algorithm, which maintains a set of active experts (which have not made a mistake so far) is proposed. At every round of the game the algorithm selects the prediction with majority vote. If the prediction is correct the game continues, otherwise the algorithm updates the set of active experts by removing all experts who have made a mistake. The regret of this algorithm is $O(\log(|\mathcal{A}|))$.

While the halving algorithm is natural, we can not expect that there is always an expert (or action) which has zero loss. Littlestone and Warmuth (1994) propose the Randomized Weighted Majority algorithm. The idea behind the weighted majority algorithm is to keep a set of weights $w_{t,i}$, one per each expert and at every round update the weights as $w_{t+1,i} = (1 - \eta\ell_t(i))w_{t,i}$, where η is some fixed parameter. The expert which is chosen at time t is sampled with probability proportional to their weight. The regret of this strategy is bounded by $O(\sqrt{\log(|\mathcal{A}|)\,T})$ for appropriately set η. Freund and Schapire (1997) later proposed the Hedge algorithm (Algorithm 1). Hedge can be seen as a version of the weighted majority algorithm, however, instead of updating the weights through a linear function of the losses, one updates them with an exponential function. Hedge has similar regret guarantees to the weighted majority algorithm and again enjoys a $O(\sqrt{\log(|\mathcal{A}|)\,T})$ regret bound for appropriately set η. The Hedge algorithm has become a staple in Online Learning literature and

```
Algorithm 1: Hedge
  Input: Step size $\eta$, time horizon $T$
  Output: Sequence of sampled experts $i_1, \ldots, i_T$
  Initialize $w_1 = \mathbf{1}, p_1 = Unif(\mathcal{A})$
  for t=1,...,T do
      Sample $i_t \sim p_t$ and observe loss vector $\ell_t$
      $w_{t+1,i} = w_{t,i} \exp(-\eta\ell_t(i))$, $p_{t+1,i} = \frac{w_{t+1,i}}{\sum_{j\in\mathcal{A}} w_{t+1,j}}, \forall i \in \mathcal{A}$.
  end for
```

many algorithms have used it as a building block, including some of the approaches presented in the current thesis. This algorithm is, in fact, part of a larger family of algorithms known as Online Mirror Descent which we discuss next.

1.2.2.1 Online Mirror Descent and Follow the Regularized Leader

Mirror descent (Nemirovskij and Yudin, 1983) is a generalization of the gradient descent algorithm to Banach spaces (normed vector spaces). We have already seen in our convex optimization example that we can indeed treat convex optimization as an instance of the online learning game. However, the opposite direction is not clear – how do we treat the online learning game as a convex optimization problem. Algorithm 1 suggests the following: treat the losses ℓ_t as gradients and treat the weights w_t as iterates, which are projected to the convex set of distributions over $|\mathcal{A}|$ known as the probability simplex.

Definition 1.2.1. The K-dimensional ***probability simplex*** is the set $\Delta^{K-1} := \{p \in [0,1]^K : \sum_{i=1}^K p_i = 1\}$ of all probability distributions over K items.

The key ingredient of the mirror descent algorithm is a ***potential function***, Ψ, which maps between the space of iterates and the space of gradients. For the rest of this work we will only consider $\Psi : \Delta^{|\mathcal{A}|-1} \to \mathbb{R}$ which are proper, lower semi-continuous and strictly convex. For definitions of the above terms and an introduction to basic convex optimization we refer the reader to Appendix A. The Online Mirror Descent (OMD) algorithm can now be summarized as follows:

$$\begin{aligned} w_{t+1} &= \operatorname*{argmin}_{w} \langle w, \ell_t \rangle + D_{\Psi_t}(w||p_t) \\ p_{t+1} &= \operatorname*{argmin}_{p \in \Delta^{|\mathcal{A}|-1}} D_{\Psi_t}(p||w_{t+1}). \end{aligned} \tag{1.4}$$

In Equation 1.4, $D_{\Psi_t}(\cdot||\cdot)$ is the ***Bregman divergence*** induced by the potential Ψ_t. A formal definition of the Bregman divergence can be found in Appendix A. Algorithm 1 can now be seen as an instance of OMD, where $\Psi_t(w) = \frac{\sum_{i \in \mathcal{A}} w_i \log(w_i)}{\eta}$ is the re-scaled negative entropy.

A related algorithm to OMD is Follow the Regularized leader (FTRL), also known as lazy OMD. The FTRL update is given by

$$\begin{aligned} w_{t+1} &= \operatorname*{argmin}_{w} \langle w, \sum_{s=1}^{t} \ell_s \rangle + \Psi_t(w) \\ p_{t+1} &= \operatorname*{argmin}_{p \in \Delta^{|\mathcal{A}|-1}} D_{\Psi_t}(p||w_{t+1}). \end{aligned} \tag{1.5}$$

It can be shown that for $\Psi_t = \frac{\Psi}{\eta}$, i.e., constant step-size updates, OMD and FTRL follow the same trajectory if initialized at the appropriate point. Somewhat surprisingly if the step-sizes η_t are decreasing throughout the game, it turns out that FTRL and OMD can follow different trajectories (Orabona and Pál, 2018).

The OMD and FTRL algorithms are applicable to a large number of online learning games, and as we will see soon, enjoy meaningful regret guarantees. The general problem of when sublinear regret is achievable for an online learning game is addressed by Rakhlin et al. (2015); Bhatia and Sridharan (2020).

1.2.2.2 Stochastic multi-armed bandits

As alluded to previously, games with ***bandit feedback*** are harder due to the amount of information observed by the player. In general we are going to distinguish two types of bandit games – one in which the losses are generated from some unknown distribution and the goal of the player is to compete against the action with smallest expected loss. This problem is known as the ***Stochastic Multi-armed Bandit*** (stochastic MAB) problem. The stochastic bandit problem dates back to Thompson (Thompson, 1933), and the motivation for the problem was to determine if in a healthcare scenario some treatment is better than placebo or not online, without having to wait for the complete trial to conclude. The second type of bandit game is one in which we do not make any assumptions about how the losses are generated, but only assume that they are in a bounded range (usually $[0, 1]$). We note that actions in the bandit game are referred to as ***arms***.

For convenience we will discuss the stochastic bandit problem in terms of rewards r_t rather than losses. The stochastic K-armed bandit problem is characterized by the expected reward vector μ, with $\mathbb{E}[r_t] = \mu \in \mathbb{R}^K$. We will focus on distributions which are sub-Gaussian with variance proxy equal to 1 or with bounded support in $[0,1]^K$. The mean of the arm with highest reward will be denoted as μ_{i^*} or μ_1. The regret for the stochastic MAB problem can now be written as

$$R(T) = T\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t}\right], \tag{1.6}$$

where i_t is the arm played at time t and the expectation is with respect to the randomness in the player's strategy and the sampling of the rewards.

In general there are three types of algorithms used for solving the stochastic bandit problem. The first type is based on a Bayesian view of the world, in which the player maintains a prior over possible distributions from which the rewards are sampled and further updates a posterior based on the observed rewards. The posterior is then used to sample a distribution and the player plays the best arm of the sampled distribution. This approach is known as Thompson Sampling (Thompson, 1933). While Thompson Sampling has enjoyed wide applications in practice, its regret was not completely understood until recently (Agrawal and Goyal, 2012, 2013; Kaufmann et al., 2012; Agrawal and Goyal, 2017). First, Agrawal and Goyal (2013) show that Thompson sampling with a Beta distribution prior achieves regret $O(\sqrt{TK\log(T)})$, and regret $O(\sqrt{TK\log(K)})$ with a Gaussian prior. The second result is in general not improvable for the Thompson sampling strategy as shown by the authors. Kaufmann et al. (2012) provide another type of regret upper bound for Thompson sampling which only depends on the problem instance. In particular, they show a regret bound of order $O\left(\sum_{i\neq i^*} \frac{\log(T)\Delta_i}{kl(\mu_i||\mu_{i^*})}\right)$, where $kl(a||b)$ is the KL-divergence between two Bernoulli random variables with means a and b respectively and $\Delta_i = \mu_{i^*} - \mu_i$ is the ***sub-optimality gap***. This regret is asymptotically optimal.

The second type of algorithms are the ones based on optimism in the face of uncertainty (OFU) principle, which has become a powerful tool in different extensions of the MAB problem such as contextual bandits and Reinforcement Learning. The principle suggests that if the player is uncertain about which the best action is, she should proceed to play the action which could attain the highest reward, based on her observations so far. More formally, at every round of the game, for every arm i, the player keeps empirical estimates of the mean, $\widehat{\mu}_i$, and further adds a bonus term depending on how many times the arm have been observed. The player then selects the arm with highest empirical mean plus bonus term. This strategy was proposed by Auer et al. (2002a) in the form of the UCB-I algorithm and comes with $O(\sqrt{TK\log(T)})$ and $O\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i}\right)$ regret guarantees.

The final type of strategies are ones that we have already discussed for the full information setting in the form of OMD and FTRL. The idea is to construct unbiased estimators of each of the reward vectors at time t and feed them to the OMD algorithm with appropriate potential function. As long as the second moment of the estimators is well controlled, regret bounds of the order $\tilde{O}(\sqrt{KT})$ are obtainable, where the $\tilde{O}$ notation hides poly-logarithmic factors. In particular Auer et al. (2002b) propose the Exp3 algorithm which is just Algorithm 1 with $\widehat{r}_t = e_{i_t}\frac{r_{i_t}}{p_{t,i_t}}$. The Exp3 algorithm enjoys a regret bound of the order $O(\sqrt{KT\log(K)})$. Gap-dependent bounds are harder to show for OMD strategies, however, recent works have made significant progress on the matter (Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Wei and Luo, 2018; Zimmert and Seldin, 2019), with Zimmert and Seldin (2019) solving the problem optimally, under the condition that there exists a unique best arm.

1.2.2.2.1 Min-max lower bounds and instance-dependent lower bounds.

All of the three approaches described so far enjoy upper bounds on their regret of the order $\tilde{O}(\sqrt{KT})$ or an instance dependent regret bound which depends on the distribution of the unknown reward vector. Are these bounds optimal? As it turns out, the min-max regret (maximizing over all possible problem instances and minimizing over all possible algorithms) for the stochastic MAB problem is $\Theta(\sqrt{KT})$. The first algorithm to achieve this rate is MOSS (Audibert and Bubeck, 2009), which falls into the category of OFU algorithms.

The results for instance dependent lower bounds are more involved. To understand the bounds, we first need to define the set of confusing bandit environments for a fixed arm $i \neq i^*$. Let $\mathbb{P}_i$ denote the distribution of arm i with mean μ_i in our original bandit problem. We define the set

$$\Lambda_i = \{\mathbb{P}'_i : \mathbb{E}_{\mathbb{P}'_i}[r_t(i)] > \mu_{i^*}\}.$$

This is precisely the set of distributions for the reward of the i-th arm, which have larger expected reward than the optimal μ_{i^*}. Lai and Robbins (1985) show that any strategy which incurs regret at most T^α for all $\alpha > 0$ must suffer regret at least $\Omega\left(\sum_{i\neq i^*} \frac{\log(T)\Delta_i}{\inf_{\mathbb{P}'_i\in\Lambda_i} KL(\mathbb{P}_i||\mathbb{P}'_i)}\right)$ as $T \to \infty$, where $KL(P||Q)$ is the KL-divergence between distributions P and Q. Formally the lower bound states that

$$\liminf_{T\to\infty} \frac{R(T)}{\log{(T)}} \geq \Omega\left(\sum_{i\neq i^*} \frac{\Delta_i}{\inf_{\mathbb{P}'_i\in\Lambda_i} KL(\mathbb{P}_i||\mathbb{P}'_i)}\right). \tag{1.7}$$

Equation 1.7 implies that Thompson Sampling is asymptotically optimal for rewards sampled from Bernoulli or Gaussian distributions. Further, the UCB-I algorithm is asymptotically optimal for rewards sampled from Gaussian distributions as the KL-term in the bound evaluates to Δ_i^2. Another algorithm from the OFU family which achieves the same guarantees as Thompson Sampling and has favorable practical performance is KL-UCB (Garivier and Cappé, 2011).

Finally, we mention another asymptotically optimal strategy which is used in the problem of best arm identification known as successive elimination (Even-Dar et al., 2002). The strategy is similar to the UCB-I algorithm as it also keeps confidence intervals around the empirical means for each arm, however, every arm is played until we can confidently say that it is worse than another arm at which point the worse arm is discarded and never played until the end of the game again.

1.2.2.3 Adversarial multi-armed bandits

In the adversarial MAB problem, there are no assumptions made regarding the observed losses, except that they are bounded in $[0, 1]$. The problem was first studied by Auer et al. (2002b), who propose the Exp3 algorithm, which as we already mentioned is a modification of Algorithm 1. Exp3 comes with a $O(\sqrt{KT \log(K)})$ regret bound in the adversarial setting. Other popular potential functions with which the OMD (or FTRL) algorithms are initialized and enjoy adversarial regret guarantees are the Log-Barrier potential defined as $\Psi(p) = -\sum_{i=1}^{K} \log(p_i)$ and the 1/2-Tsallis entropy defined as $\Psi(p) = -\sum_{i=1}^{K} \sqrt{p_i}$. The log-barrier based OMD algorithm comes with a $O(\sqrt{KT \log(T)})$ guarantee and the FTRL algorithm based on the 1/2-Tsallis entropy introduced by Audibert and Bubeck (2009) and later also studied in Zimmert and Seldin (2019) enjoys $O(\sqrt{KT})$ regret. Because the adversarial bandit problem subsumes the stochastic bandit problem, the same min-max lower bound applies and hence the 1/2-Tsallis entropy FTRL algorithm, known as Tsallis-INF, satisfies the desired regret bound. It is natural to ask if there are "instance dependent" upper bounds for the adversarial setting. The work of (Hazan and Kale, 2011) answers this question to the affirmative by defining a notion of total variation of the losses and bounding the regret in terms this variation. Further extensions and similar quantities can be found for the full information setting in (Rakhlin and Sridharan, 2013) and for the bandit setting in (Bubeck et al., 2018, 2019; Wei and Luo, 2018).

The topic of bandit learning is broad and we will only scratch the surface in this thesis. For a more complete introduction we refer the reader to the works of Bubeck et al. (2012) and Lattimore and Czepesvari (2018).

1.3 Graph theory concepts

Part of this thesis will address a setting for the online learning game which interpolates between full information and bandit feedback. Such feedback is modeled by a ***feedback graph*** with vertices corresponding to actions and edges describing which actions reveal information about each other. We now revisit some graph theoretic concepts which are used in Chapter 2.

We begin by recalling the definition of a graph.

Definition 1.3.1 (Undirected graph)**.** An ***undirected graph*** $G = (V, E)$ is a tuple containing a set of vertices V and a set of edges E, where an edge $\{u, v\} \in E, u, v \in V$ is an unordered pair of vertices.

We will only work with undirected graphs, however, all of the results in Chapter 2 extend to directed graphs as well.

Definition 1.3.2 (Directed graph)**.** A ***directed graph*** $G = (V, E)$ is a tuple containing a set of vertices V and a set of edges E, where an edge $(u, v) \in E, u, v \in V$ is an ordered pair of vertices.

We say that u, v are neighbors (or adjacent) in an undirected graph if $\{u, v\}$ is an edge of the graph. Further, we say that u is an in-neighbor to v in a directed graph if $(u, v) \in E$ and is an out-neighbor to v if $(v, u) \in E$.

The next set of definitions are for undirected graphs.

Definition 1.3.3 (Complete graph)**.** A ***complete graph*** G is a graph such that there is an edge $\{u, v\} \in E$ for every two vertices $u, v \in V$.

We call a sub-graph of G a graph with a vertex set which is a subset of V and an edge set which a subset of E.

Definition 1.3.4 (Clique partition (Erdöos et al., 1988))**.** A ***clique*** of a graph G is a complete sub-graph of G. A ***clique partition*** is a set of cliques of G such that each vertex of G is contained in exactly one clique.

Definition 1.3.5 (Clique partition number)**.** The ***clique partition number*** $\bar{\chi}(G)$ for a graph G is the size of the minimum clique partition, that is the smallest size clique partition over all clique partitions.

Finding a clique partition of G is known to be NP-hard. Further, approximating the clique partition number to a factor of $O(|V|^{1-\epsilon})$ for any $\epsilon > 0$ is also likely computationally hard (Hastad, 1999; Engebretsen and Holmerin, 2000). Cliques are tightly related to the notion of independent set of a graph.

Definition 1.3.6 (Independent set)**.** A subset of vertices of G is an ***independent set*** if no two vertices are adjacent.

Definition 1.3.7 (Independence number)**.** The ***independence number***, $\alpha(G)$ of a graph G is the size of a maximum independent set.

An clique of size k in G corresponds to an independent set of size k in the graph constructed from G by taking the same vertex set and an edge set consisting of edges $\{u, v\}$ iff $\{u, v\}$ is not part of the edge set of G. This suggests that approximating the independence number to a factor better than $|V|^{1-\epsilon}$ is also likely computationally hard (Hastad, 1999).

Another important graph theoretic quantity for our work is the domination number.

Definition 1.3.8 (Dominating set)**.** A ***dominating set*** for a graph G is a subset of V such that every vertex in V is adjacent to some vertex in the dominating set.

Definition 1.3.9 (Domination number)**.** The ***domination number***, $\gamma(G)$ of a graph G is the size of a minimum dominating set.

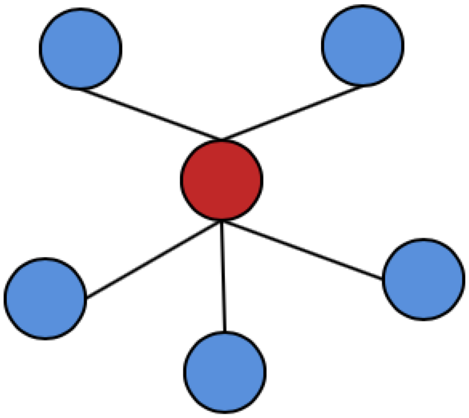


Figure 1-1: Graph example

The graph in Figure 1-1 has a maximum independent set of size 5 with vertices in blue and a minimum dominating set of size 1 in red. Further, the clique partition number for the graph consists of the blue vertices and is again equal to 5. It turns out that the domination number is always no larger than the independence number and the independence number is always no larger than the clique partition number, that is $\gamma(G) \leq \alpha(G) \leq \bar{\chi}(G)$ (Bollobás and Cockayne, 1979; Goddard and Henning, 2013). We note that there is a fourth quantity of interest which is the size of the maximum acyclic graph denoted by $\mathbf{mas}(G)$. And acyclic graph is a graph which contains no cycles, that is a sequence of edges $\{u_1, u_2\}, \{u_2, u_3\}, \ldots, \{u_n, u_1\}$ which trace a closed path in the graph starting and ending at the same vertex. It turns out that in undirected graphs $\mathbf{mas}(G) = \alpha(G)$.

Unlike the clique partition number and the independence number, there is a very simple algorithm which can approximate the domination number up to a $\log(|V|)$ factor. The pseudo code is given in Algorithm 2.

The following notes http://ac.informatik.uni-freiburg.de/teaching/ss_12/netalg/lectures/chapter7.pdf provide us with a proof that the greedy Algorithm 2 returns a dominating set R which is $2 + \log(\Delta)$ approximation to the smallest size minimal dominating set, where Δ is the maximum degree if G. It

Algorithm 2: Greedy algorithm for minimum dominating set

```
Input: An undirected graph $G(V, E)$
Output: A dominating set $S$
$R = \emptyset$
if $V == \emptyset$ then
    Return $S$
else
    Find $v \in V$ s.t. $deg(v)$ is maximized
    $R = S \bigcup \{v\}$
    $V = V \setminus \{\{v\} \bigcup N(v)\}$ and update $G$ to be the induced graph on the new set of vertices $V$.
end if
```

is possible to implement the algorithm so that it has total runtime of the order $O((|V| + |E|) \log(|V|))$ (e.g. http://homepage.cs.uiowa.edu/~sriram/3330/spring17/greedyMDS.pdf). We note that this is essentially the Greedy Set Cover algorithm of Chvatal (1979) and that it is possible to extend to directed graphs, by replacing the degree of v by the out-degree of v and the neighbours of v by just the vertices which have in-going edge from v.

1.4 Algorithmic Game theory

Algorithmic game theory is a broad topic which in general studies the performance of self-interested agents in games, what strategies lead to good performance guarantees in games and how to design games where self-interested agents enjoy good performance guarantees. Here good performance guarantees pertain to some type of socially-economic metrics such as maximizing the overall utility of players participating in the game.

In this work we focus on the following type of multi-player games. The game consists of n players. The i-th player has a finite action set $\mathcal{A}_i$, $|\mathcal{A}_i| = k_i$ and a reward (or utility) function $u_i : \mathcal{A} \to [0, 1]$, where $\mathcal{A} = \times_{i=1}^{n} \mathcal{A}_i$. Further, we will assume that the game proceeds in T rounds and during each round the players get to interact with each other only through choosing actions in their own action set and observing the utility for the selected actions at the given round. Our goal is to study what player strategies would lead to a steady-state of the game in which no player will have an incentive to deviate from the prescribed strategy. Such states are known as an ***equilibrium*** of the game.

Existence of different types of equilibrium has been a fundamental question in game theory. Perhaps the most natural equilibrium in multi-player games is the one in which *no player has incentive to change the action they are playing as long as every other player continues to play the same action.* This is known as a ***Pure Nash equilibrium*** (PNE). Even though PNEs are easy to describe and seem like a reasonable equilibrium for self-interested agents, PNEs are not even guaranteed to exist in some types of games. For example consider the simple two-player game of Rock, Paper, Scissors, in which rock beats scissors, scissors beats paper and paper beats rock. It turns out that there is no one fixed action a player can choose, such that that action wins all the time, while the other player has no incentive to switch from their losing/tie action. Instead of considering only the best fixed action it is natural to extend the policy of each player to the best fixed distribution over actions. Such reasoning leads us to the concept of a Mixed Nash equilibrium (MNE).

Definition 1.4.1. A Mixed Nash equilibrium is a product distribution $\sigma = \times_{i=1}^{n}$ over $\mathcal{A}$, where σ_i is a distribution over $\mathcal{A}_i$ satisfying the following:

$$\mathbb{E}_{(a^1,\ldots,a^i,\ldots,a^n)\sim\sigma}[u_i(a^1,\ldots,a^n)] \geq \mathbb{E}_{(a^1,\ldots,a^i,\ldots,a^n)\sim\sigma}[u_i(a^1,\ldots,(a')^i,\ldots,a^n)],$$

for all actions $(a')^i \in \mathcal{A}_i$ and all players $i \in [n]$.

MNEs are natural and do not require any coordination between the players as the i-th player only needs to sample according to their own distribution σ_i. Further MNEs exist for all (reasonable) games (Nash et al., 1950; Nash, 1951). Unfortunately it is very likely that there are no polynomial time algorithms for approximating Nash equilibria

(Daskalakis et al., 2009) even for two-player games (Chen and Deng, 2006), outside of some special type of games such as zero-sum games. The fact that approximating MNE is most likely computationally intractable motivates the study of other types of natural equilibria which always exist and can also be found efficiently.

Consider the following 2-player traffic light game. Each player can choose between two actions: stop and go. The utility functions are given in Table 1-I. In the table we distinguish between a column player and a row player. The utility of the row player is shown in the first entry of the tuple and the utility of the column player is the second entry. For the purpose of our example we allow negative utilities. Suppose a traffic signal instructs the column player to stop and the row player to go. The column player only sees their own signal, however, they can deduce that the row player has the right of way and so the column player's best action is to obey the signal. The same reasoning holds for the row player. Suppose now that the traffic light is randomized in the sense that with probability 1/2 it signals the column player to stop and the row player to go, and with probability 1/2 it signals the column player to go and the row player to stop. If both players decide to obey the traffic signal they observe, then they are at an equilibrium as if either player deviates, their utility will decrease. The traffic signal plays the role of a central agent which coordinates and hence correlates the actions of the two players. The type of equilibrium which we just described is known as a ***Correlated equilibrium*** (CE)(Aumann, 1987). Notice that this equilibrium also can not be a Nash, as it does not factor into product distributions over the action sets of each player. Indeed, the induced distribution over $\mathcal{A}$ is $\sigma((stop, go)) = 1/2, \sigma((go, stop)) = 1/2$.

	stop	go
stop	(0,0)	(0,1)
go	(1,0)	(-5,-5)

Table 1-I: Traffic light game

Definition 1.4.2. A Correlated equilibrium is a distribution σ over the action space $\mathcal{A}$ such that for every player i it holds that

$$\mathbb{E}_\sigma[u_i(a^1,\dots,a^n)|a^i] \geq \mathbb{E}_\sigma[u_i(a^1,\dots,(a')^i,\dots,a^n)|a^i], \forall (a')^i \in \mathcal{A}_i.$$

The interpretation of a CE is that if the i-th player is shown their action, sampled according to σ, then they have no incentive to deviate, given that every other player behaves according to σ. The existence of CEs can be confirmed by observing that any MNE is also a CE. Further, CEs are computable in polynomial time, for example by linear programming (Papadimitriou and Roughgarden, 2008).

The final equilibrium concept we discuss is that of a ***Coarse correlated equilibrium*** (CCE). Similarly to MNE and CE, a CCE is a distribution σ over $\mathcal{A}$. Unlike CEs, however, the agent must commit before hand to following the action sampled according to σ. Formally, a CCE is defined as follows.

Definition 1.4.3. A Coarse correlated equilibrium is a distribution σ over the action space $\mathcal{A}$ such that for every player i it holds that

$$\mathbb{E}_\sigma[u_i(a^1,\dots,a^n)] \geq \mathbb{E}_\sigma[u_i(a^1,\dots,(a')^i,\dots,a^n)], \forall (a')^i \in \mathcal{A}_i.$$

CCEs are also computable in polynomial time and in fact every CE is also a CCE. To conclude we have the following nestedness of equilibria PE $\subset$ MNE $\subset$ CE $\subset$ CCE, where each of the inclusions are non-strict.

Even though CEs and CCEs are efficiently computable, the question of: "How good this equilibria are compared to Nash?" remains. There are multiple ways to quantify the goodness of an equilibrium of a game. Two such notions are the Price of Anarchy (PoA) and Price of Stability (PoS). For a given social welfare function, mapping players policies to a real value, the PoA is defined as the ratio between the welfare of the best overall policy to the welfare of the worst policy belonging to an equilibrium class. The PoS is defined similarly, however, we compare the best overall policy to the best overall policy in the equilibrium class. Part of algorithmic Game theory studies how PoA and PoS of Nash equilibrium compare to PoA and PoS of other equilibrium. For example in the special type of “smooth” games, the worst Nash is no worse than the worst CCE (Roughgarden, 2015).

1.4.1 Equilibrium and no-regret algorithms

CCEs and CEs are also efficient to approximate through natural strategies in which every player minimizes their own regret in the repeated game (Foster and Vohra, 1997; Fudenberg and Levine, 1999; Hart and Mas-Colell, 2000; Blum and Mansour, 2007). In particular minimizing swap regret leads players to a correlated equilibrium and minimizing external regret leads to coarse correlated equilibrium. Hazan and Kale (2008) show a very general correspondence between finding equilibria, fixed-point computation and the existence of certain types of no-regret algorithms. In followup work Mohri and Yang (2014) and Mohri and Yang (2017) showed that minimizing their notions of conditional swap regret and transductive regret also leads to new notions of equilibrium.

We now outline how to convert a no-external regret algorithm to an algorithm which approximates a CCE. First, assume that the i-th player acts according to a no-regret algorithm with action space $\mathcal{A}_i$ and reward function at time t given by $r_t(\cdot) = u(a_t^1, \ldots, a_t^{i-1}, \cdot, a_t^{i+1}, \ldots, a_t^n)$, where a_t^j denotes the action taken by player j at time t. The no-regret algorithm guarantees that after T rounds of the game for any action $a^i \in \mathcal{A}_i$ it holds that $\mathbb{E}\left[\sum_{t=1}^T r_t(a^i) - r_t(a_t^i)\right] \leq O(\sqrt{T})$. After T rounds of the game we define the distribution $\widehat{\sigma}_T$ on $\mathcal{A}$ as follows. A central agent who coordinates the players samples a round $t \in [T]$ uniformly at random. Next the i-th player acts according to the strategy which prescribes they play action a_t^i at round t of the no-regret game. Thus the expected reward of player i is $\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T r_t(a_t^i)\right] =$

$\mathbb{E}_{(a^1,\dots,a^n)\sim\widehat{\sigma}_T}[u_i(a^1,\dots,a^n)]$. Further, the no-regret guarantee implies that

$$\begin{aligned}\mathbb{E}_{(a^1,\dots,a^n)\sim\widehat{\sigma}_T}[u_i(a^1,\dots,a^n)] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t(a_t^i)\right] &\geq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t((a^i)')\right] - O\left(1/\sqrt{T}\right) \\ &= \mathbb{E}_{(a^1,\dots,a^n)\sim\widehat{\sigma}_T}[u_i(a^1,\dots,(a^i)',\dots,a^n)] - O\left(1/\sqrt{T}\right).\end{aligned}$$

Thus the empirical distribution of play converges to the set of CCE in a weak convergence sense. Because we are considering finite action-space games, weak convergence is equivalent to strong convergence and we arrive at the fact that any convergent sub-sequence of $(\widehat{\sigma}_T)_{T=1}^{\infty}$ will have a limit point in the set of CCEs. Showing that no-swap regret play leads to a CE is slightly more involved, however, follows the same general idea.

Finally, we note that it is possible to derive John von Neumann's min-max theorem for two-player zero sum games using the fact that no-regret algorithms exist and following their empirical play would lead to a MNE in a two-player zero sum game.

1.5 Reinforcement Learning

Reinforcement Learning also studies the online learning game, however, additional complexity is introduced into the problem by considering it as a ***Markov Decision Process*** (MDP) (Puterman, 1994). Formally, an MDP consists of an action set $\mathcal{A}$, a set of states $\mathcal{S}$, a (deterministic) reward functions $r : \mathcal{S}\times\mathcal{A}\to[0,1]$ mapping state-action pairs to rewards, a transition kernel $P : \mathcal{S}\times\mathcal{A}\to\Delta^{|\mathcal{S}|-1}$ mapping state-action pairs to a distribution over states, and a distribution over starting states σ_0. Without loss of generality we can assume that there is a single starting state s_0 with a single available action a_0 such that $P(\cdot|s_0,a_0)\equiv\sigma_0$. The number of states is $|\mathcal{S}| = S$ and the number of actions is $|\mathcal{A}| = A$. Further, after playing a state-action pair, (s,a), the player only observes a random variable $R(s,a)$ with expectation $\mathbb{E}[R(s,a)] = r(s,a)$ and her transition to the state $s' \sim P(\cdot|s,a)$. The player does not know r or P and the goal of the player is to maximize her cumulative reward. We have already encountered a simple type of MDP in the form of the stochastic MAB problem. Indeed the stochastic bandit problem is an MDP with a single state and actions $\mathcal{A}$.

There are several types of RL problems one can consider based on the interaction protocol, cumulative reward and size of the state-action space. In this work we will consider the ***episodic***, ***finite horizon***, ***tabular*** setting. The protocol is as follows. The game proceeds in K episodes (not to be confused with the size of the action set in the online game). Each episode is of length H. At every episode the agent chooses a (deterministic[4]) policy $\pi : \mathcal{S} \to \mathcal{A}$, that is a function mapping states to actions and proceeds to observe the rewards and transitions generated by following π for H rounds. At the end of the episode the player's total reward is $\mathbb{E}_{\pi}[\sum_{t=1}^{H} R(S_t, A_t)]$, where S_t, A_t are random variables sampled according to the Markov process induced by following π and the transition kernel P. After the episode has concluded, the player updates her policy and proceeds onto the next episode. We will again measure the success of the player through regret, where the comparator is the policy with highest cumulative reward over a fixed episode, denoted by π^*. The regret is formally defined as

$$\begin{aligned} R(K) &= \sum_{k=1}^{K} \left(\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[R(S_h, A_h)\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[R(S_h, A_h)\right] \right) \\ &= \max_{\pi} \sum_{k=1}^{K} \left(\sum_{h=1}^{H} \mathbb{E}_{\pi}\left[R(S_h, A_h)\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[R(S_h, A_h)\right] \right), \end{aligned} \tag{1.8}$$

where π_k is the policy selected by the player at episode k. We additionally assume that the MDP has a ***layered*** structure. That is, every state s belongs to some layer $\kappa(s) = h \in [H]$ and the only non-zero transition probabilities are between states s and s' such that $\kappa(s) + 1 = \kappa(s')$, i.e., $P(s'|s, a) > 0 \implies \kappa(s) + 1 = \kappa(s')$.

1.5.1 Value function, Q-function, Bellman optimality

The ***value*** function of a policy π is a functional mapping from states to real numbers and it assigns to each state the expected reward of policy π starting from state s.

[4] This is WLOG as for any reasonable MDP there exists a deterministic policy which is optimal (Puterman, 1994).

Formally,

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} R(S_h, A_h)|S_{\kappa(s)} = s\right]. \tag{1.9}$$

With this notation we can concisely write the cumulative reward of the strategy of a player as $\sum_{k=1}^{K} V^{\pi_k}(s_0)$.

The second quantity of interest is the Q-function which maps a state-action pair to a real number. It is defined as follows

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} R(S_h, A_h)|S_{\kappa(s)} = s, A_{\kappa(s)} = a\right], \tag{1.10}$$

and captures the expected reward of the player if she plays according to π after playing action a in state s. Note that the following identities hold from the definition of the value function and Q-function

$$V^{\pi}(s) = Q^{\pi}(s, \pi(s)) = r(s, \pi(s)) + \mathbb{E}_{S' \sim P(\cdot|s,\pi(s))}\left[V^{\pi}(S')\right].$$

Further, we can think of V^{π} as a vector in $\mathbb{R}^S$ indexed by states, and similarly think of Q^{π} as a vector in $\mathbb{R}^{S \times A}$ indexed by state-action pairs. This implies we can write the expectation $\mathbb{E}_{S' \sim P(\cdot|s,a)}\left[V^{\pi}(S')\right]$ as an inner product between the transition kernel and the value function as

$$\mathbb{E}_{S' \sim P(\cdot|s,a)}\left[V^{\pi}(S')\right] = \langle P(\cdot|s,a), V^{\pi}\rangle.$$

Further, using $P : \mathbb{R}^S \to \mathbb{R}^{S \times A}$ as a linear operator with (s,a)-th row equal to $P(\cdot|s,a)$, we can write in vector notation

$$Q^{\pi} = r + PV^{\pi}.$$

We also adopt the following standard notation $V_k = V^{\pi_k}$ and $Q_k = Q^{\pi_k}$.

We defined the regret of the player's strategy by comparing against an optimal policy π^*, but it is not entirely clear that such an optimal policy exists. Such a policy exists and can be thought of as the ***deterministic*** policy maximising the value function on s_0, that is $\pi^* = \operatorname{argmax}_\pi V^\pi(s_0)$. The optimal policy is not necessarily unique, however, the value of the optimal policy $V^* \equiv V^{\pi^*}$ is unique. Further, one can show that the greedy policy with respect to the Q-function is an optimal policy. That is, π^* satisfies the following equation

$$V^{\pi^*}(s) = \max_{a \in \mathcal{A}} r(s,a) + \langle P(\cdot|s,a), V^{\pi^*} \rangle, \forall s \in \mathcal{S}.$$

This equation is known as the ***Bellman optimality*** equation and is at the core of designing algorithms for regret minimization (Puterman, 1994). Finally, we use the notation $Q^* = Q^{\pi^*}$.

1.6 Adaptive adversaries and Policy regret

So far we have seen several different measures of regret – external regret, internal regret, swap regret, etc. Further, even though we made a point to distinguish between expected regret for adaptive adversaries and expected regret for oblivious adversaries the main idea behind our comparator class remained the same – compare with the best fixed action or policy in hind sight. Consider the following simple example: the player is a self-driving car and the available actions determine whether the car turns, accelerates, decelerates, uses a turn signal, etc. If the player decides to go through a red light at an intersection then the next available actions and respective losses she would observe are all going to be poor as the player would have to avoid a car crash. Given all of the observed losses are poor it might turn out that the best action in hind sight would be to just go as fast as possible through the intersection. However, a better player, who stops at the red light, would have observed a different sequence of losses for which the best action in hind sight was to stop and wait for the traffic light to turn green. We can now see that external regret might not be the best choice for how we measure performance of the agent, as from a practical point of view we would much rather prefer the agent stopping at the red light rather than going full speed ahead. In general, because adaptive adversaries tailor their loss sequence to the player's past actions and overall strategy, it is not necessary for the best action in hind sight on the observed loss sequence to be the best overall action for the game, as a different strategy chosen by the player can result in a different loss sequence.

The above problem was first investigated by Merhav et al. (2002) in the full information setting and by Arora et al. (2012a) in the bandit setting. To address the critical issue described above, Arora et al. (2012a), propose ***Policy regret*** defined as follows

$$P(T) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_{1:t}) - \ell_t(a, \dots, a)\right], \tag{1.11}$$

where we use the shorthand $a_{i:j}$ to denote $a_i, a_{i+1}, \dots, a_j$. It is not hard to see that if the adversary is not constrained in any other way then it is impossible to obtain sublinear policy regret bounds. This follows by observing that all the losses could just depend on the very first action the player has chosen and it is impossible for the player to know which is the best action before the game begins. Most work on policy regret has constrained the adversary to have ***bounded memory***, that is the adversary can only base its losses on the last m actions the player has chosen. In our notation this is equivalent to saying that $\ell_t(a_{1:t}) = \ell_t(a_{t-m+1:t}), \forall t \in [T]$.

1.6.1 Policy regret bounds in the full information game

Merhav et al. (2002) consider the full information game, where the player gets to observe the full loss for all available sequences of m actions. They propose a strategy with regret bounded by $O(T^{2/3}(\log(K))^{1/3})$ for $m = 2$. Later, Anava et al. (2013) observe that the policy regret minimization when the adversary is m-memory bounded can be reduced to minimizing external regret when there is an associated cost for switching between different actions in consecutive rounds and propose an algorithm for the general online convex optimization problem with regret $\tilde{O}(\sqrt{mT})$ (the asymptotic notation hides the complexity of the decision space for the convex optimization game).

The first algorithm in the full information game which was able to minimize regret in the presence switching costs (and therefore policy regret) was proposed by Kalai and Vempala (2005) and is based on the Follow the Perturbed Leader (FTPL) strategy. The same algorithm can be modified to achieve a $O(\sqrt{mT\log(K)})$ policy regret bound. Another algorithm which enjoys $O(\sqrt{mT\log(K)})$ regret is the shrinking dartboard algorithm proposed by Geulen et al. (2010). Finally, the more general question of when sublinear policy regret is achievable is addressed in the work of Bhatia and Sridharan (2020).

1.6.2 Policy regret bounds in the bandit game

In the bandit game, Arora et al. (2012a) show how to reduce any sub-linear external regret algorithm to a sub-linear policy regret algorithm via a mini-batching trick. The proposed algorithms enjoy a $\tilde{O}(m^{1/3}K^{1/3}T^{2/3})$ regret bound in the multi-armed bandit problem. There is a discrepancy between the $\sqrt{T}$ regret bounds achievable in the full information setting versus the above $T^{2/3}$ bound. Somewhat surprisingly, Dekel et al. (2014), show that this gap can not be breached, by showing a lower bound of the order $\tilde{\Omega}(K^{1/3}T^{2/3})$ for the setting of bandits with switching costs. This lower bound also extends to the policy regret setting, because the two settings are equivalent whenever m is constant, compared to T. The dichotomy between the achievable regret in the full information setting versus the achievable regret in the bandit setting introduces the following interesting question – is it possible to achieve improved policy regret guarantees if the player is not faced with the worst-case setting. The following chapters try to answer this question in the positive.

Chapter 2

Policy regret in the presence of side observations

We study the adversarial multi-armed bandit problem where the learner is supplied with partial observations modeled by a ***feedback graph*** and where shifting to a new action incurs a fixed ***switching cost***. We give two new algorithms for this problem in the informed setting. Our best algorithm achieves a pseudo-regret of $\tilde{O}(\gamma(G)^{\frac{1}{3}}T^{\frac{2}{3}})$, where $\gamma(G)$ is the domination number of the feedback graph. This significantly improves upon the previous best result for the same problem, which was based on the independence number of G. We also present matching lower bounds for our result that we describe in detail. Finally, we give a new algorithm with improved policy regret bounds when partial counterfactual feedback is available. The main contributions of this chapter are based on Arora et al. (2019). This work was done in collaboration with Dr. Raman Arora and Dr. Mehryar Mohri.

2.1 Online learning with partial feedback

In Chapter 1 we discussed to types of feedback for the online learning game – full information, in which the player observes the losses of all her actions, and bandit feedback in which the player observes only the loss of the action she has played at the current round. We further saw that policy regret minimization for bounded memory adversaries has very different min-max rates, i.e. $\Theta(\sqrt{T\log(K)})$ in the full information game versus $\Theta(K^{1/3}T^{2/3})$ in the bandit game. In this chapter we study a third model of partial feedback which lays between full information and bandit feedback. In the partial feedback setting after playing a action, the player observes the loss of her action (as in the bandit setting), and also can observe the loss of other "neighboring" actions. A motivating example is as follows. A commercial bank issues various credit card products, many of which are similar, e.g., different branded cards with comparable fees and interest rates. At each round of the game, the bank offers a specific product to a particular sub-population (e.g., customers at a store). The payoff observed for this action also reveals feedback for related cards and similar sub-populations.

Formally, which actions reveal information about each other is determined by a ***feedback graph***, or more generally a sequence of feedback graphs $\{G_t\}_{t=1}^T$. In an undirected feedback graph, each vertex represents an action and an edge between vertices a and a' indicates that the loss of action a' is observed when action a is selected and vice-versa. The bandit setting corresponds to a feedback graph reduced to only self-loops at each vertex, the full information setting to a complete graph.

The work of Mannor and Shamir (2011) is the first to study the online learning problem when feedback graphs model which losses the player gets to observe after choosing an action. Their work proposes two algorithms, the ExpBan, which has regret $O(\sqrt{\sum_{t=1}^T \bar{\chi}(G_t)})$, where $\bar{\chi}(G)$ is the clique partition number, and the ELP algorithm which has regret $O(\sqrt{\sum_{t=1}^T \alpha(G_t)})$. They also show a regret lower bound when $G_t = G$ for all G of the order $\Omega(\sqrt{\alpha(G)T})$. The work of Alon et al. (2013) improves on that Mannor and Shamir (2011) in two significant ways. First the authors consider a setting in which the feedback graphs are directed and can be observed only after taking an action. Secondly the provided algorithms even for the informed setting are more efficient than the ones in Mannor and Shamir (2011). Their algorithm Exp3-SET has regret $\tilde{O}(\sqrt{\sum_{t=1}^T \mathbf{mas}(G_t)})$ for the uninformed setting with directed feedback graphs. Here $\mathbf{mas}(G_t)$ is the size of the maximum acyclic subgraph of G_t. When considering the undirected setting $\mathbf{mas}(G_t)$ can be replaced by $\alpha(G_t)$. In the informed setting Alon et al. (2013) propose the algorithm Exp3-DOM, which requires approximating or computing a minimum dominating set of G_t. Kocák et al. (2014) avoid such tedious computation with their algorithm Exp3-IX. The regret achieved by their algorithm is of the order $\tilde{O}(\sqrt{\sum_{t=1}^{T} \alpha(G_t)})$ even in the uninformed setting. The paper also extends the implicit exploration trick used by Exp3-IX to Follow the Perturbed Leader and solves the combinatorial bandit problem with side observations, where at each round the player is permitted to select n out of the $|\mathcal{A}|$ available actions. The achieved regret is of the order $\tilde{O}(n^{2/3}\sqrt{\sum_{t=1}^{T} \alpha(G_t)})$. In Alon et al. (2015) the authors consider a setting where the feedback graph system is fixed i.e. $G_t = G$ for all $t \in [T]$, however, the graph need not have self loops. The authors distinguish between three settings. First a setting in which each vertex either has a self loop or is revealed by all other vertices, called the strongly observable setting. The second setting assumes that every vertex is revealed by some other vertex but there exists at least one vertex which is not strongly observable. This setting is called the weakly observable setting. The third setting is that of some vertex not being revealed by any other vertex. This is called the not observable setting. Alon et al. (2015) show that the regret bounds are respectively $\tilde{\Theta}(\sqrt{\alpha(G)T})$ in the strongly observable setting, $\tilde{\Theta}(\gamma(G)^{1/3}T^{2/3})$ in the weakly observable setting and $\Theta(T)$ in the not observable setting. The work of Cohen et al. (2016) studies a setting where the feedback graph is never fully revealed to the player. They show that if the feedback graph and the losses are generated by the adversary a lower bound for the regret of any strategy is $\Omega(\sqrt{|\mathcal{A}|T})$, which matches the lower bound of the bandit setting. In contrast it is possible to recover a $\tilde{\Theta}(\sqrt{\alpha(G)T})$ regret bound if the losses are stochastic.

Very recently the works of (Lykouris et al., 2018; Lee et al., 2020b) derive regret bounds depending on the loss of the best action, and in the case of (Lee et al., 2020b), the average loss of a revealing set for the best action, in the weakly observable setting. Lykouris et al. (2018) achieve their regret bounds through a black-box reduction from general small-loss bound algorithms. In the fixed, observable feedback graph setting their bounds either scale sub-optimally with the value of the smallest loss across time or scales with the clique-partition number of the graph, instead of the independence number. Lee et al. (2020b) improve on such results by recovering optimal dependence on both the small loss quantity and independence number. Their results can be extended to time-varying feedback graphs and weakly observable graphs.

We note that online learning with feedback graphs has also been studied in the setting of stochastic losses by numerous works (Caron et al., 2012; Buccapatnam et al., 2014a; Wu et al., 2015a,b; Tossou et al., 2017; Liu et al., 2018), however, we chose not to discuss these works here as our focus is on the adversarial case. For more extensive discussion on bandits with graph feedback we refer the reader to the work of Valko (2016).

The main regret upper bound of this chapter can be summarized as follows.

Theorem 2.1.1. *There exists an algorithm for the setting of policy regret minimization with partial feedback provided by a graph G with regret bounded by $\tilde{O}((m\gamma(G))^{1/3}T^{2/3})$, where $\gamma(G)$ is the domination number of the graph and m is the memory of the adversary.*

2.2 Regret minimization in the presence of switching costs and policy regret

In Section 1.6.2 we mentioned that policy regret minimization for memory-bounded adversaries is equivalent to regret minimization in the presence of switching costs. The regret for the switching costs problem is defined as follows

$$R_S(T) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a) + \chi_{a_t \neq a_{t-1}}\right], \tag{2.1}$$

where χ_A is the characteristic function of the event A.

Let us formally argue the equivalence between m-memory bounded adversaries (for $m > 1$) and the switching cost problem. First notice that an m-memory bounded adversary for $m > 1$ can model a switching cost so policy regret minimization is at least as hard as minimizing $R_S(T)$. Suppose that we have some algorithm $\mathbb{A}$ which minimizes $R_S(T)$ and works with losses $\widehat{\ell}_t : \mathcal{A} \to [0, 1]$. The following reduction solves the policy regret problem. We split the stream of T losses into mini-batches of size m such that $\widehat{\ell}_t(\cdot) = \frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(\cdot)$. Now we would simply feed the sequence $(\widehat{\ell}_t)_{t=1}^{T/m}$ to Algorithm $\mathbb{A}$ if it were not for intervals of length m at which there is a switched action, that is $a_t \neq a_{t+1}$. Suppose that between the t-th mini-batch and the $t+1$-st mini-batch Algorithm $\mathbb{A}$ decides to switch actions so that $a_t \neq a_{t+1}$. In this case no additional feedback is available for $\widehat{\ell}_{t+1}(a_{t+1})$ and the algorithm can not proceed as normal. To fix this minor problem, the provided feedback to the algorithm is that the loss of all actions in $\mathcal{A}$ is 0. This modification can not occur more times than the number of switches Algorithm $\mathbb{A}$ does. Pseudocode for the above algorithm can be found in Algorithm 3.

Algorithm 3: Policy regret with side observations

```
Input: Parameters required by Algorithm $\mathbb{A}$ and memory $m$
Output: Action sequence $(a_t)_t$.
Initialize Algorithm $\mathbb{A}$.
for $t = 1, \dots, \lfloor T/m \rfloor$ do
    Receive action $a_t$ from Algorithm $\mathbb{A}$ and play it for the next $m$ rounds.
    if $a_{t-1} == a_t$ then
        Observe mini-batched loss $\widehat{\ell}_t(a_t) = \frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(a_t)$ (and additional side
        observations). Feed mini-batched loss (and additional side observations) to
        Algorithm $\mathbb{A}$.
    else
        Set $\widehat{\ell}_t = 0$ and feed losses to Algorithm $\mathbb{A}$.
    end if
end for
```

Algorithm 3 enjoys the following regret guarantee.

Theorem 2.2.1. *The Policy regret of Algorithm 3 is bounded by $mR_S(T/m)$.*

Proof of Theorem 2.2.1. The regret of Algorithm $\mathbb{A}$ is bounded as

$$\mathbb{E}\left[\sum_{t=1}^{T/m} \widehat{\ell}_t(a_t) - \sum_{t=1}^{T/m} \widehat{\ell}_t(a) + \sum_{t=1}^{T/m} \chi_{a_{t-1}\neq a_t}\right] \leq \tilde{R}_S(T),$$

for any action a. On the other hand we have

$$\mathbb{E}\left[\sum_{t=1}^{T/m} \widehat{\ell}_t(a_t) - \sum_{t=1}^{T/m} \widehat{\ell}_t(a)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T/m} \widehat{\ell}_t(a_t) - \sum_{t=1}^{T/m} \frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(a)\right]$$
$$=\mathbb{E}\left[\sum_{t=1}^{T/m} \frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(a_t) - \sum_{t=1}^{T/m} \frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(a) - \sum_{t=1}^{T/m} \chi_{a_{t-1}\neq a_t}\frac{1}{m}\sum_{j=1}^{m} \ell_{(t-1)m+j}(a_t)\right].$$

Combined with the regret bound, the above implies

$$\frac{1}{m}\mathbb{E}[R(T)] \leq R_S(T/m) + \mathbb{E}\left[\sum_{t=1}^{T/m} \chi_{a_{t-1}\neq a_t}\right]. \tag{2.2}$$

The second term in the right hand side bounded by the number of switches bound number of switches and hence the regret bound of Algorithm $\mathbb{A}$ as

$$\mathbb{E}\left[\sum_{t=1}^{T/m} \chi_{a_{t-1}\neq a_t}\right] \leq R_S(T/m).$$

Multiplying Inequality 2.2 by m on both sides finishes the proof. □

Thus if we have an algorithm with bounded $R_S(T)$ we also have a reasonable policy regret-minimizing algorithm for bounded memory adversaries. The standard bounds for $R_S(T)$ scale as $O(T^{2/3})$ and hence Algorithm 3 enjoys a $O(m^{1/3}T^{2/3})$ regret bound in that case which is meaningful for all $m = o(T)$.

2.3 Bandits with feedback graphs and switching Costs

The reduction in Section 2.2 allows us to focus on the problem of minimizing the switching costs regret $R_S(T)$. For the rest of this chapter we are only going to consider minimizing $R_S(T)$, or deriving information theoretic lower bounds on $R_S(T)$.

The only work we are familiar with, which studies both bandits with switching costs and side information is that of Rangi and Franceschetti (2019). The authors propose two algorithms for time-varying feedback graphs in the uninformed setting. When reduced to the fixed feedback graph setting, their regret bound becomes $\tilde{O}(\alpha(G)^{\frac{1}{3}}T^{\frac{2}{3}})$. We note that, in the informed setting with a fixed feedback graph, this bound can be achieved by applying the mini-batching technique of Arora et al. (2012a) to the EXP3-SET algorithm of Alon et al. (2013).

The main contributions outlined in this chapter are two-fold. First, we propose two algorithms for online learning in the informed setting with a fixed feedback graph G and switching costs. Our best algorithm admits a pseudo-regret bound in $\tilde{O}(\gamma(G)^{\frac{1}{3}}T^{\frac{2}{3}})$, where $\gamma(G)$ is the domination number of G. We note that the domination number $\gamma(G)$ can be substantially smaller than the independence number $\alpha(G)$ and therefore that our algorithm significantly improves upon previous work by Rangi and Franceschetti (2019) in the informed setting. We also extend our results to achieve a policy regret bound in $\tilde{O}(\gamma(G)^{\frac{1}{3}}T^{\frac{2}{3}})$ when partial counterfactual feedback is available. The $\tilde{O}(\gamma(G)^{\frac{1}{3}}T^{\frac{2}{3}})$ regret bound in the switching costs setting might seem at odds with a lower bound stated by Rangi and Franceschetti (2019). However, the lower bound of Rangi and Franceschetti (2019) can be shown to be technically inaccurate. Our second main contribution is a lower bound in $\tilde{\Omega}(T^{\frac{2}{3}})$ for any non-complete feedback graph. We also extend this lower bound to $\tilde{\Omega}(\gamma(G)^{\frac{1}{3}}T^{\frac{2}{3}})$ for a class of feedback graphs that we will describe in detail. We show a lower bound for the setting of evolving feedback graphs, matching the originally stated lower bound in (Rangi and Franceschetti, 2019).

2.3.1 Problem setup and notation

The main quantity of interest is going to be the switching cost regret $R_S(T)$. We assume that the player has access to an undirected graph $G = (\mathcal{A}, E)$, which determines which expert losses can be observed at each round. The vertex set $\mathcal{A}$ is the set of experts (or actions) and the graph specifies that, if at round t the player selects action a_t, then, the losses of all experts whose vertices are adjacent to that of a_t can be observed: $\ell_t(a)$ for $a \in N(a_t)$, where $N(a_t)$ denotes the neighborhood of a_t in G defined for any $u \in \mathcal{A}$ by: $N(u) = \{v \colon (u, v) \in E\}$. We will denote by $deg(u) = |N(u)|$ the degree of $u \in \mathcal{A}$ in graph G. We assume that G admits a self-loop at every vertex, which implies that the player can at least observe the loss of their own action (bandit information). In all our figures, self-loops are omitted for the sake of simplicity.

We assume that the feedback graph is available to the player at the beginning of the game (***informed setting***). The ***independence number*** of G is the size of a ***maximum independent set*** in G and is denoted by $\alpha(G)$. The ***domination number*** of G is the size of a ***minimum dominating set*** and is denoted by $\gamma(G)$. The following inequality holds for all graphs G: $\gamma(G) \leq \alpha(G)$ (Bollobás and Cockayne, 1979; Goddard and Henning, 2013). In general, $\gamma(G)$ can be substantially smaller than $\alpha(G)$, with $\gamma(G) = 1$ and $\alpha(G) = |\mathcal{A}| - 1$ in some cases. We note that all our results can be straightforwardly extended to the case of directed graphs.

2.3.2 An adaptive mini-batch algorithm

In this section, we describe an algorithm for online learning with switching costs, using adaptive mini-batches.

The standard exploration versus exploitation dilemma in the bandit setting is further complicated in the presence of a feedback graph: if a poor action reveals the losses of all other actions, do we play the poor action? The lower bound construction of Mannor and Shamir (2011) suggests that we should not, since it might be better to just switch between the other actions.

Adding switching costs, however, modifies the price of exploration and the lower bound argument of Mannor and Shamir (2011) no longer holds. It is in fact possible to show that EXP3 and its graph feedback variants switch too often in the presence of two good actions, thereby incurring $\Omega(T)$ regret, due to the switching costs. One way to deal with the switching costs problem is to adapt the fixed mini-batch technique of Arora et al. (2012a). That technique, however, treats all actions equally while, in the presence of switching costs, actions that provide additional information are more valuable.

We deal with the issues just discussed by adopting the idea that the mini-batch sizes could depend both on how favorable an action is and how much information an action provides about good actions.

2.3.2.1 Algorithm for Star Graphs

We start by studying a simple feedback graph case in which one action is adjacent to all other actions with none of these other actions admitting other neighbors. For an example see Figure 2-1.

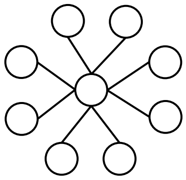


Figure 2-1: Example of a star graph.

We call such graphs ***star graphs*** and we refer to the action adjacent to all other actions as the ***revealing action***. The revealing action is denoted by r. Since only the revealing action can convey additional information about other actions, we will select our mini-batch size to be proportional to the quality of this action. Also, to prevent our algorithm from switching between two non-revealing actions too often, we will simply disallow that and allow switching only between the revealing action and a non-revealing action. Finally, we will disregard any feedback a non-revealing action provides us. This simplifies the analysis of the regret of our algorithm. The pseudocode of the algorithm is given in Algorithm 4.

The following intuition guides the design of our algorithm and its analysis. We need to visit the revealing action sufficiently often to derive information about all

Algorithm 4: Algorithm for star graphs

```
Input: Star graph $G(\mathcal{A}, E)$, learning rates $(\eta_t)$, exploration rate $\beta \in [0,1]$, maximum
    mini-batch $\tau$.
Output: Action sequence $(a_t)_{t=1}^{T}$.
$q_1 = \frac{\mathbf{1}}{|\mathcal{A}|}$.
while $\sum_t \lfloor \tau_t \rfloor \le T$ do
    $p_t = (1-\beta) q_t + \beta \delta(r)$        % $\delta(r)$ is the Dirac distribution on $r$
    Draw $a_t \sim p_t$, set $\tau_t = p_t(r)\tau$
    if $a_{t-1} \neq r$ and $a_t \neq r$ then
        Set $a_t = a_{t-1}$
    end if
    Play $a_t$ for the next $\lfloor \tau_t \rfloor$ iterations
    Set $\widehat{\ell}_t(i) = \sum_{j=t}^{t+\lfloor \tau_t \rfloor - 1} \mathbb{I}(a_t = r) \frac{\ell_j(i)}{p_t(r)}$
    For all $i \in \mathcal{A}$, $q_{t+1}(i) = \frac{q_t(i) \exp\left(-\eta_t \widehat{\ell}_t(i)\right)}{\sum_{j \in \mathcal{A}} q_t(j) \exp\left(-\eta_t \widehat{\ell}_t(j)\right)}$
    $t = t + 1$
end while
```

other actions, which is determined by the explicit exploration factor β. If r is a good action, our regret will not be too large if we visit it often and spent a large amount of time in it. On the other hand if r is poor, then the algorithm should not sample it often and, when it does, it should not spend too much time there. Disallowing the algorithm to directly switch between non-revealing actions also prevents it from switching between two good non-revealing actions too often. The only remaining question is: do we observe enough information about each action to be able to devise a low regret strategy? The following regret guarantee provides a precise positive response.

Theorem 2.3.1. *Suppose that the inequality $\mathbb{E}[\ell_t^2(i)] \le \rho$ holds for all $t \le T$ and all $i \in \mathcal{A}$, for some ρ and $\beta \ge \frac{1}{\tau}$. Then, for any action $a \in \mathcal{A}$, Algorithm 4 admits the following guarantee:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] \le \frac{\log(|\mathcal{A}|)}{\eta} + T\eta\tau\rho + T\beta.$$

Furthermore, the algorithm does not switch more than $2\frac{T}{\tau}$ times, in expectation.

Proof sketch. The complete proof can be found in Section 2.4. The key elements of the proof are as follows. First we analyze the standard regret of an algorithm which excludes lines 5 and 6 from Algorithm 4. The analysis follows standard arguments for bounding the regret of the Exp3 algorithm. Next, we show that the analyzed algorithm will follow the same trajectory as Algorithm 4. Finally, we show that Algorithm 4 will not switch between the revealing action r and another action too often due to our choice of τ_t. In particular, either τ_t is large when we have high probability to sample r and hence we spend a large number of iterations at r or the probability to visit r is small and the algorithm continues playing the same non-revealing action. □

The exploration parameter β is needed to ensure that $\tau_t = p_t(r)\tau \geq 1$, so that at every iteration of the while loop Algorithm 4 plays at least one action. The bound assumed on the second moment $\mathbb{E}[\ell_t^2(i)]$ might seem unusual since in the adversarial setting we do not assume a randomization of the losses. For now, the reader can just assume that this is a bound on the squared loss, that is, $\ell_t^2(i) \leq \rho$. The role of this expectation and the source of the randomness will become clear in Section 2.3.2.3. We note that the star graph admits independence number $\alpha(G) = |\mathcal{A}| - 1$ and domination number $\gamma(G) = 1$. In this case, the algorithms of Rangi and Franceschetti (2019) and variants of the mini-batching algorithm only guarantee a regret bound of the order $\tilde{O}(\alpha(G)^{\frac{1}{3}}T^{\frac{2}{3}})$, while Algorithm 4 guarantees a regret bound of the order $\tilde{O}(T^{\frac{2}{3}})$ when we set $\eta = 1/T^{\frac{2}{3}}$, $\tau = T^{\frac{2}{3}}$, and $\beta = 1/T^{\frac{1}{3}}$.

2.3.2.2 Algorithm for General Feedback Graphs

We now extend Algorithm 4 to handle arbitrary feedback graphs. The pseudocode of this more general algorithm is given in Algorithm 5.

The first step of Algorithm 5 consists of computing an approximate minimum dominating set for G using the Greedy Set Cover algorithm (Chvatal, 1979). The Greedy Set Cover algorithm naturally partitions G into disjoint star graphs with

Algorithm 5: Algorithm for general feedback graphs

```
Input: Graph $G(\mathcal{A}, E)$, learning rates $(\eta_t)$, exploration rate $\beta \in [0,1]$, maximum
    mini-batch $\tau$.
Output: Action sequence $(a_t)_t$.
Compute an approximate dominating set $R$
$q_1 \equiv Unif(\mathcal{A}), u \equiv Unif(R)$
while $\sum_t \tau_t \leq T$ do
    $p_t = (1-\beta)q_t + \beta u$.
    Draw $i \sim p_t$, set $\tau_t = p_t(r_i)\tau$, where $r_i$ is the dominating vertex for $i$ and set
    $a_t = i$.
    if $a_{t-1} \notin R$ and $a_t \notin R$ then
        Set $a_t = a_{t-1}$
    end if
    Play $a_t$ for the next $\lfloor \tau_t \rfloor$ iterations.
    Set $\widehat{\ell}_t(i) = \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r_i)\frac{\ell_j(i)}{p_t(r_i)}$.
    For all $i \in \mathcal{A}$, $q_{t+1}(i) = \frac{q_t(i)\exp\left(-\eta_t \widehat{\ell}_t(i)\right)}{\sum_{j \in \mathcal{A}} q_t(j)\exp\left(-\eta_t \widehat{\ell}_t(j)\right)}$.
    $t = t + 1$.
end while
```

revealing actions/vertices in the dominating set R. Next, Algorithm 5 associates with each star-graph its revealing arm $r \in R$. The mini-batch size at time t now depends on the probability $p_t(r)$ of sampling a revealing action r, as in Algorithm 4. There are several key differences, however, that we now point out. Unlike Algorithm 4, the mini-batch size can change between rounds even if the action remains fixed. This occurs when the newly sampled action is associated with a new revealing action in R, however, it is different from the prior revealing action. The above difference introduces some complications, because τ_t conditioned on all prior actions $a_{1:t-1}$ is still a random variable, while it is a deterministic in Algorithm 4. We also allow switches between any action and any vertex $r \in R$. This might seem to be a peculiar choice. For example, allowing only switches within each star-graph in the partition and only between revealing vertices seems more natural. Allowing switches between any vertex and any revealing action benefits exploration while still being sufficient for controlling the number of switches. If we further constrain the number of switches by using the more natural approach, it is possible that not enough information is received about each action, leading to worse regret guarantees. We leave the investigation of such more natural approaches to future work. Algorithm 5 admits the following regret bound.

Theorem 2.3.2. *For any $\beta \geq \frac{|R|}{\tau}$ The expected regret of Algorithm 5 is*

$$\frac{\log(|\mathcal{A}|)}{\eta} + \eta\tau T + \beta T.$$

Further, if the algorithm is augmented similar to Algorithm 8, then it will switch between actions at most $\frac{2T|R|}{\tau}$ times.

Setting $\eta = 1/(|R|^{\frac{1}{3}}T^{\frac{2}{3}})$, $\tau = |R|^{\frac{2}{3}}T^{\frac{1}{3}}$ and $\beta = |R|^{\frac{1}{3}}/T^{\frac{1}{3}}$, recovers a pseudo-regret bound of $\tilde{O}(|R|^{\frac{1}{3}}T^{\frac{2}{3}})$, with an expected number of switches bounded by $2|R|^{\frac{1}{3}}T^{\frac{2}{3}}$. We note that $|R| = O(\gamma(G)\log(|\mathcal{A}|))$ and thus the regret bound of our algorithm scales like $\gamma(G)^{\frac{1}{3}}$. Further, this is a strict improvement over the results of Rangi and Franceschetti (2019) as their result shows a scaling of $\alpha(G)^{\frac{1}{3}}$. The proof of Theorem 2.3.2 follows the ideas in the proof of Theorem 2.3.1 while carefully handling the construction the unbiased estimators of the losses. The proof can be found in Section 2.4.

2.3.2.3 Corralling Star Graph Algorithms

An alternative natural method to tackle the general feedback graph problem is to use the recent corralling algorithm of Agarwal et al. (2016). In this section, we describe that technique, even though it does not seem to achieve an optimal rate. Here too, the first step consists of computing an approximate minimum dominating set. Next, we initialize an instance of Algorithm 4 for each star graph. Finally, we combine all of the star graph algorithms via a mini-batched version of the corralling algorithm of Agarwal et al. (2016). Mini-batching is necessary to avoid switching between star graph algorithms too often. The pseudocode of this algorithm is given in

Algorithm 6: Corralling star-graph algorithms

```
Input: Feedback graph $G(\mathcal{A}, E)$, learning rate $\eta$, mini-batch size $\tau$
Output: Action sequence $(a_t)_{t=1}^T$.
Compute an approximate minimum dominating set $R$ and initialize $|R|$ base star-graph algorithms, $B_1, B_2, \ldots, B_{|R|}$, with step size $\frac{\eta'}{2|R|}$, mini-batch size $\tau$ and exploration rate $\frac{1}{\tau}$ (Algorithm 4).
$T' = \frac{T}{\tau}$, $\beta = \frac{1}{T'}$, $\tilde{\beta} = \exp\left(\frac{1}{\log(T)}\right)$, $\eta_{1,i} = \eta$, $\rho_{1,i} = 2|R|$ for all $i \in [|R|]$, $q_1 = p_1 = \frac{\mathbf{1}}{|R|}$
for $t = 1, \ldots, T'$ do
    Draw $i_t \sim p_t$
    for $j_t = (t-1)\tau + 1, \ldots, (t-1)\tau + \tau$ do
        Receive action $a_{j_t}^i$ from $B_i$ for all $i \in [|R|]$.
        Set $a_{j_t} = a_{j_t}^{i_t}$, play $a_{j_t}$ and observe loss $\ell_{j_t}(a_{j_t})$.
        Send $\frac{\ell_{j_t}(a_{j_t})}{p_t(i_t)} \mathbb{I}\{i = i_t\}$ as loss to algorithm $B_i$ for all $i \in [|R|]$.
        Update $\widehat{\ell}_t(i) = \widehat{\ell}_t(i) + \frac{1}{\tau} \frac{\ell_{j_t}(a_{j_t})}{p_t(i_t)} \mathbb{I}\{i = i_t\}$.
    end for
    Update $q_{t+1} =$ Algorithm 7$(q_t, \widehat{\ell}_t, \eta_t)$.
    Set $p_{t+1} = (1-\beta) q_{t+1} + \beta \frac{\mathbf{1}}{|R|}$.
    for $i = 1, \ldots, |R|$ do
        if $\frac{1}{p_t(i)} > \rho_{t,i}$ then
            Set $\rho_{t+1,i} = \frac{2}{p_t(i)}$, $\eta_{t+1,i} = \tilde{\beta} \eta_{t,i}$ and restart $i$-th star-graph algorithm, with updated step-size $\frac{\eta'}{\rho_{t+1,i}}$
        else
            Set $\rho_{t+1,i} = \rho_{t,i}$, $\eta_{t+1,i} = \eta_{t,i}$.
        end if
    end for
end for
```

Algorithm 6. Since during each mini-batch we sample a single star graph algorithm, we need to construct appropriate unbiased estimators of the losses ℓ_{j_t}, which we feed back to the sampled star graph algorithm. The bound on the second moment of these estimators is exactly what Theorem 2.3.1 requires. Our algorithm admits the following guarantees.

Theorem 2.3.3. *Let $\tau = T^{\frac{1}{3}}/|R|^{\frac{1}{4}}, \eta = |R|^{\frac{1}{4}}/(40c \log(T') T^{\frac{1}{3}} \log(|\mathcal{A}|))$, and $\eta' = 1/T^{\frac{2}{3}}$, where c is a constant independent of T, τ, $|\mathcal{A}|$ and $|R|$. Then, for any $a \in \mathcal{A}$,*

Algorithm 7: Log-Barrier-OMD(q_t, ℓ_t, η_t)

```
Input: Previous distribution $q_t$, loss vector $\ell_t$, learning rate vector $\eta_t$.
Output: Updated distribution $q_{t+1}$.
Find $\lambda \in [\min_i \ell_t(i), \max_i \ell_t(i)]$ such that $\sum_{i=1}^{|R|} \frac{1}{\frac{1}{q_t(i)} + \eta_{t,i}(\ell_t(i) - \lambda)} = 1$
Return $q_{t+1}$ such that $\frac{1}{q_{t+1}(i)} = \frac{1}{q_t(i)} + \eta_{t,i}(\ell_t(i) - \lambda)$.
```

the following inequality holds for Algorithm 6:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] \leq \tilde{O}\left(\sqrt{|R|}\, T^{\frac{2}{3}}\right).$$

Furthermore, the expected number of switches of the algorithm is bounded by $T^{\frac{2}{3}}|R|^{\frac{1}{3}}$.

Sketch of proof. The basic idea behind the proof-sketch is to first partition the graph G into star graphs using the greedy algorithm for finding a dominating set. Next, we know that running Algorithm 4 on each star graph comes with an $R_S(T) \leq \tilde{O}(T^{2/3})$ regret guarantee. We could directly use the corralling algorithm on top of the star graph algorithms, however, we still need to make sure that the corralling algorithm does not switch too often. To this end we just use the reduction outlined in Arora et al. (2012a) to obtain a corralling algorithm with bounded switches. Unfortunately, this reduction leads to the sub-optimal dependence on the size of the dominating set. □

This bound is suboptimal compared to the $\gamma(G)^{\frac{1}{3}}$-dependency achieved by Algorithm 5. We conjecture that this gap is an artifact of the analysis of the corralling algorithm of Agarwal et al. (2016). However, we were unable to improve on the current regret bound by simply corralling.

2.4 Detailed proofs from Section 2.3.2

2.4.1 Adaptive Mini-batching for Star Graphs

The proof of Theorem 2.3.1 begins by considering a slightly modified version of Algorithm 4. In particular we remove lines 5 through 7 which disallow switching between non-revealing actions. This intuitively should not change the policy which Algorithm 4 produces as such switches do not provide any new information to the algorithm. For convenience of the reader we give the pseudo-code of the modified algorithm in Algorithm 8, where the lines in red are commented out and are not part of the algorithm.

Algorithm 8: Algorithm for star graphs (modified)

```
Input: Star graph $G(\mathcal{A}, E)$, learning rate sequence $(\eta_t)$, exploration rate $\beta \in [0, 1]$,
    maximum mini-batch $\tau$.
Output: Action sequence $(a_t)_t$.
$q_1 \equiv Unif(\mathcal{A})$.
while $\sum_t \tau_t \leq T$ do
    $p_t = (1 - \beta)q_t + \beta\delta(r)$.
    Draw $a_t \sim p_t$, set $\tau_t = p_t(r)\tau$.
    if $a_{t-1} \neq r$ and $a_t \neq r$ then
        Set $a_t = a_{t-1}$
    end if
    Play $a_t$ for the next $\lfloor \tau_t \rfloor$ iterations.
    Set
        $$\widehat{\ell}_t(i) = \sum_{j=t}^{t+\lfloor \tau_t \rfloor - 1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}.$$
    For all $i \in \mathcal{A}$, $q_{t+1}(i) = \frac{q_t(i)\exp\left(-\eta_t \widehat{\ell}_t(i)\right)}{\sum_{j \in \mathcal{A}} q_t(j)\exp\left(-\eta_t \widehat{\ell}_t(j)\right)}$.
    $t = t + 1$.
end while
```

Algorithm 8 comes with the following regret guarantee.

Theorem 2.4.1. *Suppose that for all $t \leq T$ and all $i \in \mathcal{A}$ it holds that $\mathbb{E}[\ell_t(i)^2] \leq \rho$ and $\beta \geq \frac{1}{\tau}$. Then Algorithm 8 produces an action sequence $(a_t)_{t=1}^T$ satisfying:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] \leq \frac{\log\left(|\mathcal{A}|\right)}{\eta} + T\eta\tau\rho + T\beta,$$

for any $a \in \mathcal{A}$.

Proof. Since $\beta \geq \frac{1}{\tau}$, this implies that $\lfloor \tau_t \rfloor \geq 1$ and the algorithm terminates, producing an action sequence $(a_t)_{t=1}^T$. Let i_t^* be the best action at time t and let $L_{t,*} = \sum_{s=1}^t \widehat{\ell}_s(i_t^*)$.

Let $w_t(i) = \exp\left(-\eta \sum_{j=1}^{t-1} \widehat{\ell}_j(i)\right)$ and $W_t = \sum_{i \in \mathcal{A}} w_t(i)$. We have

$$\begin{aligned}
\log\left(\frac{W_{t+1}}{w_{t+1}(i^*_{t+1})}\right) - \log\left(\frac{W_t}{w_t(i^*_t)}\right) &= \eta\left(L_{t+1,*} - L_{t,*}\right) \\
&\quad + \log\left(\frac{\sum_{i\in\mathcal{A}} w_t(i) \exp\left(-\eta \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}\right)}{W_t}\right) \\
&= \eta\left(L_{t+1,*} - L_{t,*}\right) \\
&\quad + \log\left(\sum_{i\in\mathcal{A}} q_t(i) \exp\left(-\eta \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}\right)\right) \\
&\le \eta\left(L_{t+1,*} - L_{t,*}\right) - 1 \\
&\quad + \sum_{i\in\mathcal{A}} q_t(i) \exp\left(-\eta \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}\right) \\
&\le \eta\left(L_{t+1,*} - L_{t,*}\right) - \eta\frac{\mathbb{I}(a_t = r)}{p_t(r)} \sum_{i\in\mathcal{A}} q_t(i) \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \ell_j(i) \\
&\quad + \frac{\eta^2}{2}\frac{\mathbb{I}(a_t = r)}{p_t(r)^2} \sum_{i\in\mathcal{A}} q_t(i) \left(\sum_{j=t}^{t+\tau_t-1} \ell_j(i)\right)^2,
\end{aligned}$$

where the first inequality follows from $\log(x) \le x - 1$ for all $x > 0$ and the second inequality follows from $e^{-x} \le 1 - x + x^2/2$ for $x \ge 0$. Rearranging terms in the above and taking expectation we have

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)}\sum_{i\in\mathcal{A}}q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)|a_{1:t-1}\right]\right]$$
$$\leq\frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_t^*)}\right)-\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right)\right]$$
$$+\frac{\eta}{2}\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\sum_{i\in\mathcal{A}}q_t(i)\left(\sum_{j=t}^{t+\tau_t-1}\ell_j(i)\right)^2|a_{1:t-1}\right]\right]+\mathbb{E}[L_{t+1,*}-L_{t,*}]$$
$$\implies$$
$$\mathbb{E}\left[\sum_{i\in\mathcal{A}}q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)\right]\leq\frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_t^*)}\right)-\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right)\right]$$
$$+\frac{\eta}{2}\mathbb{E}\left[\frac{1}{p_t(r)}\sum_{i\in\mathcal{A}}q_t(i)\left(\sum_{j=t}^{t+\tau_t-1}\ell_j(i)\right)^2\right]+\mathbb{E}[L_{t+1,*}-L_{t,*}]$$
$$\implies$$
$$\mathbb{E}\left[\sum_{i\in\mathcal{A}}q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)\right]\leq\frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_t^*)}\right)-\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right)\right]$$
$$+\frac{\eta}{2}\mathbb{E}\left[\frac{1}{p_t(r)}\sum_{i\in\mathcal{A}}q_t(i)\tau_t\sum_{j=t}^{t+\tau_t-1}\ell_j(i)^2\right]+\mathbb{E}[L_{t+1,*}-L_{t,*}]$$
$$\implies$$
$$\mathbb{E}\left[\sum_{i\in\mathcal{A}}q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)\right]\leq\frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_t^*)}\right)-\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right)\right]$$
$$+\frac{\eta}{2}\mathbb{E}\left[\frac{1}{p_t(r)}\sum_{i\in\mathcal{A}}q_t(i)\tau_t\sum_{j=t}^{t+\tau_t-1}\mathbb{E}[\ell_j(i)^2|a_{1:t-1}]\right]+\mathbb{E}[L_{t+1,*}-L_{t,*}]$$
$$\implies$$
$$\mathbb{E}\left[\sum_{i\in\mathcal{A}}q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)\right]\leq\frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_t^*)}\right)-\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right)\right]$$
$$+\frac{\eta}{2}\mathbb{E}\left[\rho\frac{p_t(r)^2\tau^2}{p_t(r)}\sum_{i\in\mathcal{A}}q_t(i)\right]+\mathbb{E}[L_{t+1,*}-L_{t,*}].$$

Notice that $\mathbb{E}[L_{T,*}]=\mathbb{E}[\sum_{t=1}^{T'}\frac{\mathbb{I}(a_t=r)}{p_t(r)}\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i^*)]=\mathbb{E}[\sum_{t=1}^{T'}\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i^*)]$. Summing over $t = 1$ through T and using the fact $\log\left(\frac{W_1}{w_1(i^*)}\right) = \log(|\mathcal{A}|)$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T'}\sum_{i\in\mathcal{A}} q_t(i) \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} (\ell_j(i) - \ell_j(i^*))\right] \leq \frac{\log(|\mathcal{A}|)}{\eta} + \frac{\eta}{2}\tau\mathbb{E}\left[\rho\sum_{t=1}^{T'} p_t(r)\tau\right]$$
$$\leq \frac{\log(|\mathcal{A}|)}{\eta} + T\eta\tau\rho,$$

where T' is the random variable equaling the number of mini-batches. The last inequality in the above follows since $\tau_T \in o(T)$ and from our while loop we know that $\sum_{t=1}^{T'-1}\tau_t \leq T$, thus we can bound $\mathbb{E}[\sum_{t=1}^{T'}\tau_t] \leq 2T$. Notice that the LHS in the above inequality is almost equal to the expected regret of our algorithm. We have $q_t(i) \leq p_t(i) - \beta$ and thus the expected regret is bounded by

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t) - \ell_t(a)\right] \leq \frac{\log(|\mathcal{A}|)}{\eta} + T\eta\tau\rho + T\beta.$$

□

Lemma 2.4.2. *Algorithm 8 switches between a revealing and a non-revealing action at most $\frac{T}{\tau}$ times in expectation.*

Proof. The number of switches can be upper bounded by twice the number of times a_t is equal to r. Thus the expected number of switches is bounded by $\mathbb{E}[\sum_{t=1}^{T'}\mathbb{I}(a_t = r)] = \frac{1}{\tau}\mathbb{E}[\sum_{t=1}^{T'}p_t(r)\tau] = \frac{1}{\tau}\mathbb{E}[\sum_{t=1}^{T'}\tau_t] \leq \frac{2T}{\tau}$. □

To finish the proof of Theorem 2.3.1 we need to verify that the expected regret of Algorithm 8 is the same as the expected regret of Algorithm 4.

Lemma 2.4.3. *Algorithm 8 and Algorithm 4 have the same expected regret bound.*

Proof. Let $(p_t)_{t=1}^T$ be the sequence of random vectors generated by Algorithm 8 and let $(p'_t)_{t=1}^T$ be the sequence of random vectors generated by Algorithm 4. First we show by induction that the distribution of p_t is the same as that of p'_t. The base case is trivial as $p_1 = p'_1$. To see that the induction step holds we just notice that if we condition on p_t either both algorithms update p_{t+1} and p'_{t+1} because action r was sampled, in which case the updates are exactly the same, or both algorithms do not update p_{t+1}, respectively p'_{t+1}. Let a_t and a'_t denote the t-th action of Algorithm 8 and Algorithm 4 respectively. We now show that $\mathbb{E}[\ell_t(a_t)] = \mathbb{E}[\ell_t(a'_t)]$. Let X_t denote the random variable indicating the last time before t in which action r was played by Algorithm 8 and let X'_t be the random variable indicating the last time before t in which action r was played by Algorithm 4. Since X_t is function of $p_1, \ldots, p_{t-1}$ and X'_t is a function of $p'_1, \ldots, p'_{t-1}$, then X_t and X'_t have the same distribution. Now we can write

$$\begin{aligned}
\mathbb{E}[\ell_t(a_t)] &= \sum_{j=1}^{t-1} \mathbb{P}(X_t = j)\mathbb{E}[\ell_t(a_t)|X_t = j] = \sum_{j=1}^{t-1} \mathbb{P}(X_t = j)\mathbb{E}[\sum_{i\in\mathcal{A}} p_t(i)\ell_t(i)|X_t = j] \\
&= \sum_{j=1}^{t-1} \mathbb{P}(X_t = j)\mathbb{E}[\sum_{i\in\mathcal{A}} p_{j+1}(i)\ell_t(i)|X_t = j] \\
&= \sum_{j=1}^{t-1} \mathbb{P}(X_t = j)\mathbb{E}[\sum_{i\in\mathcal{A}} p'_{j+1}(i)\ell_t(i)|X'_t = j] \\
&= \sum_{j=1}^{t-1} \mathbb{P}(X'_t = j)\mathbb{E}[\ell_t(a'_t)|X'_t = j] = \mathbb{E}[\ell_t(a'_t)].
\end{aligned}$$

□

Proof of Theorem 2.3.1. Lemma 2.4.3 together with Theorem 2.4.1 imply the bound

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] \leq \tilde{O}\left(\sqrt{\rho}T^{2/3}\right).$$

Lemma 2.4.2 together with the fact that Algorithm 4 can only switch between the revealing action and non-revealing actions imply the bound on number of switches. □

2.4.2 Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. First note that because of the condition $\beta \geq \frac{|R|}{\tau}$ each of the mini-batches $\lfloor \tau_t \rfloor$ is at least 1, since for any $r \in R$ we have $p_t(r) \geq \frac{\beta}{|R|} \geq \frac{1}{\tau}$, and thus the algorithm will terminate in at most $2T$ iterations. Next, similarly to Lemma 2.4.3, we can analyze the regret of Algorithm 5 by removing lines 6 and 7 when bounding the cumulative loss of the algorithm and then use lines 6 and 7 to guarantee that the algorithm does not switch too often. Let $w_{t+1}(i) = w_t(i)\exp\left(-\eta_t \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r_i)\frac{\ell_j(i)}{p_t(r_i)}\right)$ and $W_t = \sum_{i\in\mathcal{A}} w_t(i)$, so that $q_t(i) = \frac{w_t(i)}{W_t}$. Let $\mathcal{A}_r$ be the subset of actions dominated by the vertex r. Let i_t^* be the best action at time t and let $L_{t,*} = \sum_{s=1}^t \widehat{\ell}_s(i_t^*)$. We consider the difference $\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right) - \log\left(\frac{W_t}{w_t(i_t^*)}\right)$.

$$
\begin{aligned}
&\log\left(\frac{W_{t+1}}{w_{t+1}(i_{t+1}^*)}\right) - \log\left(\frac{W_t}{w_t(i_t^*)}\right) = \eta_t(L_{t+1,*} - L_{t,*}) \\
&+ \log\left(\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\exp\left(-\eta_t \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r_i)\frac{\ell_j(i)}{p_t(r_i)}\right)\right) \\
&\leq \eta_t(L_{t+1,*} - L_{t,*}) - 1 \\
&+ \sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\exp\left(-\eta_t \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r_i)\frac{\ell_j(i)}{p_t(r_i)}\right) \\
&\leq \eta_t(L_{t+1,*} - L_{t,*}) - \eta_t \sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i) \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r_i)\frac{\ell_j(i)}{p_t(r_i)} \\
&+ \frac{\eta_t^2}{2}\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\left(\sum_{j=t}^{t+\tau_t-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}\right)^2,
\end{aligned}
$$

where the first inequality follows from the fact that $\log\left(()\, x\right) \leq x - 1$ for all $x \geq 0$ and the second inequality follows from the fact that $e^{-x} \leq 1 - x + x^2/2$, for all $x \geq 0$. Set $\eta_t = \eta$ and divide both sides by η. Shuffling terms around, taking expectation and noting that if one drops the floor function from the quadratic term it will only get larger we arrive at the following

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i) \sum_{j=t}^{t+\lfloor\tau_t\rfloor-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)} + L_{t+1,*} - L_{t,*}\right] \\
&\leq \frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{W_t}{w_t(i_{r^*}^*)}\right) - \log\left(\frac{W_{t+1}}{w_{t+1}(i_{r^*}^*)}\right)\right] \\
&+ \frac{\eta}{2}\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\left(\sum_{j=t}^{t+\tau_t-1} \mathbb{I}(a_t = r)\frac{\ell_j(i)}{p_t(r)}\right)^2\right].
\end{aligned}
\tag{2.3}
$$

Consider the term on the LHS.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\mathbb{I}(a_t=r)\frac{\ell_j(i)}{p_t(r)}+L_{t+1,*}-L_{t,*}\right]\\
&=\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)+L_{t+1,*}-L_{t,*}\right],
\end{aligned}$$

where in the last inequality we used that $\ell_j(i)\le 1$ for all $i\in\mathcal{A}$. Now we consider the second term on the RHS of the inequality.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\left(\sum_{j=t}^{t+\tau_t-1}\mathbb{I}(a_t=r)\frac{\ell_j(i)}{p_t(r)}\right)^2\right]\\
&=\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\left(\sum_{j=t}^{t+\tau_t-1}\ell_j(i)\right)^2|a_{1:t-1}\right]\right]\\
&\le\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\tau_t^2|a_{1:t-1}\right]\right]
\end{aligned}$$

Consider the term $\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\tau_t^2|a_{1:t-1}\right]$. We have $a_t=r$ with probability $p_t(r)$ and so $\tau_t=p_t(r)\tau$. Otherwise we have $\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\tau_t^2=0$. Thus the RHS is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\left(\sum_{j=t}^{t+\tau_t-1}\mathbb{I}(a_t=r)\frac{\ell_j(i)}{p_t(r)}\right)^2\right]\\
&\le\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\mathbb{E}\left[\frac{\mathbb{I}(a_t=r)}{p_t(r)^2}\tau_t^2|a_{1:t-1}\right]\right]=\mathbb{E}\left[\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)p_t(r)\tau^2\right]\\
&=\tau\mathbb{E}\left[\sum_{r\in R}p_t(r)\tau\mathbb{P}[\tau_t=p_t(r)\tau]\right]=\tau\mathbb{E}[\tau_t].
\end{aligned}$$

Summing the LHS and RHS of Equation 2.3 and using our respective bounds, we get:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}\sum_{i\in\mathcal{A}_r} q_t(i)\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i)-\sum_{j=t}^{t+\lfloor\tau_t\rfloor-1}\ell_j(i^*_{r^*})\right]\\
&\le\frac{\log\left(|\mathcal{A}|\right)}{\eta}+\frac{\eta}{2}\tau\mathbb{E}\left[\sum_{t=1}^{T'}\tau_t\right]\le\frac{\log\left(|\mathcal{A}|\right)}{\eta}+\eta\tau T.
\end{aligned}$$

Next we notice that the LHS is almost the expected regret of the algorithm, except we need to replace $q_t(i)$ by $p_t(i)$. This is done at the cost of an additional βT term, since $q_t(r)\le p_t(r)-\frac{\beta}{|R|}$ for $r\in R$. Finally we upper bound the number of times the algorithm switches by the number of times it samples a revealing arm which is equal to $\mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}\mathbb{I}(a_t=r)\right]$. To bound this term we do the following

$$\begin{aligned}
2T &\geq \mathbb{E}\left[\sum_{t=1}^{T'}\tau_t\right] = \mathbb{E}\left[\sum_{t=1}^{T'}\mathbb{E}\left[\tau_t|p_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}p_t(r)\tau\sum_{i\in\mathcal{A}_r}p_t(i)\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}\tau p_t(r)^2\right] = \tau\mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}p_t(r)^2\right] \geq \frac{\tau}{|R|}\mathbb{E}\left[\sum_{t=1}^{T'}\left(\sum_{r\in R}p_t(r)\right)^2\right] \\
&\geq \frac{\tau}{|R|}\mathbb{E}\left[\sum_{t=1}^{T'}\left(\mathbb{E}\left[\sum_{r\in R}p_t(r)|a_{1:(t-1)}\right]\right)^2\right] = \frac{\tau}{|R|}\mathbb{E}\left[\sum_{t=1}^{T'}\left(\sum_{r\in R}\mathbb{I}(a_t=r)\right)^2\right] \\
&= \frac{\tau}{|R|}\mathbb{E}\left[\sum_{t=1}^{T'}\sum_{r\in R}\mathbb{I}(a_t=r)\right],
\end{aligned}$$

where the second inequality follows from the fact that $\sum_{i\in\mathcal{A}_r}p_t(i)\geq p_t(r)$, the third inequality follows from the fact that $(\sum_{r\in R}p_t(r))^2\leq|R|\sum_{r\in R}p_t(r)^2$ and the fourth inequality follows from Jensen's inequality for conditional expectations. □

2.4.3 Proof of Theorem 2.3.3

We use a mini-batch version of Algorithm 1 in Agarwal et al. (2016) where each of the base algorithms is Algorithm 4. We note that the greedy algorithm for computing an approximate minimum dominating set gives a natural way to partition the feedback graph G into star graphs. In particular, whenever the greedy algorithm adds a vertex v to the dominating set, we create a new instance of the star graph algorithm with revealing vertex v and leaf nodes all neighbors of v which have not already been assigned to a star graph algorithm.

Lemma 2.4.4. *For any $i\in[|R|]$, Algorithm 6 ensures that:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t)-\ell_t(a_t^i)\right]\leq O\left(\frac{\tau|R|\log(T')}{\eta}+T\eta\right)-\mathbb{E}\left[\frac{\tau\rho_{T',i}}{40\eta\log(T')}\right]$$

Proof. From the proof of Lemma 13 in Agarwal et al. (2016) it follows that for any $i\in[|R|]$

$$\sum_{t=1}^{T'}\langle p_t-e_i,\widehat{\ell}_t\rangle\leq O\left(\frac{|R|\log(T')}{\eta}+T'\eta\right)+\sum_{t=1}^{T'}\frac{2\widehat{\ell}_t(a_t)}{T'|R|}-\frac{\rho_{T',i}}{40\eta\log(T')}.$$

Notice that by construction we have $\mathbb{E}[\widehat{\ell}_t(a_t)] = \sum_{i\in[|R|]} \frac{1}{\tau} \sum_{j=t}^{t+\tau-1} \ell_j(a_j^i) \leq |R|$. Also notice that $\mathbb{E}[\langle p_t, \widehat{\ell}_t\rangle] = \mathbb{E}[\frac{1}{\tau}\sum_{j=t}^{t+\tau-1} \ell_j(a_j)]$ and $\mathbb{E}[\widehat{\ell}_t(i)] = \frac{1}{\tau}\sum_{j=t}^{t+\tau-1} \ell_t(a_j^i)$. These imply $\mathbb{E}\left[\sum_{t=1}^{T'} \frac{1}{\tau}\sum_{j=t}^{t+\tau-1} \ell_j(a_j) - \frac{1}{\tau}\sum_{j=t}^{t+\tau-1} \ell_t(a_j^i)\right] \leq O\left(\frac{|R|\log(T')}{\eta} + T'\eta\right) + \sum_{t=1}^{T'} \frac{2\widehat{\ell}_t(a_t)}{T'|R|} - \frac{\rho_{T',i}}{40\eta\log(T')}$. Multiplying by τ and using the fact that $T'\tau = T$ finishes the proof. □

The following theorem from Agarwal et al. (2016) shows that restarting the i-th algorithm in line 16 of Algorithm 6 does not hinder the regret bound by too much.

Theorem 2.4.5 (Theorem 15 (Agarwal et al., 2016)). *Suppose a base algorithm B_i is such that if the loss sequence $(\ell_t)_{t=1}^T$ is replaced by $\ell'_t = \rho_t\ell_t$ such that $\mathbb{E}[\ell'_t] = \ell_t$, its regret bound changes from $R(T)$ to $\mathbb{E}[\rho^\alpha]R(T)$, where $\rho = \max_{t\leq T}\rho_t$. Let $(a_t^i)_{t\leq T}$ be the action sequence generated by B_i ran under Algorithm 6. Then for any action a in the action set of B_i, it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell'_t(a_t^i) - \ell'_t(a)\right] \leq \frac{2^\alpha}{2^\alpha - 1}\mathbb{E}[\rho^\alpha]R(T).$$

Proof of Theorem 2.3.3. For any action $a \in \mathcal{A}$, let i_a be the star-graph algorithm which has a in its actions and let its regret be $R_{i_a}(T)$. Notice that the loss estimators $\ell'_t(i) = \frac{\ell_{t+j}(a_{t+j})}{p_t(i_t)}\mathbb{I}\{i = i_t\}$ we feed the algorithm are such that $\mathbb{E}[\ell'_t(i)^2] \leq \rho_T$. Now Theorem 2.3.1 implies that the condition of Theorem 2.4.5 is satisfied with $\alpha = 1/2$. Thus, Theorem 2.4.5 implies that

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell'_t(a_t) - \ell'_t(a)\right] \leq \sqrt{2}(\sqrt{2}+1)\mathbb{E}[\rho_{T',i_a}^{1/2}]3T^{2/3}\log(|\mathcal{A}|).$$

Combining the above with Lemma 2.4.4 we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(a)\right] \leq O\left(\frac{\tau|R|\log(T')}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\tau\rho_{T',i_a}}{40\eta\log(T')}\right] + 3\sqrt{2}(\sqrt{2}+1)\mathbb{E}[\rho_{T',i_a}^{1/2}]T^{2/3}\log(|\mathcal{A}|)$$

Let $c = 3\sqrt{2}(\sqrt{2}+1)$. We now consider the terms containing ρ_{T',i_a} in the above inequality.

$$c\mathbb{E}[\rho_{T',i_a}^{1/2}]T^{2/3}\log(|\mathcal{A}|) - \mathbb{E}\left[\frac{\tau\rho_{T',i_a}}{40\eta\log(T')}\right] = \mathbb{E}\left[\rho_{T',i_a}^{1/2}\left(cT^{2/3}\log(|\mathcal{A}|) - \frac{\tau\rho_{T',i_a}^{1/2}}{40\eta\log(T')}\right)\right].$$

Set $\tau = \frac{T^{1/3}}{|R|^{1/4}}, \eta = \frac{|R|^{1/4}}{40\log(T')T^{1/3}c\log(|\mathcal{A}|)}$ to get

$$\mathbb{E}\left[\rho_{T',i_a}^{1/2}\left(cT^{2/3}\log\left(|\mathcal{A}|\right) - \frac{\tau\rho_{T',i_a}^{1/2}}{40\eta\log\left(T'\right)}\right)\right] = cT^{2/3}\log\left(|\mathcal{A}|\right)\mathbb{E}\left[\rho_{T',i_a}^{1/2}\left(1 - \frac{\rho_{T',i_a}^{1/2}}{|R|^{1/2}}\right)\right]$$
$$\leq c\sqrt{|R|}\log\left(|\mathcal{A}|\right)T^{2/3}.$$

Plugging in the the values of η and τ in the rest of the bound finishes the regret bound.

The number of switches is bounded from the fact that Algorithm 6 can switch between star-graph algorithms at most $T^{2/3}|R|^{1/3}$ times and Lemma 2.4.2. □

2.5 Policy regret bound

For deriving Theorem 2.1.1 we assume that we are provided with a feedback graph for losses with memory m. We restrict the feedback graph to only have vertices for repeated m-tuples of actions in $\mathcal{A}$. In particular we can only observe additional feedback for losses of the type $\ell_t(a, a, \ldots, a)$, where $a \in \mathcal{A}$. The algorithm for this setting is based on Algorithm 5. The feedback graph we provide to our policy regret algorithm is the same as for the m-memory bounded losses, however, each m-tuple vertex is replaced by a copy of a single action e.g. the vertex $(a, \ldots, a)$ is replaced by a. Next we split the stream of T losses into mini-batches of size m such that $\widehat{\ell}_t(\cdot) = \frac{1}{m}\sum_{j=1}^{m}\ell_{(t-1)m+j}(\cdot)$. Now we would simply feed the sequence $(\widehat{\ell}_t)_{t=1}^{T/m}$ to Algorithm 5 if it were not for the constraint on the additional feedback. Suppose that between the t-th mini-batch and the $t+1$-st mini-batch Algorithm 5 decides to switch actions so that $a_t \neq a_{t+1}$. In this case no additional feedback is available for $\widehat{\ell}_{t+1}(a_{t+1})$ and the algorithm can not proceed as normal. To fix this minor problem, the provided feedback to Algorithm 5 is that the loss of action a_{t+1} was 0 and all actions adjacent to a_{t+1} also incurred 0 loss. This modification can not occur more times than the number of switches Algorithm 5 does. Since the expected number of switches is bounded by $O(\gamma(G)^{1/3}T^{2/3})$, intuitively the modification becomes benign to the total expected regret. Formally, we use Algorithm 5 as Algorithm $\mathbb{A}$ in the reduction provided by Algorithm 3. Theorem 2.2.1 now implies Theorem 2.1.1 because $R_S(T) = \tilde{O}(\gamma(G)^{1/3}T^{2/3})$ (by Theorem 2.3.2) and the reduction (Theorem 2.2.1) guarantees that $P(T) = m\tilde{O}(\gamma(G)^{1/3}(T/m)^{2/3} = \tilde{O}((m\gamma(G))^{1/3}T^{2/3})$.

2.6 Lower bounds

The main tool for constructing lower bounds when switching costs are involved is the stochastic process constructed by Dekel et al. (2014). The crux of the proof consists of a carefully designed multi-scale random walk. The two characteristics of this random walk are its depth and its width. At time t, the depth of the walk is the number of previous rounds on which the value of the current round depends. The width of the walk measures how far apart two rounds that depend on each other are in time. The loss of each action is equal to the value of the random walk at each time step, and the loss of the best action is slightly better by a small positive constant. The depth of the process controls how well the losses concentrate in the interval $[0, 1]$[1]. The width of the walk controls the variance between losses of different actions and ensures it is impossible to gain information about the best action, unless one switches between different actions.

2.6.1 Lower bound for non-complete graphs

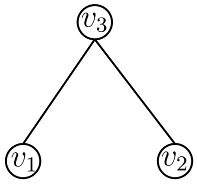


Figure 2-2: Feedback graph for switching costs

We first verify that the dependence on the time horizon cannot be improved from $T^{\frac{2}{3}}$ for any feedback graph in which there is at least one edge missing, that is, in which there exist two vertices that do not reveal information about each other. Without loss of generality,

[1] Technically, the losses are always clipped between $[0, 1]$.

assume that the two vertices not joined by an edge are v_1 and v_2. Take any vertex that is a shared neighbor and denote this vertex by v_3 (see Figure 2-2 for an example). We set the loss for action v_3 and all other vertices to be equal to one. We now focus the discussion on the subgraph with vertices $\{v_1, v_2, v_3\}$. The losses of actions v_1 and v_2 are set according to the construction in (Dekel et al., 2014). Since $\{v_1, v_2\}$ forms an independent set, the player would need to switch between these vertices to gain information about the best action. This is also what the lower bound proof of Rangi and Franceschetti (2019) is based upon. However, it is important to realize that the construction in Dekel et al. (2014) also allows for gaining information about the best action if its loss is revealed together with some other loss constructed from the stochastic process. In that case, playing vertex v_3 would provide such information. This is a key property which Rangi and Franceschetti (2019) seem to have missed in their lower bound proof. We discuss this mistake carefully and provide a lower bound matching what the authors claim in the ***uninformed*** setting in Section 2.6.3. Our discussion suggests that we should set the price for revealing information about multiple actions according to the switching cost and this is why the losses of all vertices outside of the independent set are equal to one. We note that the losses of the best action are much smaller than one sufficiently often, so that enough instantaneous regret is incurred when pulling action v_3. Our main result follows.

Theorem 2.6.1. *For any non-complete feedback graph G, there exists a sequence of losses on which any algorithm $\mathcal{A}$ in the informed setting incurs expected regret at least*

$$R_S(T) \geq \Omega\left(\frac{T^{\frac{2}{3}}}{\log(T)}\right).$$

Before proceeding with the proof of Theorem 2.6.1, we introduce the stochastic process defined in Dekel et al. (2014).

Stochastic process definition. We denote by $\xi_{1:T}$ a sequence of i.i.d. zero-mean Gaussian random variables with variance σ^2 and $\rho : [T] \to \{0\} \bigcup [T]$ the parent function, which assigns to $t \in [T]$ a parent $\rho(t) \in [T]$ with $\rho(t) < t$. The stochastic process W_t associated with $\rho(t)$ is defined as

$$\begin{aligned} W_0 &= 0 \\ W_t &= W_{\rho(t)} + \xi_t. \end{aligned} \tag{2.4}$$

The set of ancestors of t is the set $\rho^*(t) = \rho^*(\rho(t)) \bigcup \{\rho(t)\}$ with $\rho^*(0) = \{\}$. The depth of ρ is $d(\rho) = \max_{t \in [T]} |\rho^*(t)|$. The cut of ρ is $cut(t) = \{s \in [T] : \rho(s) < t \leq s\}$ i.e. the set of rounds which are separated from their parent by t. The width of ρ is defined as $\omega(\rho) = \max_{t \in [T]} |cut(t)|$. The specific random walk which Dekel et al. (2014) consider has both depth and width logarithmic in T. In particular the parent function is defined as

$$\rho(t) = t - 2^{\delta(t)}, \text{where }, \delta(t) = \max\{i \geq 0 : t \equiv 0 \bmod 2^i\} \tag{2.5}$$

Let us consider two examples of a stochastic processes defined by Equation 2.4. The first one is just setting $\rho(t) = 0$, so that W_t is just a standard Gaussian variable. The width of this process is just T and its depth is 1. While we have good concentration guarantees over the maximum value of W_t uniformly over all $t \in [T]$, which is important for controlling the losses, it is very easy to gain information about actions 1 and 2 without switching. Indeed one can just first play 1 for a sufficient number of iteration and then play 2 for fixed number of iterations to be able, with high probability, to distinguish between the two losses. Now consider a Gaussian random walk where $\rho(t) = t - 1$. In this case the cut is 1 but the depth is T. It turns out that to distinguish between two processes with small width, we require that we observe both the processes at the same time (or times differing by a small amount). This is intuitively because of the large drift of the process that occurs between W_t and W_{t+k}. We note that the simple Gaussian walk is not a good process for the losses, since its depth is too large for us to be able to control the size of the (unclipped) losses.

The feedback graph we work for the reset of this section is $G(\mathcal{A}, E)$, where $\mathcal{A} = \{1, 2, 3\}$ and $E = \{(1,3), (2,3), (1,1), (2,2), (3,3)\}$ (see Figure 2-2).

Constructing the losses. We consider the following adversarial sequence of losses. First sample an action uniformly at random from $\{1, 2\}$. WLOG we condition on the event that the sampled action is 1. Next set $\ell_t(3) = 1$, $\ell_t(2) = clip(W_t + \frac{1}{2})$, $\ell_t(1) = clip(W_t + \frac{1}{2} - \epsilon)$, where $clip(\alpha) = \min\{\max\{\alpha, 0\}, 1\}$. The intuition behind our lower bound is very simple and holds for a general feedback graph. It is as follows: if we do not have a complete feedback graph then there are at least two actions which do not tell us anything about each other. We leverage this by selecting one of the two actions uniformly at random to be the ***best*** action. If we play an action which is not 1 or 2 we incur constant regret in that turn but we can gain information about the losses of both 1 and 2. If we play 2, then we do not learn anything about 1 and if we play 1 we do not learn anything about 2. In these two cases the per round regret incurred is ϵ, however, because of the loss construction, we need to switch between these actions to be able to distinguish them and thus we will incur regret from switching. Overall the loss construction together with the result in Dekel et al. (2014) implies that to distinguish between 1 and 2 we need to observe the losses of both actions at the same time or switch between them at least $\tilde{\Omega}(T^{2/3})$ rounds. This is what we formally argue below.

Let Y_t be the observed loss vector associated with the action at time t, a_t, i.e. if $a_t = 2$ then $Y_t = W_t + \frac{1}{2}$, if $a_t = 1$ then $Y_t = W_t + \frac{1}{2} - \epsilon$ and if $a_t = 3$ then $Y_t = \begin{pmatrix} W_t + \frac{1}{2} \\ W_t + \frac{1}{2} - \epsilon \end{pmatrix}$. We let $Y_0 = 1/2$. We let $\mathcal{Q}_1$ be the probability measure on the σ-field $\mathcal{F}$ generated by $\{Y_t\}_{t=0}^T$. Let $\mathcal{Q}_0$ be the probability measure on the same σ-field if $\ell_t(1) = \ell_t(2) = clip(W_t + \frac{1}{2})$ i.e. there is no best action. In this case $Y_t = W_t + \frac{1}{2}$ for $a_t = 1$ or $a_t = 2$ and $Y_t = \begin{pmatrix} W_t + \frac{1}{2} \\ W_t + \frac{1}{2} \end{pmatrix}$ if $a_t = 2$. Denote by $\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1)$ the total variational distance between $\mathcal{Q}_0$ and $\mathcal{Q}_1$ on the σ-field $\mathcal{F}$. Let $\mathrm{D}_{\mathrm{KL}}(\mathcal{Q}_0 || \mathcal{Q}_1)$ be the KL-divergence between $\mathcal{Q}_0$ and $\mathcal{Q}_1$. We now show that a sufficiently large number of switches between actions 1 and 2 or choosing action 3 is required to distinguish between $\mathcal{Q}_0$ and $\mathcal{Q}_1$. As it was discussed above, the width of the process plays an important role, which is clarified by the lemma below. It essentially is an upper bound on the number of switches required to distinguish between $\mathcal{Q}_0$ and $\mathcal{Q}_1$.

Lemma 2.6.2. *Let M be the number of times the player's strategy switched between actions 1 and 2. Let N be the number of times the payer chose to play action 3. Then* $\mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) \leq \frac{\epsilon}{2\sigma}\sqrt{\omega(\rho)\mathbb{E}_{\mathcal{Q}_0}[M + N]}$.

Next we show that, because of the depth of the random walk, we are able to say that with high probability most of the non-clipped losses will be equal to the clipped losses. The implications of this result are two-fold. First the regret incurred on the non-clipped versions is close to the regret incurred on the clipped version. Secondly, we are able to say that loss of action 3 is worse by a constant from the losses of actions 1 and 2 often enough, so that we also incur constant regret when playing action 3 as compared to the other two actions. Let ℓ'_t denote the non-clipped version of ℓ_t and define

$$R' = \sum_{t=1}^{T} \ell'_t(a_t) + M - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell'_t(a)$$
$$R = \sum_{t=1}^{T} \ell_t(a_t) + M - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t(a)$$

Lemma 4 in Dekel et al. (2014) compares R' to R

Lemma 2.6.3. *For $T \geq 6$, $\mathbb{E}[R] \geq \mathbb{E}[R'] - \epsilon T/6$.*

The lower bound for $\mathbb{E}[R']$ is given by the following lemma.

Lemma 2.6.4. *Let $\mathcal{Q}_2$ be the conditional distribution induced by sampling the best action to be equal to 2. Then*

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{\epsilon T}{2}(\mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2)) + \mathbb{E}\left[M + \frac{N}{7}\right]$$

Putting the above two lemmas together, we are able to show Theorem 2.6.1.

Proof of Theorem 2.6.1. First assume that the event $M + N/7 > \epsilon T$ does not occur on losses generated from $\mathcal{Q}_0$ or $\mathcal{Q}_i$. This implies $\mathcal{Q}_0(M+N/7 > \epsilon T) = \mathcal{Q}_i(M+N/7 > \epsilon T) = 0$. Then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Q}_0}[M + N/7] - \mathbb{E}[M + N/7] \\
&= \frac{\mathbb{E}_{\mathcal{Q}_0}[M + N/7] - \mathbb{E}_{\mathcal{Q}_1}[M + N/7] + \mathbb{E}_{\mathcal{Q}_0}[M + N/7] - \mathbb{E}_{\mathcal{Q}_2}[M + N/7]}{2} \\
&\leq \frac{\epsilon T}{2}(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2)).
\end{aligned}$$

The above, together with Lemma 2.6.4 implies

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \epsilon T(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2)) + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{N}{7}\right].$$

Applying Lemma 2.6.3 now gives

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \epsilon T(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2)) + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{N}{7}\right].$$

On the other hand we can bound $(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2))/2$ by Lemma 2.6.2 as

$$(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_1) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_2))/2 \leq \frac{\epsilon}{\sigma\sqrt{2}}\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]\log(T)}.$$

This implies

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \frac{\sqrt{2}\epsilon^2 T}{\sigma}\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]\log(T)} + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{N}{7}\right].$$

Let $x = \sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]}$. Then we have

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \frac{\sqrt{2}\epsilon^2 T\sqrt{\log(T)}}{\sigma}x + \frac{x^2}{7}.$$

The quadratic $\frac{x^2}{7} - \frac{\sqrt{2}\epsilon^2 T\sqrt{\log(T)}}{\sigma}x$ has minimum $-\frac{7\log(T)\epsilon^4 T^2}{2\sigma^2}$. We set $\epsilon = c\frac{1}{T^{1/3}\log(T)}$ for a constant c to be determined later. We then have

$$\mathbb{E}[R] \geq \frac{cT^{2/3}}{3\log(T)} - \frac{7c^4}{2}\frac{T^{2/3}}{\log(T)^3\sigma^2}.$$

Set $\sigma = \frac{1}{\log(T)}$. The above implies

$$\mathbb{E}[R] \geq \frac{T^{2/3}}{\log(T)}\left(\frac{c}{3} - \frac{7c^4}{2}\right).$$

Choosing $c = \frac{1}{42^{1/3}}$ gives $\frac{c}{3} - \frac{7c^4}{2} \geq \frac{1}{16}$.

Suppose there is some strategy for which $M + N/7 \geq c\frac{T^{2/3}}{\log(T)}$ occurs. Let this strategy have regret R. We change the strategy in the following way. Keep track of $M + N/7$ and the moment it exceeds $c\frac{T^{2/3}}{\log(T)}$ pick an action which has had loss smaller than 5/6. If there is no such action, pick any action and play it until the end of the game. With probability at least $1/T$ we know that such an action exists and that it was set according to the stochastic process construction. Thus the regret of the new strategy R^* is bounded by $\mathbb{E}[R^*] \leq \mathbb{E}[R] + (1 - 1/T)\epsilon T + 1/T \times T \leq 2\mathbb{E}[R] + 1$. Since the lower bound holds for $\mathbb{E}[R^*]$ the proof is complete. □

2.6.2 Lower Bound for Disjoint Union of Star Graphs

Let G be the graph which is a union of star graphs. Let R be the set of revealing vertices for the star graphs. We denote by $\mathcal{A}_i$ the set of vertices associated with the star graph with revealing vertex v_i. First for each star graph we sample an ***active*** vertex uniformly at random from its leaves. Next we sample the best vertex uniformly at random from the set of active vertices. We set the loss of the best vertex to be $clip(W_t + 1/2 - \epsilon)$ and the loss of all other active vertices to $clip(W_t + 1/2)$. For any star graph consisting of a single vertex, we treat the vertex as a leaf. The following theorem follows as an easy reduction from the proof of Dekel et al. (2014).

Theorem 2.6.5. *The expected regret of any algorithm $\mathcal{A}$ on a disjoint union of star graphs is lower bounded as follows:*

$$R_T(\mathcal{A}) \geq \Omega\left(\frac{\gamma(G)^{1/3}T^{2/3}}{\log(T)}\right).$$

Proof of Theorem 2.6.5. Let $\mathcal{I}$ be the set of all possible ways to sample a set of active vertices. Let $\mathbb{E}_i$ be the expectation conditioned on the event that the set of active vertices indexed by $i \in \mathcal{I}$ is sampled in the beginning of the game. Consider the subgraph induced by the active vertices I and all of their neighbors R. Suppose that there exists a player's strategy such that $\mathbb{E}_i[R] \leq o\left(\frac{\gamma(G)^{1/3}T^{2/3}}{\log(T)}\right)$. We claim this strategy implies a regret upper bound for bandits with switching costs of the order $o\left(\frac{\gamma(G)^{1/3}T^{2/3}}{\log(T)}\right)$. We convert the player's strategy over $I \bigcup R$ to a strategy over I. For every time that $a_t \in R$ is played, we replace a_t by the ***unique*** neighbor of a_t in I. This updated strategy's regret is at most the regret of the original strategy and thus by our assumption it has regret at most $o\left(\frac{\gamma(G)^{1/3}T^{2/3}}{\log(T)}\right) = \left(\frac{|I|^{1/3}T^{2/3}}{\log(T)}\right)$. This is in contradiction with the result of Dekel et al. (2014) since the subgraph induced by I is precisely modeling bandit feedback and the losses of actions in I are exactly constructed as in Dekel et al. (2014). Thus we have $\mathbb{E}[R] \geq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathbb{E}_i[R] = \tilde{\Omega}\left(\frac{\gamma(G)^{1/3}T^{2/3}}{\log(T)}\right)$. □

Even though the above theorem is a trivial consequence of the result in Dekel et al. (2014) it can also be proved in another way. Let I denote the set of conditional distributions induced by the observed losses, where the conditioning is with respect to the random sampling of vertices as described in the beginning of the section. The general idea of the complicated proof is to count the number of distributions which each strategy of the player gains information about. For example a strategy which switches between two revealing vertices v_i and v_j will gain information about $deg(v_i)deg(v_j)$ distributions. Now the lower bound follows from a careful counting of the number of distributions for which we gain information by switching between revealing vertices. This counting argument can be generalized beyond union of star graphs, by considering an appropriate pair of minimal dominating/maximal independent sets. We leave a detailed argument for future work.

2.6.2.1 Counting Argument for Theorem 2.6.5

Let $\mathcal{I}$ denote the set of all possible ways to sample active vertices. The cardinality of this set is $|\mathcal{I}| = \prod_{v_i \in R} deg(v_i)$. Denote by $\mathcal{Q}_0^i$ the conditional distribution generated by the observed losses if all losses for active vertices indexed by $i \in \mathcal{I}$ were set to $clip(W_t + 1/2)$. Denote by $\mathcal{Q}_j^i$ the conditional distribution generated by the observed losses when active vertex j is chosen to be the best given the active vertices are indexed by $i \in \mathcal{I}$. Let M_j^i denote the random variable counting the number of times the player switched from and to an action adjacent to j. Let N_j^i denote the random variable counting the number of times the player played an action adjacent to j.

Lemma 2.6.6. *For all $i \in \mathcal{I}$ and $j \in [|R|]$ it holds that* $\mathrm{d}_{TV}^{\mathcal{F}}\left(\mathcal{Q}_0^i, \mathcal{Q}_j^i\right) \leq \frac{\epsilon}{2\sigma}\sqrt{\omega(\rho)\mathbb{E}_{\mathcal{Q}_0^i}[M_j^i + N_j^i]}$.

Let M_i denote the random variable measurable with respect to the draw of $i \in \mathcal{I}$ which counts the total number of switches. Similarly let N_i count the total number of times a revealing vertex of degree at least 2 was played.

Lemma 2.6.7. *The following holds*

$$\frac{1}{|R||\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{j\in[|R|]} \mathrm{d}_{TV}^{\mathcal{F}}\left(\mathcal{Q}_0^i, \mathcal{Q}_j^i\right) \leq \frac{\epsilon}{\sigma\sqrt{2|R|}}\sqrt{\frac{\omega(\rho)}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathbb{E}_{\mathcal{Q}_0^i}[M_i + N_i]}.$$

Proof. Notice that conditioned on the draw of $i \in \mathcal{I}$ we have $\sum_{j\in[|R|]} N_i^j \leq N_i$. This happens because there is only one revealing vertex adjacent to the best vertex for every $\mathcal{Q}_i^j$, i.e., the revealing vertex indexed by $j \in [|R|]$. Similarly we have $\sum_{j\in[|R|]} M_i^j \leq 2M_i$, where the constant two appears because we have counted each switch twice – once from action j and once to action j. Using Lemma 2.6.6 with concavity of the square root finishes the proof. □

The above lemma was easy to prove because we did not have two vertices which are dominated simultaneously by two different neighbors in R. This allowed us to count very easily the number of times we might have over-count N_i for two different choices of the best action. We were also lucky that it was impossible to gain information about the best action proportional to the degree of a revealing vertex. For a general graph both of these events can happen and the counting argument would have to be more careful.

Lemma 2.6.8. *The following holds*

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{\epsilon T}{|\mathcal{I}||R|} \sum_{i \in \mathcal{I}} \sum_{j \in [|R|]} \mathrm{d}_{TV}^{\mathcal{F}} \left(\mathcal{Q}_0^i, \mathcal{Q}_j^i \right) + \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_i \left[M_i + \frac{N_i}{7} \right]$$

Let M denote the random variable counting the total number of switches and N the random variable denoting the total number of times a revealing action with degree at least 2 was played. We can write $\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_i[M_i] \leq \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_i[M] = \mathbb{E}[M]$ and similarly $\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_i[N_i] \leq \mathbb{E}[N]$. The proof of Theorem 2.6.5 can now be completed by following the proof of Theorem 2.6.1. We note that bounding M_i by M is in general tight for disjoint union of star graphs and equality occurs for all strategies which switch only between revealing vertices. For general graphs this upper bound can become very loose and we should exercise caution when constructing an upper bound. In particular we should carefully count how many distributions are covered by a single switch.

2.6.3 Lower bound for a sequence of feedback graphs in the uninformed setting.

While going through the proof of Theorem 1 in Rangi and Franceschetti (2019), we came across an important technical mistake. In page 2 of the supplementary material, in the paragraph after Equation 8, the authors state that, at a single time instance, the loss of only one single action can be observed from the independent set in their construction. This is not correct since a player's strategy can play an action that is not in the independent set but is adjacent to two or more vertices in the independent set.

The problem with this statement becomes apparent when one considers a fixed feedback graph system, i.e., $G_t = G, \forall t \in [T]$, where G is a star graph. In that case, the construction of the losses by Rangi and Franceschetti (2019) amounts to sampling a best action from the leaves of G, setting its loss to be ϵ_1 smaller than the loss of all other actions in the leaves of G, and setting the revealing action to be ϵ_2

larger than the losses in the leaves of G. The losses of the remaining actions are set according to the stochastic process of Dekel et al. (2014). With these choice of losses and ϵ_1 and ϵ_2 set according to what the authors suggest, a very simple strategy is information-theoretically optimal: the player only needs to play the revealing action $T^{2/3}$ times to distinguish which of the leaves of G contains the best action. This strategy would actually incur expected regret of the order $\tilde{\Theta}(\sqrt{T})$.

Let $\alpha(G_{1:T})$ denote the largest cardinality among all intersections of independent sets of the sequence $(G_t)_{t=1}^T$. A lower bound of $\tilde{\Omega}(\alpha(G_{1:T})^{1/3}T^{2/3})$ is still possible under additional assumptions about how the feedback graph system is generated in the ***uninformed*** setting. In particular, we show that if we allow the feedback graphs to be chosen by the adversary, there still exists a sequence of feedback graphs for which the lower bound is $\tilde{\Omega}(\alpha(G_{1:T})^{1/3}T^{2/3})$, while for each G_t, we have $\gamma(G_t) = 1$.

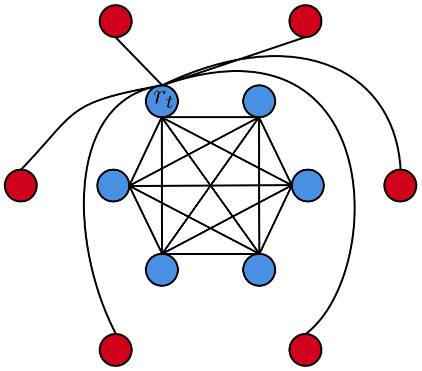


Figure 2-3: G_t

Formally, we show that in the uninformed setting, when we allow the graphs to be chosen by the adversary, there exists a sequence $(G_t)_{t=1}^T$ such that for all $t \in [T]$, $\gamma(G_t) = 1$, $\alpha(G_t) \gg 1$ and $\alpha(G_{1:t}) = \Theta(\alpha(G_t))$, for which any player's strategy will incur regret of the order $\tilde{\Omega}(\alpha(G_{1:t})^{1/3}T^{2/3})$. In particular, there is a non-trivial example of a sequence of graphs for which the independence number is arbitrarily larger than the domination number and every strategy has to incur regret depending on the independence number.

We now present our construction. Fix $\alpha \gg 1$ and let $|\mathcal{A}| = 2\alpha$. Let I be a subset of $\mathcal{A}$ of size α and let $R = \mathcal{A} \setminus I$. Set the losses of actions in I according to the construction of Dekel et al. (2014), as described in Section 2.6.1. Set the losses of actions in R equal to one. The edges of the graph $G_t = (\mathcal{A}, E_t)$ at round t are defined as follows. The vertices in R form a clique. A vertex r is sampled uniformly at random from R to be the revealing action and all edges $(r, v_i), v_i \in I$ are also added to E_t. We note that $\alpha(G_t) = \alpha + 1$, $\gamma(G_t) = 1$ for all $t \in [T]$ and $\alpha(G_{1:T}) = \alpha$. We present an illustration for our construction in Figure 2-3. Here $\alpha = 6$, the set I are the vertices in red, the set R are the vertices in blue.

The intuition behind our construction is that the player needs on average α rounds to observe the losses of all actions, due to the randomization over the revealing vertex r. The switching cost again contributes to the $T^{2/3}$ time-horizon regret.

Again assume that the strategy of the player is deterministic. We let $\mathcal{Q}_i$ denote the conditional distribution generated by the observed losses, when the best action was sampled to be $v_i \in I$ and $\mathcal{Q}_0$ denotes the distribution over observed losses when there is no best action in I. Let M_i be the number of times the player's strategy switched between an action in $I \setminus \{i\}$ and i. Let M_i' be the number of times that the player switched between i and the revealing action. Let N be the total number of times a vertex in R was played and let N' be the total number of times a revealing vertex was played. We have the following.

Lemma 2.6.9. *For all* $i \in [|I|] \bigcup \{0\}$

$$\frac{1}{\alpha}\mathbb{E}_{\mathcal{Q}_i}[N] = \mathbb{E}_{\mathcal{Q}_i}[N'].$$

Let M denote the random variable counting the total number of switches.

Lemma 2.6.10. *The following inequality holds:* $\frac{1}{\alpha}\sum_{v_i \in I} \mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) \leq \frac{\epsilon}{\sigma}\sqrt{\frac{\omega(\rho)}{2\alpha}}\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]}$.

Repeating the rest of the arguments in Section 2.6.4 with $\phi(G)$ replaced by $\frac{1}{\alpha}$ shows the following theorem.

Theorem 2.6.11. *For any* $\alpha > 1, \alpha \in \mathbb{N}$, *there exists an adversarially generated sequence of feedback graphs* $(G_t)_{t=1}^T$, *with* $\alpha(G_t) = \alpha + 1, \gamma(G_t) = 1, \forall t \in [T]$ *and* $\alpha(G_{1:T}) = \alpha$, *such that the expected regret of any strategy in the uninformed setting is* *at least*

$$\mathbb{E}[R] \geq \frac{\alpha^{1/3}T^{2/3}}{16\log{(T)}}.$$

2.6.4 Lower Bound for Arbitrary Graphs

In this section we propose a construction leading to a non-tight lower bound for general graphs. We choose to present this construction due to it developing tools which can be useful for a tight generic bound. In particular the way we use Lemma 2.6.12 in the proof of Lemma 2.6.6 can be mimicked for general graphs when coupled with a careful counting argument.

Let $G = (\mathcal{A}, E)$ be a feedback graph with vertex set $\mathcal{A}$ and edge set E. Let $\mathcal{I}$ denote the set of all maximal independent sets I of G. For any I we say that I is dominated by $S \subseteq \mathcal{A}$ if for every $v \in I$, there exists a neighbor of v in S. For any I let S_I be a minimal set of vertices which dominates I and let $\mathcal{S}_I$ be the set of all such S_I. Let $\delta(S_I)$ equal the maximum number of neighbors in I, which a vertex in S_I can have. Let $\delta(\mathcal{S}_I)$ be the maximum over all $\delta(S_I)$ and let $\phi(G) = \min_{I\in\mathcal{I}} \frac{\delta(\mathcal{S}_I)}{|I|}$. Let I^* be a maximal independent set for which $|S_{I^*}| = \phi(G)$. To construct our adversarial loss sequence we begin by uniformly sampling an action i from I^* and setting it to be the action with smallest loss. Let $\mathcal{Q}_i$ denote the conditional probability measure given the sampled best action was i and let $\mathcal{Q}_0$ be the probability distribution when all of the actions in I^* are equal i.e. there is no best action. Let W_t be the stochastic process as defined in Section 2.6.1. We set the losses for actions in I^* to be $clip(W_t + 1/2)$ for $v \in I^* \setminus \{i\}$ and the loss of i to be $clip(W_t + 1/2 - \epsilon)$. The loss of all other actions is set to be 1. We let Y_t denote the loss vector of observed losses only on I^*. WLOG we can disregard other losses, since they will not let us distinguish between $\mathcal{Q}_i$ and $\mathcal{Q}_0$. We denote by $Y_t(j)$ the loss of action $j \in I^*$ if that loss was observed at time t. Let $\mathcal{F}$ be the σ-field generated by $(Y_t)_{t=1}^T$.

Our intuition behind the definition of $\phi(G)$ and the above construction is the following. First we require that the losses based on the stochastic process $(W_t)_{t=1}^T$ be assigned to vertices in an independent set. Otherwise, there would exist a setting in which the best action would be adjacent to another action with losses generated from $(W_t)_{t=1}^T$ and in this case it is information theoretically possible to obtain $O(\sqrt{T})$ regret by playing the best action or its adjacent action enough times, without switching. For every independent set, once a best action is fixed, from the lower bound in Section 2.6.1 we know two ways to distinguish it. First we switch between the best action and some other action in the independent set (or more generally switch between actions giving information about the best action and another action in the independent set), or play an action which is adjacent to the best action and another action in the independent set. In the general setting there might be an action which is adjacent to multiple actions in the independent set and not adjacent to the best action. In such cases switching between the best action and said action, reveals information proportional to the degree of said action. Similarly if there is an action adjacent to the best action and multiple other actions, selecting it again reveals information proportional to its degree. Since we do not want to assume anything about the strategy of the player, it is natural to select an independent set, such that minimum amount of vertices have a common neighbor. Because the size of the independent set also gives freedom to hide information from the player, we would simultaneously like to maximize its size. This suggests that we search for and independent set which minimizes the ratio in the definition of $\phi(G)$. In Figure 2-4 we give three examples of graphs with different $\phi(G)$. For the first example the independent set $|I^*|$ is the set of all vertices. The set S_{I^*} is also the set of all vertices and $\delta(S_{I^*}) = 1$ thus $\phi(G) = 1/|\mathcal{A}|$ and this is exactly equal to $\gamma(G)^{-1}$. For the second example I^* is the set of leafs of the star graph and S_{I^*} is the vertex adjacent to all other vertices. In this case $\delta(S_{I^*}) = |I^*|$ and $\phi(G) = 1$ which again equals the inverse of the dominating number of G. Our final example

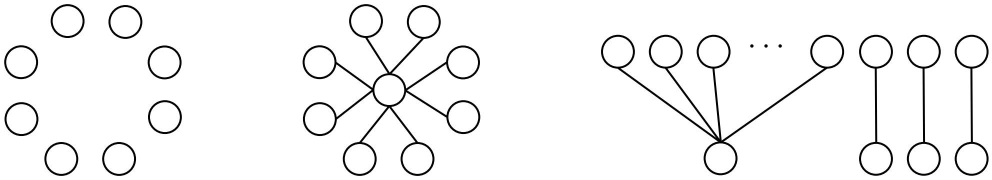


Figure 2-4: Example of feedback graphs with different $\phi(G)$.

shows that $\phi(G)$ can be arbitrary close to 1 even though $\gamma(G)^{-1} < 1$. In particular S_{I^*} consists of the bottom 4 vertices and this is also the minimum dominating set of G. However, there exists a vertex (the first vertex of the bottom four) of arbitrary large degree so that $\frac{\delta(S_{I^*})}{|I^*|}$ can be arbitrary close to 1. The problem with our lower bound construction becomes clear from this example. The player has a strategy in which too much information is revealed by playing the action of arbitrary large degree. To try and fix this problem we could set only one of the vertices adjacent to the action of large degree according to $(W_t)_{t=1}^T$ and the rest of the adjacent actions are set to have loss equal to 1. This construction can fail for general graphs, as it might happen that there exists another action which is adjacent to exactly the four actions whose losses were chosen according to $(W_t)_{t=1}^T$ in the right most graph of Figure 2-4.

Lemma 2.6.12. *Let M_i be the number of times the player's strategy switched between action adjacent only to i and another action not adjacent to i but adjacent to at least one other action in I^*. Let N_i be the number of times the player chose to play an action adjacent to i and another action in I^*. Then* $\mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) \leq \frac{\epsilon}{2\sigma}\sqrt{\omega(\rho)\mathbb{E}_{\mathcal{Q}_0}[|I^*|\phi(G)M_i + N_i]}$.

Let M denote the total number of switches and N the total number of times an action revealing adjacent to at least two vertices in I^* is played.

Lemma 2.6.13. *It holds that* $\frac{1}{|I^*|}\sum_{i\in I^*} \mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) \leq \frac{\epsilon}{\sigma}\sqrt{\frac{\omega(\rho)\phi(G)}{2}}\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M+N]}$.

Lemma 2.6.14. *It holds that*

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \epsilon T \frac{1}{|I^*|}\sum_{i\in I^*} \mathrm{d}_{TV}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) + \mathbb{E}\left[M + \frac{N}{7}\right].$$

Theorem 2.6.15. *The expected regret of a deterministic player is at least*

$$\mathbb{E}[R] \geq 4\frac{T^{2/3}}{\log(T)\,\phi(G)^{1/3}}$$

Proof. First assume that the event $M + N/7 > \epsilon T$ does not occur on losses generated from $\mathcal{Q}_0$ or $\mathcal{Q}_i$ for a deterministic player strategy. This implies $\mathcal{Q}_0(M + N/7 > \epsilon T) = \mathcal{Q}_i(M + N/7 > \epsilon T) = 0$. Then

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}_0}[M + N/7] - \mathbb{E}[M + N/7] &= \frac{1}{|I^*|}\sum_{i \in I^*}\left(\mathbb{E}_{\mathcal{Q}_0}[M + N/7] - \mathbb{E}_{\mathcal{Q}_i}[M + N/7]\right)\\ &\leq \frac{\epsilon T}{|I^*|}\sum_{i \in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right).\end{aligned}$$

The above, together with Lemma 2.6.14 implies

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{2\epsilon T}{|I^*|}\sum_{i \in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{1}{7}N\right].$$

Applying Lemma 2.6.3 now gives

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \frac{2\epsilon T}{|I^*|}\sum_{i \in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{1}{7}N\right].$$

On the other hand we can bound $\frac{1}{|I^*|}\sum_{i \in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right)$ by Lemma 2.6.13 as

$$\frac{1}{|I^*|}\sum_{i \in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) \leq \frac{\epsilon}{\sigma}\sqrt{\frac{\log(T)\,\phi(G)}{2}}\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]}.$$

This implies

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \frac{\sqrt{2}\epsilon^2 T}{\sigma}\sqrt{\phi(G)\log(T)\,\mathbb{E}_{\mathcal{Q}_0}[M + N]} + \mathbb{E}_{\mathcal{Q}_0}\left[M + \frac{1}{7}N\right].$$

Let $x = \sqrt{\mathbb{E}_{\mathcal{Q}_0}\left[M + N\right]}$. Then we have

$$\mathbb{E}[R] \geq \frac{\epsilon T}{3} - \frac{\sqrt{2}\epsilon^2 T\sqrt{\log(T)\,\phi(G)}}{\sigma}x + \frac{x^2}{7}.$$

The quadratic $\frac{x^2}{7} - \frac{\epsilon^2 T\sqrt{2\log(T)\phi(G)}}{\sigma}x$ has global minimum $-\frac{\epsilon^4 T^2 \log(T)\phi(G)}{14}$ We set $\epsilon = c\frac{1}{T^{1/3}\log(T)}$ for a constant c to be determined later. We then have

$$\mathbb{E}[R] \geq \frac{cT^{2/3}}{3\log(T)} - \frac{c^4}{14}\frac{T^{2/3}\phi(G)}{\log(T)^3\,\sigma^2}.$$

Set $\sigma = \frac{1}{\log(T)}$. The above implies

$$\mathbb{E}[R] \geq \frac{T^{2/3}}{\log(T)} \left(\frac{c}{3} - \frac{c^4 \phi(G)}{14} \right).$$

Choosing $c = \left(\frac{7}{6\phi(G)} \right)^{1/3}$ guarantees $\mathbb{E}[R] \geq \frac{T^{2/3}}{16 \log(T) \phi(G)^{1/3}}$.

The case when $M + N/7 > \epsilon T$ is treated in the same way as in the proof of Theorem 2.6.1 □

2.7 Detailed proofs for Section 2.6

2.7.1 Detailed proofs for Section 2.6.1

Proof of Lemma 2.6.2. Let $Y_{0:t}$ denote $(Y_0, Y_1, \ldots, Y_t)$ and whenever Y_t is a vector, let $Y_t(i)$ be its i-th coordinate. We assume that the player is deterministic. By Yao's minimax principle this is without loss of generality. Thus we have that a_t is a deterministic function of $Y_{0:t-1}$. Using the chain rule for relative entropy and by the construction of W_t, we have:

$$\mathrm{D}_{\mathrm{KL}} \left(\mathcal{Q}_0(Y_{0:T}) || \mathcal{Q}_1(Y_{0:T}) \right) = \mathrm{D}_{\mathrm{KL}} \left(\mathcal{Q}_0(Y_0) || \mathcal{Q}_1(Y_1) \right) + \sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}} \left(\mathcal{Q}_0(Y_t | Y_{\rho*(t)}) || \mathcal{Q}_1(Y_t | Y_{\rho*(t)}) \right).$$

Let us consider the term $\mathrm{D}_{\mathrm{KL}} \left(\mathcal{Q}_0(Y_t | Y_{\rho*(t)}) || \mathcal{Q}_1(Y_t | Y_{\rho*(t)}) \right)$. First assume that $a_t = a_{\rho(t)} \neq 3$. Then $Y_t = \mathcal{N}(Y_{\rho(t)}, \sigma^2)$ under both $\mathcal{Q}_0$ and $\mathcal{Q}_1$. Next consider the case when $a_t = a_{\rho(t)} = 3$. In this case $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)}(2) \\ Y_{\rho(t)}(2) \end{pmatrix}, \sigma^2 \mathrm{I}_2 \right)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)}(2) - \epsilon \\ Y_{\rho(t)}(2) \end{pmatrix}, \sigma^2 \mathrm{I}_2 \right)$ under $\mathcal{Q}_1$. If $a_t \neq a_{\rho(t)}$ we have 6 options:

1. $a_{\rho(t)} = 3$
   (a) $a_t = 1$, in this case $Y_t = \mathcal{N}(Y_{\rho(t)}(2), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}(Y_{\rho(t)}(2) - \epsilon, \sigma^2)$ under $\mathcal{Q}_1$;

   (b) $a_t = 2$ in this case $Y_t = \mathcal{N}(Y_{\rho(t)}(2), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}(Y_{\rho(t)}(2), \sigma^2)$ under $\mathcal{Q}_1$;

2. $a_{\rho(t)} = 1$

   (a) $a_t = 3$, in this case $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)} \\ Y_{\rho(t)} \end{pmatrix}, \sigma^2 \mathrm{I}_2\right)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)} \\ Y_{\rho(t)} + \epsilon \end{pmatrix}, \sigma^2 \mathrm{I}_2\right)$ under $\mathcal{Q}_1$;

   (b) $a_t = 2$ in this case $Y_t = \mathcal{N}(Y_{\rho(t)}, \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}(Y_{\rho(t)} + \epsilon, \sigma^2)$ under $\mathcal{Q}_1$;

3. $a_{\rho(t)} = 2$

   (a) $a_t = 3$, in this case $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)} \\ Y_{\rho(t)} \end{pmatrix}, \sigma^2 \mathrm{I}_2\right)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}\left(\begin{pmatrix} Y_{\rho(t)} - \epsilon \\ Y_{\rho(t)} \end{pmatrix}, \sigma^2 \mathrm{I}_2\right)$ under $\mathcal{Q}_1$;

   (b) $a_t = 1$ in this case $Y_t = \mathcal{N}(Y_{\rho(t)}, \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t = \mathcal{N}(Y_{\rho(t)} - \epsilon, \sigma^2)$ under $\mathcal{Q}_1$.

Thus we have

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_1(Y_t|Y_{\rho*(t)})\right) &= \mathcal{Q}_0(a_t = a_{\rho(t)} = 3)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(-\epsilon,\sigma^2)\right) \\
&+ \mathcal{Q}_0(a_{\rho(t)=3}, a_t = 1)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(-\epsilon,\sigma^2)\right) \\
&+ \mathcal{Q}_0(a_{\rho(t)=1}, a_t = 3)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(\epsilon,\sigma^2)\right) \\
&+ \mathcal{Q}_0(a_{\rho(t)=1}, a_t = 2)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(\epsilon,\sigma^2)\right) \\
&+ \mathcal{Q}_0(a_{\rho(t)=2}, a_t = 3)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(-\epsilon,\sigma^2)\right) \\
&+ \mathcal{Q}_0(a_{\rho(t)=2}, a_t = 1)\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(0,\sigma^2)||\mathcal{N}(-\epsilon,\sigma^2)\right) \\
&= \frac{\epsilon^2}{2\sigma^2}\mathcal{Q}_0(A_t),
\end{aligned}$$

where A_t is the event that either action 3 was played at round t or there were odd number of switches between actions 1 and 2. Let N denote the random number of times action 3 was played and let M denote the random number of switches between action 1 and action 2. Let $S_{1:M}$ denote the random sequence of times during which there was a switch. Then we have

$$\sum_{t=1}^{T} \chi_{A_t} \le \sum_{r=1}^{M} \sum_{t \in \mathrm{cut}(S_r)} \chi_{A_t} + N \le \omega(\rho)(M + N),$$

where $\mathrm{cut}(t)$ and $\omega(\rho)$ are defined in Dekel et al. (2014). Thus

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_1(Y_t|Y_{\rho*(t)})\right) \leq \frac{\epsilon^2\omega(\rho)}{2\sigma^2}\mathbb{E}_{\mathcal{Q}_0}[M+N].$$

Pinsker's inequality that $\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_1\right) \leq \frac{\epsilon}{2\sigma}\sqrt{\omega(\rho)\mathbb{E}_{\mathcal{Q}_0}[M+N]}$ □

Proof of Lemma 2.6.4. First let us consider the amount of regret the player incurs for picking action 3 N times. To do this we consider the number of times $1/2 + W_t > 5/6$. The expected number of times this occurs is

$$\mathbb{E}\sum_{t=1}^{T}\chi_{1/2+W_t>5/6} \leq \sum_{t=1}^{T}\mathbb{P}\left(|W_t| + \frac{1}{2} \geq \frac{5}{6}\right) \leq \sum_{t=1}^{T} e^{-\frac{1}{d(\rho)\sigma^2}} \leq \sum_{t=1}^{T} e^{-\frac{9\log(T)}{2}} \leq 1.$$

Thus in expectation the regret for picking action 2 N times is at least $(1/6+\epsilon)(N-1)$. Since we choose $\epsilon = \tilde{\Theta}(T^{-1/3})$, for sufficiently large T we have that in expectation the regret for picking action 3 N times is at least $(N-1)/6$. Let χ denote the uniform random variable over actions $\{1, 2\}$, which picks the best action in the beginning of the game. Denote by B_i the number of times action i was played. Then $\mathbb{E}[R'] \geq \mathbb{E}[\epsilon(T-N-B_\chi) + M + (N-1)/6]$ (this is a lower bound since M only tracks the switches between actions 1 and 2, so the switches to and from action 2 are left out). Thus we have

$$\begin{aligned}\mathbb{E}[R'] &= \frac{\mathbb{E}[\epsilon(T-N-B_1) + M + (N-1)/6|\chi = 1] + \mathbb{E}[\epsilon(T-N-B_2) + M + (N-1)/6|\chi = 2]}{2} \\ &= \epsilon T - \frac{\epsilon}{2}\left(\mathbb{E}_{\mathcal{Q}_1}[B_1] + \mathbb{E}_{\mathcal{Q}_2}[B_0]\right) + \mathbb{E}\left[M + \frac{N-1}{6} - \epsilon N\right].\end{aligned}$$

Since $\epsilon = \tilde{\Theta}(T^{-1/3})$ we have $\frac{N-1}{6} - \epsilon N \leq \frac{N}{7}$. Consider $\mathbb{E}_{\mathcal{Q}_1}[B_1]$, we have

$$\mathbb{E}_{\mathcal{Q}_1}[B_1] - \mathbb{E}_{\mathcal{Q}_0}[B_1] = \sum_{t=1}^{T}(\mathcal{Q}_1(a_t = 1) - \mathcal{Q}_0(a_t = 1)) \leq T\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_1\right).$$

A similar inequality holds for $\mathbb{E}_{\mathcal{Q}_2}[N_0]$ and thus we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}_1}[B_1] + \mathbb{E}_{\mathcal{Q}_2}[B_0] &\leq T(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_1\right) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_2\right)) + \mathbb{E}_{\mathcal{Q}_0}[B_0 + B_1] \\ &\leq T(\mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_1\right) + \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_2\right)) + T - \mathbb{E}_{\mathcal{Q}_0}[N].\end{aligned}$$

The above implies

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{\epsilon T}{2}(\mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}\left(\mathcal{Q}_0, \mathcal{Q}_1\right) + \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}\left(\mathcal{Q}_0, \mathcal{Q}_2\right)) + \mathbb{E}\left[M + \frac{N}{7}\right] + \frac{\epsilon}{2}\mathbb{E}_{\mathcal{Q}_0}[N].$$

□

2.7.2 Detailed proofs from Section 2.6.2

Proof of Lemma 2.6.8. Let $\mathbb{E}_i$ denote the conditional distribution for sampling the active vertex set indexed by $i \in \mathcal{I}$. We have $\mathbb{E}[R'] = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathbb{E}_i[R']$. First let us consider the amount of regret the player incurs for picking a revealing action N_i times. To do this we consider the number of times $1/2 + W_t > 5/6$. The expected number of times this occurs is

$$\mathbb{E}\sum_{t=1}^{T} \chi_{1/2+W_t>5/6} \leq \sum_{t=1}^{T} \mathbb{P}\left(|W_t| + \frac{1}{2} \geq \frac{5}{6}\right) \leq \sum_{t=1}^{T} e^{-\frac{1}{d(\rho)\sigma^2}} \leq \sum_{t=1}^{T} e^{-\frac{9\log(T)}{2}} \leq 1.$$

Thus in expectation the regret for picking a revealing action N_i times is at least $(1/6 + \epsilon)(N_i - 1)$. Let χ_i denote the uniform random variable over R which picks the best action. Denote by B^i_j the number of times action j was played from the active vertices. Then $\mathbb{E}_i[R'] \geq \mathbb{E}_i[\epsilon(T - N_i - B^i_{\chi_i}) + M_i + N_i/6 - 1/6]$. Thus we have

$$\begin{aligned}\mathbb{E}[R'] &= \frac{\sum_{i\in|\mathcal{I}|} \mathbb{E}_i[\epsilon(T - N_i - B^i_{\chi_i}) + M_i + N_i/6 - 1/6]}{|\mathcal{I}|}\\ &= \epsilon T - \frac{\epsilon}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathbb{E}_i[B^i_{\chi_i}] + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathbb{E}_i\left[M_i + \frac{N_i}{6} - 1/6 - \epsilon N_i\right].\end{aligned}$$

Consider $\mathbb{E}_i[B^i_{\chi_i}] = \frac{1}{|R|}\sum_{j\in[|R|]} \mathbb{E}_{\mathcal{Q}^i_j}[B^i_j]$. For each term of the sum we have

$$\mathbb{E}_{\mathcal{Q}^i_j}[B^i_j] - \mathbb{E}_{\mathcal{Q}^i_0}[B^i_j] = \sum_{t=1}^{T}(\mathcal{Q}^i_j(a_t = j) - \mathcal{Q}_0(a_t = j)) \leq T\mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}\left(\mathcal{Q}^i_0, \mathcal{Q}^i_j\right).$$

Thus we get

$$\begin{aligned}\sum_{i\in\mathcal{I}} \mathbb{E}_i[B^i_{\chi_i}] &\leq T\frac{1}{|R|}\sum_{i\in\mathcal{I}}\sum_{j\in[|R|]} \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}\left(\mathcal{Q}^i_0, \mathcal{Q}^i_j\right) + \frac{1}{|R|}\sum_{i\in\mathcal{I}}\sum_{j\in[|R|]} \mathbb{E}_{\mathcal{Q}^i_0}[B^i_j]\\ &\leq \frac{T}{|R|}\sum_{i\in\mathcal{I}}\sum_{j\in[|R|]} \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}\left(\mathcal{Q}^i_0, \mathcal{Q}^i_j\right) + T - \frac{1}{|R|}\sum_{i\in\mathcal{I}} \mathbb{E}_{\mathcal{Q}^i_0}[N_i].\end{aligned}$$

Using the assumption that $|\mathcal{I}| \geq 2$, the above implies

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{\epsilon T}{|\mathcal{I}||R|} \sum_{i \in \mathcal{I}} \sum_{j \in [|R|]} \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}} \left(\mathcal{Q}^i_0, \mathcal{Q}^i_j \right) + \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_i \left[M_i + \frac{N_i}{6} - 1/6 - \epsilon N_i \right]$$

Since $\epsilon = \tilde{\Theta}(T^{-1/3})$ we have $\mathbb{E}_i \left[M_i + \frac{N_i - 1}{6} - \epsilon N_i \right] \geq \mathbb{E}_i \left[M_i + \frac{N_i}{7} \right]$. □

2.7.3 Detailed proofs from Section 2.6.3

Proof of Lemma 2.6.9. Let r_t denote the revealing action at time t.

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}_i}[N'] = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{Q}_i}[\mathbb{I}(a_t = r_t)] &= \sum_{t=1}^{T} \mathcal{Q}_i(a_t \in R) \mathbb{E}_{\mathcal{Q}_i}[\mathbb{I}(a_t = r_t) | a_t \in R] \\
&+ \sum_{t=1}^{T} \mathcal{Q}_i(a_t \notin R) \mathbb{E}_{\mathcal{Q}_i}[\mathbb{I}(a_t = r_t) | a_t \notin R] \\
&= \sum_{t=1}^{T} \mathcal{Q}_i(a_t \in R) \mathbb{E}_{\mathcal{Q}_i}[\mathbb{I}(a_t = r_t) | a_t \in R] \\
&= \sum_{t=1}^{T} \mathcal{Q}_i(a_t \in R) \frac{1}{\alpha} = \frac{1}{\alpha} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{Q}_i}[\mathbb{I}(a_t \in R)] = \frac{1}{\alpha} \mathbb{E}_{\mathcal{Q}_i}[N].
\end{aligned}$$

This completes the proof. □

Proof of Lemma 2.6.10. The proof of Lemma 2.6.12 implies that for any $\mathcal{Q}_i$ we have

$$\mathrm{d}^{\mathcal{F}}_{\mathrm{TV}} \left(\mathcal{Q}_0, \mathcal{Q}_i \right) \leq \frac{\epsilon}{2\sigma} \sqrt{\omega(\rho) \mathbb{E}_{\mathcal{Q}_0}[\alpha M'_i + M_i + N']},$$

since the amount of information that can be revealed by a switch is at most α and this precisely happens when the player switches from i to the revealing action. Notice that $\sum_{v_i \in I} M'_i \leq N'$, because the number of switches between any i and a revealing action is bounded by the number of times a revealing action is played. Lemma 2.6.9 implies that $\mathbb{E}_{\mathcal{Q}_0}[\alpha M'_i + M_i + N'] \leq \mathbb{E}_{\mathcal{Q}_0}[N/\alpha + M_i + \alpha M'_i]$. Next, we note that $\sum_{i \in [|I|]} M_i \leq 2M$ as each switch is counted at most twice by M_i. Thus we have

$$\begin{aligned}
\frac{1}{\alpha} \sum_{v_i \in I} \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}} \left(\mathcal{Q}_0, \mathcal{Q}_i \right) &\leq \frac{1}{\alpha} \frac{\epsilon}{2\sigma} \sum_{v_i \in I} \sqrt{\omega(\rho) \mathbb{E}_{\mathcal{Q}_0}[N/\alpha + M_i + \alpha M'_i]} \\
&\leq \frac{\epsilon}{2\sigma} \sqrt{\frac{\omega(\rho)}{\alpha} \mathbb{E}_{\mathcal{Q}_0} \left[\sum_{v_i \in I} N/\alpha + M_i + \alpha M'_i \right]} \\
&\leq \frac{\epsilon}{\sigma} \sqrt{\frac{\omega(\rho)}{2\alpha}} \sqrt{\mathbb{E}_{\mathcal{Q}_0}[M + N]},
\end{aligned}$$

where the second to last inequality follows again from Lemma 2.6.9. □

2.7.4 Detailed proofs from Section 2.6.4

Proof of Lemma 2.6.12. Using Yao's minimax principle we can assume the player is deterministic and thus their t-th action a_t is a deterministic function of $Y_{0:t-1}$.Using the chain rule for relative entropy and by the construction of W_t, we have:

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_{0:T})||\mathcal{Q}_i(Y_{0:T})\right) = \mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_0)||\mathcal{Q}_i(Y_1)\right) + \sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_i(Y_t|Y_{\rho*(t)})\right).$$

Let us consider the term $\mathrm{D}_{\mathrm{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_i(Y_t|Y_{\rho*(t)})\right)$. First assume that $a_t = a_{\rho(t)}$ is not an action adjacent to i or $a_t = a_{\rho(t)} = i$. Then for any observed $j \in I^*$ we have $Y_t(j) = \mathcal{N}(Y_{\rho(t)}, \sigma^2)$ under both $\mathcal{Q}_0$ and $\mathcal{Q}_i$. Next consider the case when $a_t = a_{\rho(t)}$ is an action adjacent to i and some other $j \in I^*$. In this case $Y_t(j) = Y_t(i) = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(i) = \mathcal{N}(Y_{\rho(t)}(j) - \epsilon, \sigma^2)$, $Y_t(j) = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_i$ for all observed $j \in I^* \setminus \{i\}$. If $a_t \neq a_{\rho(t)}$ we have 6 options:

1. $a_{\rho(t)}$ is an action adjacent to i and another action $j \in I^* \setminus \{i\}$

   (a) a_t is an action adjacent to i, in this case $Y_t(j) = Y_t(i) = \mathcal{N}(Y_{\rho(t)}(j'), \sigma^2)$ under $\mathcal{Q}_0$ for all observed $j' \in I^*$ and $Y_t(i) = \mathcal{N}(Y_{\rho(t)}(j) - \epsilon, \sigma^2)$, $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_i$ for all observed $j' \in I^*$;

   (b) a_t is an action not adjacent to i in this case $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_i$ for all observed j' in I^*;

2. $a_{\rho(t)}$ is an action not adjacent to i but adjacent to j

   (a) a_t is an action adjacent to i, in this case $Y_t(j') = Y_t(i) = \mathcal{N}(Y_t(j), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(i) = \mathcal{N}(Y_{\rho(t)}(j) - \epsilon, \sigma^2)$, $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_i$ for all observed j';

   (b) a_t is an action not adjacent to i, in this case $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j), \sigma^2)$ under $\mathcal{Q}_i$ for all observed j';

3. $a_{\rho(t)}$ is an action only adjacent to i and no other $j \in I^*$

   (a) a_t is an action adjacent to i, in this case $Y_t(j') = Y_t(i) = \mathcal{N}(Y_{\rho(t)}(i), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(i) = \mathcal{N}(Y_{\rho(t)}(i), \sigma^2)$, $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(j') + \epsilon, \sigma^2)$ under $\mathcal{Q}_i$ for all observed j';

   (b) a_t is an action not adjacent to i, in this case $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(i), \sigma^2)$ under $\mathcal{Q}_0$ and $Y_t(j') = \mathcal{N}(Y_{\rho(t)}(i) + \epsilon, \sigma^2)$ under $\mathcal{Q}_i$ for all observed j'.

Thus we have

$$\mathrm{D_{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_i(Y_t|Y_{\rho*(t)})\right) \leq \frac{\epsilon^2}{2\sigma^2}\mathcal{Q}_0(A_t) + |I^*|\phi(G)\frac{\epsilon^2}{2\sigma^2}\mathcal{Q}_i(B_t)$$

where A_t is the event that $a_{\rho(t)}$ was adjacent to at least one action in $I^* \setminus \{i\}$ and at time t action i was observed and B_t is the event that $a_{\rho(t)}$ was adjacent only to i and the player switched at time t to an action which is adjacent to an action in $I^* \setminus \{i\}$. Let N_i denote the random number of times an action adjacent to i was played and let M_i denote the random number of switches between an action adjacent to i and an action not adjacent to i. Let $S_{1:M}$ denote the random sequence of times during which there was a switch. Then we have

$$\sum_{t=1}^{T} \chi_{A_t} + \chi_{B_t} \leq \sum_{r=1}^{M} \sum_{t \in \mathrm{cut}(S_r)} \chi_{A_t} + N_i \leq \omega(\rho)(M_i + N_i),$$

where $\mathrm{cut}(t)$ and $\omega(\rho)$ are defined in Dekel et al. (2014). Thus

$$\mathrm{D_{KL}}\left(\mathcal{Q}_0(Y_t|Y_{\rho*(t)})||\mathcal{Q}_i(Y_t|Y_{\rho*(t)})\right) \leq \frac{\epsilon^2\omega(\rho)}{2\sigma^2}\mathbb{E}_{\mathcal{Q}_0}[|I^*|\phi(G)M_i + N_i].$$

Pinsker's inequality that $\mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) \leq \frac{\epsilon}{2\sigma}\sqrt{\omega(\rho)\mathbb{E}_{\mathcal{Q}_0}[|I^*|\phi(G)M_i + N_i]}$. □

Proof of Lemma 2.6.13. From concavity of square root and Lemma 2.6.12 we have

$$\frac{1}{|I^*|}\sum_{i \in I^*} \mathrm{d}^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) \leq \frac{\epsilon\sqrt{\omega(\rho)}}{2\sigma}\sqrt{\frac{1}{|I^*|}\mathbb{E}_{\mathcal{Q}_0}\left[\sum_{i \in I^*}|I^*|\phi(G)M_i + N_i\right]}.$$

Now $\sum_{i \in I^*} M_i = 2M$ since we count each switch twice, once from i and once to i. On the other hand each action which is adjacent to n actions in I^* has been overcounted n times. Since $n \leq |I^*|\phi(G)$ we have $\sum_{i \in I^*} N_i \leq |I^*|\phi(G)N$. □

Proof of Lemma 2.6.14. First let us consider the amount of regret the player incurs for picking action adjacent to two actions in I^* N times. To do this we consider the number of times $1/2 + W_t > 5/6$. The expected number of times this occurs is

$$\mathbb{E}\sum_{t=1}^{T} \chi_{1/2+W_t>5/6} \leq \sum_{t=1}^{T} \mathbb{P}\left(|W_t| + \frac{1}{2} \geq \frac{5}{6}\right) \leq \sum_{t=1}^{T} e^{-\frac{1}{d(\rho)\sigma^2}} \leq \sum_{t=1}^{T} e^{-\frac{9\log(T)}{2}} \leq 1.$$

Thus in expectation the regret for picking an action adjacent to actions in I^* N times is at least $(1/6 + \epsilon)(N - 1)$. Let χ denote the uniform random variable over actions in I^*, which picks the best action in the beginning of the game. Denote by B_i the number of times action $i \in I^*$ was played. Then $\mathbb{E}[R'] \geq \mathbb{E}[\epsilon(T - N - B_\chi) + M + N/6]$. Thus we have

$$\begin{aligned}\mathbb{E}[R'] &= \frac{\sum_{i\in I^*} \mathbb{E}[\epsilon(T - N - B_i) + M + (N-1)/6|\chi = i]}{|I^*|} \\ &= \epsilon T - \frac{\epsilon}{|I^*|}\sum_{i\in I^*} \mathbb{E}_{\mathcal{Q}_i}[B_i] + \mathbb{E}\left[M + \frac{N-1}{6} - \epsilon N\right].\end{aligned}$$

Consider $\mathbb{E}_{\mathcal{Q}_i}[B_i]$, we have

$$\mathbb{E}_{\mathcal{Q}_i}[B_i] - \mathbb{E}_{\mathcal{Q}_0}[B_i] = \sum_{t=1}^{T}(\mathcal{Q}_i(a_t = i) - \mathcal{Q}_0(a_t = i)) \leq T \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right).$$

Thus we get

$$\begin{aligned}\sum_{i\in I^*} \mathbb{E}_{\mathcal{Q}_i}[B_i] &\leq T\sum_{i\in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) + \sum_{i\in I^*} \mathbb{E}_{\mathcal{Q}_0}[B_i] \\ &\leq T\sum_{i\in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) + T - \mathbb{E}_{\mathcal{Q}_0}[N].\end{aligned}$$

Using the assumption that $|I^*| \geq 2$, the above implies

$$\mathbb{E}[R'] \geq \frac{\epsilon T}{2} - \frac{\epsilon T}{|I^*|}\sum_{i\in I^*} \mathrm{d}_{\mathrm{TV}}^{\mathcal{F}}\left(\mathcal{Q}_0, \mathcal{Q}_i\right) + \mathbb{E}\left[M + \frac{N-1}{6} - \epsilon N\right] + \frac{\epsilon}{2}\mathbb{E}_{\mathcal{Q}_0}[N].$$

Since $\epsilon = \tilde{\Theta}(T^{-1/3})$ we have $\mathbb{E}\left[M + \frac{N-1}{6} - \epsilon N\right] + \frac{\epsilon}{2}\mathbb{E}_{\mathcal{Q}_0}[N] \geq \mathbb{E}\left[M + \frac{N}{7}\right]$ □

Chapter 3

Corralling stochastic bandit algorithms

We study the problem of corralling stochastic bandit algorithms, that is combining multiple bandit algorithms designed for a stochastic environment, with the goal of devising a corralling algorithm that performs almost as well as the best base algorithm. We give two general algorithms for this setting, which we show benefit from favorable regret guarantees. We show that the regret of the corralling algorithms is no worse than that of the best algorithm containing the arm with the highest reward, and depends on the gap between the highest reward and other rewards. The main contributions of this chapter are based on Arora et al. (2021). This work was done in collaboration with Dr. Raman Arora and Dr. Mehryar Mohri.

3.1 The corralling problem

In the corralling problem the player is tasked with selecting, at each round, one out of a fixed collection of bandit algorithms and playing the action returned by that algorithm. Note that the player does not directly select an arm, but only a base algorithm. She never requires knowledge of the action set of each base algorithm. The objective of the player is to achieve a large cumulative reward or a small pseudo-regret, over the course of her interactions with the environment. Further, the player only sends feedback to the base algorithms based on the observed reward. This complicates the problem as the state of the base learners at a fixed round depends on the player's strategy and selections made up to the round. Thus the best base learner might seem sub-optimal in the beginning of the game if it has not been selected enough times and a poor exploration strategy by the player might discard it. Solving the corralling problem amounts to policy regret minimization because the observed reward at every round of the game depends on the state of the base learners which in turn depends on all prior actions chosen by the player. Because policy regret minimization is impossible without restricting the adversary it is natural to make additional assumptions how the base learners behave. To this end we will require that the best of the base learners satisfies an any-time regret guarantee, that is at any time t, the i^*-th base algorithm, $\mathcal{A}_{i^*}$ will have pseudo-regret $\mathbb{E}[R_{i^*}(t)] \leq \bar{R}_{i^*}(t) \leq o(t)$. Further, we are going to investigate the stochastic setting, in which each of the base learners is solving an instance of the stochastic multi-armed bandit problem. The main regret upper bound in our work depends on the difference between the expected reward of the best arm of the best algorithm and the best arm of the i-th algorithm denoted by Δ_i.

Theorem 3.1.1 (Informal)**.** *If the regret of $\mathcal{A}_{i^*}$ is bounded by $\bar{R}_{i^*}(t) \leq O(\sqrt{t}), \forall t \in [T]$, then there exists an algorithm which guarantees a regret bound of the order*

$$\mathbb{E}[R(T)] = T\mu_{1,1} - \mathbb{E}\left[\sum_{t=1}^{T} r_t(a_{i_t,j_t})\right] \leq O\left(\sum_{i \neq i^*} \frac{(\log(T))^2}{\Delta_i} + \log(T)\,\bar{R}_{i^*}(T)\right),$$

where $\mu_{1,1}$ is the expected reward of the best arm of the best algorithm and $r_t(a_{i_t,j_t})$ is the reward of the j_t-th action played by the algorithm, i_t, selected at time t.

The corralling problem was first investigated by Odalric and Munos (2011) by combining Exp3 or UCB-I base learners through an Exp4 (Auer et al., 2002b) strategy with added uniform exploration. The achieved regret bounds are of the order $O(T^{2/3})$. Agarwal et al. (2016) study corralling problem in the adversarial setting, that is the rewards/losses on which the base learners compete are assumed to be adversarially generated. The authors assume that the base learners obey certain stability, which we will address shortly, and design a corralling strategy with favorable regret guarantees dependent on the worst-case regret guarantees of all learners. Their algorithm is based on OMD, however, instead of using a fixed step-size, a special non-decreasing step-size is employed, which grows whenever the probability to play a base learner becomes too small. The authors were not able to demonstrate that the proposed corralling strategy enjoys regret guarantees better than $\Omega(\sqrt{T})$. Parallel, work (Cutkosky et al., 2020), studies the same stochastic setting as our work and actually improves on the regret bounds stated in Theorem 3.1.1 by removing the additional logarithmic factors with the caveat that the time-horizon is known before the start of the game.

3.2 The model selection problem

Corralling can also be seen as a form of model selection, in which the player is tasked with selecting the best base algorithm for an apriori unknown environment, e.g., choose between a contextual bandits algorithm such as Exp4 or a MAB algorithm such as UCB-I. Recently the model selection problem has received a lot of attention in the linear stochastic bandits setting which we now describe carefully.

3.2.1 Linear contextual bandits

In the contextual bandit problem the player receives additional information before she is required to select an action from her action set. The additional information is in the form of a context $x_t \in \mathcal{X}$, where $\mathcal{X}$ is some very large (possibly infinite) set of contexts. In practice contexts can be any characteristics associated with the action. For example, in the adds allocation problem, in which the player has to choose among multiple adds to display for a given user, the player might also get additional information about the user such as their age, music preferences, if they like ice-cream, etc. Each of these attributes can be modeled by a contextual vector. Contexts can be sampled from some unknown distribution, $\mathcal{D}$, or generated adversarially. The rewards in the contextual bandit problems also take into account the current context, i.e., $r_t : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$. We are interested in the following variant of the contextual problem – there exists an embedding $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$ which maps context-action pairs into a d-dimensional vector space. The rewards now satisfy $r_t(x_t, a_t) = \langle \beta, \phi(x_t, a_t) \rangle + \xi_t$, where $\beta \in \mathbb{R}^d$ is unknown and ξ_t is random noise sampled from a sub-Gaussian distribution with variance proxy bounded by 1. Further, the contexts are sampled from some distribution which is unknown to the player. The player only observes $r_t(x_t, a_t)$ and the goal is to minimize the regret

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} \left(\max_{a \in \mathcal{A}} \mathbb{E}[\langle \beta, \phi(x_t, a) \rangle] - \mathbb{E}[\langle \beta, \phi_{i^*}(x_t, a_t) \rangle] \right).$$

The contextual stochastic bandit problem was first investigated by Abe and Long (1999), with the first algorithm based on the OFU principle proposed by Auer (2002) with regret bounded by $\tilde{O}(\sqrt{dT \log{(|\mathcal{A}|)}})$, where the $\tilde{O}$ notation only hides poly-logarithmic factors in the horizon T. Li et al. (2010); Chu et al. (2011) propose the LinUCB and SupLinUCB algorithms, respectively, for the setting in which $\mathcal{A}$ is finite. These algorithms are based on the OFU principle with bonuses based on a confidence ellipsoid around an estimator or β and enjoy a regret bound of the order $\tilde{O}(\sqrt{dT \log{(|\mathcal{A}|)}})$. For infinite action sets, Abbasi-Yadkori et al. (2011), propose the OFUL algorithm which has regret upper bounded as $\tilde{O}(d\sqrt{T})$. It turns out that both OFUL and SupLinUCB achieve the min-max optimal regret, up to poly-logarithmic factors for their respective settings. Lattimore et al. (2020) propose an approach based on Optimal design for the least squares problem (Kiefer and Wolfowitz, 1960), which can handle miss-specification in the linear model, that is the observed losses are not linear, however, can be approximated by a linear loss up to ϵ. If ϵ is known, the authors propose an algorithm which enjoys a $O(\sqrt{dT \log{(|\mathcal{A}|)}} + \epsilon T \sqrt{d} \log{(n)})$ regret bound. Foster et al. (2020a) remove the requirement that ϵ is known and achieve similar regret bounds through a corralling approach.

3.2.2 Model selection for linear bandits

In the model selection problem for linear bandits, the player is given a nested sequence of classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_K = \mathcal{F}$, where each $\mathcal{F}_i$ is defined as

$$\mathcal{F}_i = \{(x,a) \to \langle \beta_i, \phi_i(x,a)\rangle : \beta_i \in \mathbb{R}^{d_i}\},$$

for some feature embedding $\phi_i \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_i}$. It is assumed that each feature embedding ϕ_i contains ϕ_{i-1} as its first d_{i-1} coordinates. It is further assumed that there exists a smallest $i^* \leq K = |\mathcal{A}|$ to which the optimal parameter β^* belongs, that is the observed rewards for each context-action pair (x,a) satisfy $\mathbb{E}[r_t(x,a)] = \mathbb{E}\left[\langle \beta^*, \phi_{d_{i^*}}(x,a)\rangle\right]$. The goal in the model selection problem is to identify i^* and compete against the smallest loss for the t-th context in $\mathbb{R}^{d_{i^*}}$ by minimizing the regret $R_{i^*}(T)$. Foster et al. (2019) propose an algorithm which does not incur more than $\tilde{O}\left(\frac{1}{\gamma^3}(i^*T)^{2/3}(Kd_{i^*})^{1/3}\right)$, where γ^3 is the smallest eigenvalue of the covariance matrix of feature embeddings $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{K}\sum_{a\in\mathcal{A}}\phi_K(x,a)\phi_K(x,a)^\top\right]$. Pacchiano et al. (2020b) propose a different approach based on the corralling algorithm of Agarwal et al. (2016) which enjoys a $\tilde{O}(d_{i^*}\sqrt{T})$ regret bound for finite action sets and $\tilde{O}(d_{i^*}^2\sqrt{T})$ bound for arbitrary action sets $\mathcal{A}$. Later, Pacchiano et al. (2020a) design an algorithm which enjoys a gap-dependent guarantee under the assumption that all of the miss-specified models have regret $R_i(t) \geq \Delta t, \forall t \in [T]$. Under such an assumption, the authors recover a regret bounds of the order $\tilde{O}(d_{i^*}\sqrt{T} + d_{i^*}^4/\Delta)$ for arbitrary action sets. Cutkosky et al. (2020) also manage to recover the $O(d_{i^*}\sqrt{T})$ and $O(d_{i^*}^2\sqrt{T})$ bounds for the model selection problems through their corralling algorithm. Our corralling strategy is also able to recover these bounds and further enjoys a certain gap-dependent guarantee as well, similar in spirit to the bounds shown by Pacchiano et al. (2020a). Our model selection result is stated below.

Theorem 3.2.1. *Assume that every base learner $\mathcal{A}_i$, $i \geq i^*$, admits a $\tilde{O}(d_i^\alpha\sqrt{T})$ regret. Then, there exists a corralling strategy with expected regret bounded by $\tilde{O}(d_{i^*}^{2\alpha}\sqrt{T} +$*

$K\sqrt{T}$). Moreover, under the additional assumption that the following holds for any $i < i^*$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$

$$\max_{a\in\mathcal{A}} \mathbb{E}[\langle \beta^*, \phi_{i^*}(x,a)\rangle] - \mathbb{E}[\langle \beta_i, \phi_i(x,a)\rangle] \geq 2\frac{d_{i^*}^{2\alpha} - d_i^{2\alpha}}{\sqrt{T}},$$

the expected regret of the same strategy is bounded as $\tilde{O}(d_{i^*}^{\alpha}\sqrt{T} + K\sqrt{T})$.

All of the above results fall short from a true model-selection algorithm which would enjoy a regret guarantee of the order $O(d_{i^*}^{\alpha}T^{1-\alpha})$ for the finite action set case, where $\alpha \in (0, 1)$. It turns out that this is not just a shortcoming of our and prior work, but is rather information theoretically impossible as shown by Zhu and Nowak (2021). Zhu and Nowak (2021) provide a detailed characterization of the possible model selection rates by showing both information theoretic lower bounds and an algorithm with regret guarantees which nearly match the lower bounds.

3.3 Preliminaries and additional notation for the corralling problem

We consider the problem of corralling K stochastic multi-armed bandit algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_K$, which we often refer to as ***base algorithms*** (base learners). At each round t, a ***corralling algorithm*** selects a base algorithm $\mathcal{A}_{i_t}$, which plays action a_{i_t,j_t}. The corralling algorithm is not informed of the identity of this action but it does observe its reward $r_t(a_{i_t,j_t})$. The top algorithm then updates its decision rule and provides feedback to each of the base learners $\mathcal{A}_i$. We note that the feedback may be just the empty set, in which case the base learners do not update their state. We will also assume access to the parameters controlling the behavior of each $\mathcal{A}_i$ such as the step size for mirror descent-type algorithms, or the confidence bounds for UCB-type algorithms. Our goal is to minimize the cumulative pseudo-regret of the corralling algorithm as defined in Equation 3.1:[1]

$$\mathbb{E}[R(T)] = T\mu_{1,1} - \mathbb{E}\left[\sum_{t=1}^{T} r_t(a_{i_t,j_t})\right], \tag{3.1}$$

where $\mu_{1,1}$ is the mean reward of the best arm.

We denote by e_i the ith standard basis vector, by $\mathbf{0}_K \in \mathbb{R}^K$ the vector of all 0s, and by $\mathbf{1}_K \in \mathbb{R}^K$ the vector of all 1s. For two vectors $x, y \in \mathbb{R}^K$, $x \odot y$ denotes their Hadamard product. We also denote the line segment between x and y as $[x, y]$.

For the base algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_K$, let $T_i(t)$ be the number of times algorithm $\mathcal{A}_i$ has been played until time t. Let $T_{i,j}(t)$ be the number of times action j has been proposed by algorithm $\mathcal{A}_i$ until time t. Let $[k_i]$ denote the set of arms or action set of algorithm $\mathcal{A}_i$. We denote the reward of arm j in the action set of algorithm i at time t as $r_t(a_{i,j})$ and denote its mean reward by $\mu_{i,j}$. We also use a_{i,j_t} to denote the arm proposed by algorithm $\mathcal{A}_i$ during time t. Further, the algorithm played at time t is denoted as i_t, its action played at time t is a_{i_t,j_t} and the reward for that action is $r_t(a_{i_t,j_t})$ with mean μ_{i_t,j_t}. Let i^* denote the index of the base algorithm that contains the arm with the highest mean reward. Without loss of generality, we will assume that $i^* = 1$. Similarly, we assume that $a_{i,1}$ is the arm with highest reward in algorithm $\mathcal{A}_i$. We assume that the best arm of the best algorithm has a gap to the best arm of every other algorithm. We denote the gap between the best arm of $\mathcal{A}_1$ and the best arm of $\mathcal{A}_i$ as Δ_i: $\Delta_i = \mu_{i^*,1} - \mu_{i,1} > 0$ for $i \neq i^*$. Further, we denote the intra-algorithm gaps by $\Delta_{i,j} = \mu_{i,1} - \mu_{i,j}$. We denote by $\bar{R}_i(t)$ an upper bound on the regret of algorithm $\mathcal{A}_i$ at time t and by $R_i(t)$ the actual regret of $\mathcal{A}_i$, so that $\mathbb{E}[R_i(t)]$ is the expected regret of algorithm $\mathcal{A}_i$ at time t.

[1] For conciseness, from now on, we will simply write ***regret*** instead of ***pseudo-regret***.

3.4 Lower bounds without anytime regret guarantees

As we already mentioned in Section 3.1 we require that each of the base learners satisfy an any-time regret bound. We now show that this assumption is necessary through a simple lower bound. Our lower bound is based on corralling base algorithms that only admit a fixed-time horizon regret bound and do not enjoy anytime regret guarantees. We further assume that the corralling strategy cannot simulate anytime regret guarantees on the base algorithms, say by using the so-called doubling trick. This result suggests that the base algorithms must admit a strong regret guarantee during every round of the game.

The key idea behind our construction is the following. Suppose one of the corralled algorithms, $\mathcal{A}_i$, incurs a linear regret over the first $R_i(T)$ rounds. In that case, the corralling algorithm is unable to distinguish between $\mathcal{A}_i$ and an another algorithm that mimics the linear regret behavior of $\mathcal{A}_i$ throughout all T rounds, unless the corralling algorithm plays $\mathcal{A}_i$ at least $R_i(T)$ times. Formally, assume that the corralling algorithm can play one of two algorithms, $\mathcal{A}_1$ or $\mathcal{A}_2$, with the rewards of each arm played by these algorithms distributed according to a Bernoulli random variable. Algorithm $\mathcal{A}_1$ plays a single arm with expected reward μ_1 and algorithm $\mathcal{A}_2$ is defined as follows.

Let β be drawn according to the Bernoulli distribution $\beta \sim \text{Ber}(\frac{1}{2})$ and let α be drawn uniformly over the unit interval, $\alpha \sim \text{Unif}[0,1]$. If $\beta = 1$, $\mathcal{A}_2$ alternates between playing an arm with mean μ_2 and an arm with mean μ_3 every round, so that the algorithm incurs linear regret. We set μ_i such that $\mu_2 > \mu_1 > \frac{\mu_2+\mu_3}{2}$. If $\beta = 0$, then $\mathcal{A}_2$ behaves in the same way as if $\beta = 1$ for the first $T^{(1-\alpha)}$ rounds and for the remaining $T - T^{(1-\alpha)}$ rounds $\mathcal{A}_2$ only pulls the arm with mean μ_2. Notice that, in this setting, $\mathcal{A}_2$ admits sublinear regret almost surely.

We denote by $\mathbb{P}(\cdot|r_1(a_{i_1,j_1}), \ldots, r_t(a_{i_t,j_t}), \beta = i)$ the natural measure on the σ-

algebra generated by the observed rewards under the environment $\beta = i$ and all the randomness of the player's algorithm. To simplify the notation, we denote by $r_{1:t}$ the sequence $\{r_s(a_{i_s,j_s})\}_{s=1}^{t}$. Let N denote the random variable counting the number of times the corralling strategy selected $\mathcal{A}_1$. Information-theoretically, the player can obtain a good approximation of μ_1 in time $O(\log(T))$ and, therefore, for simplicity, we assume that the player knows μ_1 exactly. Note that this can only make the problem easier for the player. Given this information, we can assume that the player begins by playing algorithm $\mathcal{A}_2$ for $T - N + 1$ rounds and then switches to $\mathcal{A}_1$ for the rest of the game. In particular, we assume that $T - N + 1$ is the time when the player can figure out that $\beta = 1$. We note that at time $T^{(1-\alpha)}$ we have $\mathbb{P}(\cdot|r_{1:T^{(1-\alpha)}}, \beta = 1) = \mathbb{P}(\cdot|r_{1:T^{(1-\alpha)}}, \beta = 0)$, as the distribution of the rewards provided by $\mathcal{A}_2$ do not differ between $\beta = 1$ and $\beta = 0$. Furthermore, any random strategy would also need to select algorithm $\mathcal{A}_2$ at least $T^{(1-\alpha)} + 1$ rounds before it is able to distinguish between $\beta = 1$ or $\beta = 0$. It is also important to note that under the event that $\beta = 1$, the corralling algorithm does not receive any information about the value of α. This allows us to show that in the setting constructed above, with at least constant probability the best algorithm i.e., $\mathcal{A}_1$ when $\beta = 1$ and $\mathcal{A}_2$ when $\beta = 0$, has sublinear regret. Finally, a direct computation of the regret of this corralling strategy gives the following result.

Theorem 3.4.1. *Let algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ follow the construction in Section 3.4. Then, with probability at least $1/2$ over the random choice of α, any corralling strategy incurs regret at least $\tilde{\Omega}(T)$, while the regret of the best algorithm is at most $O(\sqrt{T})$.*

Lower bound for successive elimination. The behavior of $\mathcal{A}_2$ for the setting given by $\beta = 0$, in the construction above, may seem somewhat artificial: a stochastic bandit algorithm may not be expected to behave in that manner when the gap between μ_2 and μ_3 is large enough. Here, we describe how to set μ_1, μ_2 and μ_3 such that the successive elimination algorithm (Even-Dar et al., 2002) admits a similar behavior to $\mathcal{A}_2$ with $\beta = 0$. Recall that successive elimination needs at least $1/\Delta^2$ rounds to distinguish between the arm with mean μ_2 and the arm with mean μ_3. In other words, for at least $1/\Delta^2$ rounds, it will alternate between the two arms. Therefore, we set $\frac{1}{\Delta^2} = T^{(1-\alpha)}$ or, equivalently, $\Delta = \frac{1}{T^{(1-\alpha)/2}}$, and $\mu_1 = \mu_2 - \frac{1}{4T^{(1-\alpha)/2}}$ to yield behavior similar to $\mathcal{A}_2$. For this construction, we show the following lower bound.

Theorem 3.4.2. *Let algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ follow the construction in Section 3.4. With probability at least $1/4$ over the random choice of α any corralling strategy will incur regret at least $\tilde{\Omega}(\sqrt{T})$ while the gap between μ_2 and μ_3 is such that $\Delta > \omega(T^{-1/4})$ and hence the regret of the best algorithm is at most $o(T^{1/4})$.*

We note that, in our construction, if $\beta = 1$, then the inequality $\Delta \gg \frac{1}{\sqrt{T}}$ holds almost surely. In this setting, the instance-dependent regret bound for $\mathcal{A}_2$ and successive elimination is asymptotically smaller compared to the worst-case instance-independent regret bounds for stochastic bandit algorithms, which scale as $\tilde{O}(\sqrt{T})$ with the time horizon. This suggests that, even though $\mathcal{A}_2$ enjoys asymptotically better regret bounds than $\tilde{O}(\sqrt{T})$, the corralling algorithm will necessarily incur $\tilde{\Omega}(\sqrt{T})$ regret.

A lower bound when a worst case regret bound is known. Next, suppose that we know a worst case regret bound of $R_2(T)$ for algorithm $\mathcal{A}_2$. As before, we sample β according to a Bernoulli distribution. If $\beta = 1$, then algorithm $\mathcal{A}_2$ has a single arm with reward distributed as $\mathrm{Ber}((\mu_2+\mu_3)/2)$; in that case, $\mathcal{A}_2$ admits a regret equal to 0. If $\beta = 0$, then $\mathcal{A}_2$ has two arms distributed according to $\mathrm{Ber}(\mu_2)$ and $\mathrm{Ber}(\mu_3)$, respectively. We sample $\alpha \sim \mathrm{Unif}[0, 1]$, and let $\mathcal{A}_2$ play an arm uniformly at random for the first $R_2(T)^{(1-\alpha)}$ rounds. In particular, during each of the first $R_2(T)^{(1-\alpha)}$ rounds, $\mathcal{A}_2$ plays with equal probability the arm with mean μ_2 and the arm with mean μ_3. On round $R_2(T)^{(1-\alpha)}$, the algorithm switches to playing μ_1 until the rest of the game. Notice that the rewards up to time $R_2(T)^{(1-\alpha)}$, whether $\beta = 1$ or $\beta = 0$, have the same distribution. Hence, $\mathbb{P}(\cdot|r_{1:R_2(T)^{(1-\alpha)}}, \beta = 1) = \mathbb{P}(\cdot|r_{1:R_2(T)^{(1-\alpha)}}, \beta = 0)$. Then, following the arguments in the proof of Theorem 3.4.1, we can prove the following lower bound.

Theorem 3.4.3. *Let algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ follow the construction in Section 3.4. Suppose that the worst case known regret bound for Algorithm is $R_2(T)$. With probability at least $1/2$ over the random choice of α any corralling strategy will incur regret at least $\tilde{\Omega}(R_2(T))$ while the regret of $\mathcal{A}_2$ is at most $O(\sqrt{R_2(T)})$.*

3.5 Detailed proofs for Section 3.4

Proof of Theorem 3.4.1. Let $R(T)$ denote the regret of the corralling algorithm. Direct computation shows that if $\beta = 1$ the corralling regret is

$$\mathbb{E}[R(T)|\beta = 1, r_{1:T^{(1-\alpha)}}, \alpha] = \mathbb{E}\left[\left(\mu_1 - \frac{\mu_2+\mu_3}{2}\right)(T-N)|\beta = 1, r_{1:T^{(1-\alpha)}}, \alpha\right]$$

Further if $\beta = 0$ and $\mathcal{A}_2$ is the best algorithm the regret of corralling is

$$\begin{aligned}
&\mathbb{E}[R(T)|\beta = 0, r_{1:T^{(1-\alpha)}}, \alpha] = \mathbb{E}\left[T^{(1-\alpha)}\mu_2 + (T - T^{(1-\alpha)})\mu_2|\beta = 0, r_{1:T^{(1-\alpha)}}, \alpha\right]\\
\geq&\mathbb{E}\left[\frac{\mu_2+\mu_3}{2}T^{(1-\alpha)} + \mu_2(T - T^{(1-\alpha)})\right.\\
&- \mu_1 N - \chi_{(N \leq T - T^{(1-\alpha)})}\left(\frac{\mu_2+\mu_3}{2}T^{(1-\alpha)} + \mu_2(T - T^{(1-\alpha)} - N)\right)\\
&\left.- \chi_{(N > T - T^{(1-\alpha)})}\frac{\mu_2+\mu_3}{2}(T-N)|\beta = 0, r_{1:T^{(1-\alpha)}}, \alpha\right],
\end{aligned}$$

where the characteristic functions describe the event in which we pull $\mathcal{A}_1$ less times than is needed for $\mathcal{A}_2$ to switch to playing the best action. Notice that the total regret for corralling is at least the above as we also need to add the regret of the best algorithm to the above.

We first consider the case $\beta = 1$. Notice that in this case the corralling algorithm does not receive any information about α because $\mathcal{A}_2$ alternates between μ_2 and μ_3 at all rounds. This implies $\mathbb{E}[R(T)|\beta = 1, \alpha] = \mathbb{E}[R(T)|\beta = 1]$. Condition on the event

$N \leq T - T^{(1-\alpha)}$. We have

$$\begin{aligned}
&\mathbb{E}[R(T)|\beta = 1, N \leq T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha] = \\
&\quad \mathbb{E}[R(T)|\beta = 1, N \leq T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}] \geq \left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right) \mathbb{E}[T^{(1-\alpha)}|\beta = 1] \\
&\qquad = \left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right) \mathbb{E}[T^{(1-\alpha)}] \\
&\qquad = \left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right) \frac{T-1}{\log{(T)}},
\end{aligned}$$

where in the first inequality we have replaced N by $T - T^{1-\alpha}$. Next consider the case $\beta = 0$. Condition on the event $N > T - T^{(1-\alpha)}$. We have

$$\begin{aligned}
&\mathbb{E}[R(T)|\beta = 0, N > T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha] \\
&= \mathbb{E}\left[\frac{\mu_2 - \mu_3}{2}(T - T^{(1-\alpha)}) - \left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right) N|\beta = 0, N > T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha\right] \\
&\geq \mathbb{E}\left[(\mu_2 - \mu_1)T - \frac{\mu_2}{2}T^{(1-\alpha)}|\beta = 0, N > T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha\right] \\
&= \mathbb{E}\left[(\mu_2 - \mu_1)T - \frac{\mu_2}{2}T^{1-\alpha}|\alpha\right],
\end{aligned}$$

where in the inequality we have used the fact that $N > T - T^{1-\alpha}$ to bound $-\mu_1 N$ and $T \geq N$ to bound $\frac{\mu_1+\mu_3}{2}N$. Let A denote the event $N \leq T - T^{(1-\alpha)}$. We are now ready to lower bound the regret of the player's strategy as follows.

$$\begin{aligned}
\mathbb{E}\left[R(T)|\alpha\right] &= \frac{1}{2}\mathbb{E}\left[\mathbb{E}[R(T)|r_{1:T^{(1-\alpha)}}, \beta = 1, \alpha] + \mathbb{E}[R(T)|r_{1:T^{(1-\alpha)}}, \beta = 0, \alpha]|\alpha\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[\mathbb{P}(A|r_{1:T^{(1-\alpha)}}, \beta = 1, \alpha)\mathbb{E}[R(T)|r_{1:T^{(1-\alpha)}}, \beta = 1, A, \alpha]\right. \\
&\quad \left. + \mathbb{P}(A^c|r_{1:T^{(1-\alpha)}}, \beta = 1, \alpha)\mathbb{E}[R(T)|r_{1:T^{(1-\alpha)}}, \beta = 0, A^c, \alpha]|\alpha\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[\mathbb{P}(A|r_{1:T^{(1-\alpha)}}, \beta = 1, \alpha)\left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right)\frac{T-1}{2\log{(T)}}\right. \\
&\quad \left. + (1 - \mathbb{P}(A|r_{1:T^{(1-\alpha)}}, \beta = 1), \alpha)(\mu_2 - \mu_1)T - \frac{\mu_2}{2}T^{1-\alpha}|\alpha\right],
\end{aligned}$$

where in the first inequality we have used the fact that the conditional measures induced by $\beta = 1$ and $\beta = 0$ are equal for the first $T^{1-\alpha}$ rounds. Because $\alpha \geq 1/2$ with probability at least $1/2$ it holds that the random variable $\mathbb{E}[R(T)|\alpha] > \tilde{\Omega}(T)$ with probability at least $1/2$ and that the regret of $\mathcal{A}_2$ when $\beta = 1$ is at most $O(\sqrt{T})$. $\square$

Proof of Theorem 3.4.2. From the proof of Theorem 3.4.1 we can compute, when $\beta = 1$, we can directly compute

$$\begin{aligned}
&\mathbb{E}\left[R(T)|\beta = 1, N \leq T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha\right] \\
\geq &\mathbb{E}\left[\left(\mu_1 - \frac{\mu_2 + \mu_3}{2}\right) T^{(1-\alpha)}|\beta = 1, N \leq T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha\right] \\
= &\mathbb{E}\left[\frac{1}{4T^{(1-\alpha)/2}} T^{(1-\alpha)}|\beta = 1, N \leq T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}\right] = \frac{\sqrt{T} - 1}{2\log(T)},
\end{aligned}$$

Where in the equality we again used the fact that if $\beta = 1$, the corralling algorithm receives no information about α. Further when $\beta = 0$ we have

$$\begin{aligned}
&\mathbb{E}\left[R(T)|\beta = 0, N > T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}(a_{T^{(1-\alpha)}}), \alpha\right] \\
\geq &\mathbb{E}\left[(\mu_2 - \mu_1)T - \frac{\mu_2}{2} T^{(1-\alpha)}|\beta = 0, N > T - T^{(1-\alpha)}, r_{1:T^{(1-\alpha)}}, \alpha\right] \\
= &\mathbb{E}\left[\frac{T^{(1+\alpha)/2}}{4}|\alpha\right] - \mathbb{E}\left[\frac{T^{(1-\alpha)}}{2}|\alpha\right].
\end{aligned}$$

Again we note that with probability $1/2$ we have $\alpha \geq 1/2$ and the above expression becomes asymptotically larger than $\sqrt{T}$. The same computation as in the proof of Theorem 3.4.1 finishes the proof. □

3.6 UCB-style corralling algorithm

The negative result of Section 3.4 hinge on the fact that the base algorithms do not admit anytime regret guarantees. Therefore, we assume, for the rest of the paper, that the base algorithms, $\{\mathcal{A}_i\}$, satisfy the following:

$$\mathbb{E}\left[t\mu_{i,1} - \sum_{s=1}^{t} r_s(a_{i_s,j_s})\right] \leq \bar{R}_i(t), \tag{3.2}$$

for any time $t \in [T]$. For UCB-type algorithms, such bounds can be derived from the fact that the expected number of pulls, $T_{i,j}(t)$, of a suboptimal arm j, is bounded as $\mathbb{E}[T_{i,j}(t)] \leq c\frac{\log(t)}{(\Delta_{i,j})^2}$, for some time and gap-independent constant c (e.g., Bubeck (2010)), and take the following form, $\bar{R}_i(t) \leq c'\sqrt{k_i t \log(t)}$, for some constant c'.

Suppose that the bound in Equation 3.2 holds with probability $1-\delta_t$. Note that such bounds are available for some UCB-type algorithms (Audibert et al., 2009). We can then adopt the optimism in the face of uncertainty principle for each $\mu_{i,1}$ by overestimating it with $\frac{1}{t}\sum_{s=1}^{t} r_s(a_{i,j_s}) + \frac{1}{t}\bar{R}_i(t)$. As long as this occurs with high enough probability, we can construct an upper confidence bound for $\mu_{i,1}$ and use it in a UCB-type algorithm. Unfortunately, the upper confidence bounds required for UCB-type algorithms to work need to hold with high enough probability, which is not readily available from Equation 3.2 or from probabilistic bounds on the pseudo-regret of anytime stochastic bandit algorithms. In fact, as discussed in Section 3.6.1, we expect it to be impossible to corral any-time stochastic MAB algorithms with a standard UCB-type strategy. However, a simple boosting technique, in which we run $2\log(1/\delta)$ copies of each algorithm $\mathcal{A}_i$, gives the following high probability version of the bound in Equation 3.2.

Lemma 3.6.1. *Suppose we run $2\log(1/\delta)$ copies of algorithm $\mathcal{A}_i$ which satisfies Equation 3.2. If $\mathcal{A}_{med_i}$ is the algorithm with median cumulative reward at time t, then $\mathbb{P}[t\mu_{i,1} - \sum_{s=1}^{t} r_s(a_{med_i,j_s}) \geq 2\bar{R}_i(t)] \leq \delta$.*

Proof of Lemma 3.6.1. First note that $\mu_{med_i,1} = \mu_{i_s,1}$ and $\bar{R}_{i_s}(t) = \bar{R}_{med_i}(t)$for all s and t. The assumption in Equation 3.2 together with Markov's inequality implies that for every copy $\mathcal{A}_{i_s}$ of $\mathcal{A}_i$ at time t it holds that

$$\mathbb{P}\left[t\mu_{med_i,1} - \sum_{s=1}^{t} r_s(a_{i,j_s}) \geq 2\bar{R}_{med_i}(t)\right] \leq \frac{1}{2}.$$

Let $\mathcal{A}_{i_1}, \ldots, \mathcal{A}_{i_n}$ be the algorithms which have reward smaller than $\mathcal{A}_{med_i}$ at time t. We have

$$\begin{aligned}\mathbb{P}\left[t\mu_{med_i,1} - \sum_{s=1}^{t} r_s(a_{med_i,j_s}) \geq 2\bar{R}_{med_i}(t)\right] &\leq \mathbb{P}\left[\bigcap_{l\in[n]}\left\{t\mu_{l,1} - \sum_{s=1}^{t} r_s(a_{i_l,j_s}) \geq 2\bar{R}_{med_i}(t)\right\}\right]\\ &\leq \left(\frac{1}{2}\right)^{\log(1/\delta)} \leq \delta,\end{aligned}$$

where the first inequality follows from the definition of $\mathcal{A}_{med_i}$ and $\mathcal{A}_{i_l}$ for $l \in [n]$. □

Algorithm 9: UCB-C

```
Input: Stochastic bandit algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_K$
Output: Sequence of algorithms $(i_t)_{t=1}^{T}$.
$t = 1$
for $i = 1, \ldots, K$ do
    $\mathbb{A}_i = \emptyset$ % contains all copies of $\mathcal{A}_i$
    for $s = 1, \ldots, \lceil 2\log(T) \rceil$ do
        Initialize $\mathcal{A}_i(s)$ as a copy of $\mathcal{A}_i$, $\widehat{\mu}_i(s) = 0$
        Append $(\mathcal{A}_i(s), \widehat{\mu}_i(s))$ to $\mathbb{A}_i$
    end for
end for
for $i = 1, \ldots, K$ do
    Foreach $(\mathcal{A}_i(s), \widehat{\mu}_i(s)) \in \mathbb{A}_i$, play $\mathcal{A}_i(s)$, update empirical mean $\widehat{\mu}_i(s)$,
    $t = t + 2\log(T)$
    $\widehat{\mu}_{med_i} = \texttt{Median}(\{\widehat{\mu}_i(s)\}_{s=1}^{\lceil 2\log(T) \rceil})$
end for
while $t \leq T$ do
    $b_\ell(t) = \frac{\sqrt{2\bar{R}_{med_\ell}(T_{med_\ell}(t))} + \sqrt{2T_{med_\ell}(t)\log(t)}}{T_{med_\ell}(t)}, \forall \ell \in [K]$
    $i = \text{argmax}_{\ell \in [K]} \{\widehat{\mu}_{med_\ell} + b_\ell(t)\}$
    Foreach $(\mathcal{A}_i(s), \widehat{\mu}_i(s)) \in \mathbb{A}_i$, play $\mathcal{A}_i(s)$, update empirical mean $\widehat{\mu}_i(s)$,
    $t = t + 2\log(T)$
    $\widehat{\mu}_{med_i} = \texttt{Median}(\{\widehat{\mu}_i(s)\}_{s=1}^{\lceil 2\log(T) \rceil})$
end while
```

We consider the following variant of the standard UCB algorithm for corralling. We initialize $2\log(T)$ copies of each base algorithm $\mathcal{A}_i$. Each $\mathcal{A}_i$ is associated with the median empirical average reward of its copies. At each round, the corralling algorithm picks the $\mathcal{A}_i$ with the highest sum of median empirical average reward and an upper confidence bound based on Lemma 3.6.1. The pseudocode is given in Algorithm 9. The algorithm admits the following regret guarantees.

Theorem 3.6.2. *Suppose that algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_K$ satisfy the following regret bound $\mathbb{E}[R_i(t)] \leq \sqrt{\alpha k_i t \log(t)}$, respectively for $i \in [K]$. Algorithm 9 selects a sequence* *of algorithms* $i_1, \ldots, i_T$ *which take actions* $a_{i_1,j_1}, \ldots, a_{i_T,j_T}$*, respectively, such that*

$$\mathbb{E}[R(T)] \leq O\left(\sum_{i \neq i^*} \frac{k_i \log(T)^2}{\Delta_i} + \log(T)\, \mathbb{E}\left[R_{i^*}(T)\right]\right),$$

$$\mathbb{E}[R(T)] \leq O\left(\log(T) \sqrt{KT \log(T) \max_{i \in [K]}(k_i)}\right).$$

Proof sketch. The ideas behind the proof are very similar to the standard analysis of most UCB approaches. First one shows that because of the bonuses, each of the empirical estimators of the means of best arms plus respective bonuses, are overestimators of the true means. Next, because the bonuses are decreasing with the number of times each sub-optimal algorithm is played, it turns out that we can not play the i-th sub-optimal algorithm more than roughly k_i/Δ_i^2 times as the optimistic estimator of the best reward would become smaller than $\mu_{1,1}$. Finally, a union bound ensures that optimism holds throughout the T rounds which finishes the proof. □

We note that both the optimistic and the worst case regret bounds above involve an additional factor that depends on the number of arms, k_i, of the base algorithm $\mathcal{A}_i$. This dependence reflects the complexity of the decision space of algorithm $\mathcal{A}_i$. We conjecture that a complexity-free bound is not possible, in general. To see this, consider a setting where each $\mathcal{A}_i$, for $i \neq i^*$, only plays arms with equal means $\mu_i = \mu_{1,1} - \Delta_i$. Standard stochastic bandit regret lower bounds, e.g. (Garivier et al., 2018), state that any strategy on the combined set of arms of all algorithms will incur regret at least $\Omega(\sum_{i \neq i^*} k_i \log(T)/\Delta_i)$. The $\log(T)$ factor in front of the regret of the best algorithm comes from the fact that we are running $\Omega(\log(T))$ copies of it.

3.6.1 Discussion regarding tightness of bounds

A natural question is if it is possible to achieve bounds that do not have a $\log(T)^2$ scaling. After all, for the simpler stochastic MAB problem, regret upper bounds only scale as $O(\log(T))$ in terms of the time horizon. As already mentioned, the extra logarithmic factor comes from the boosting technique, or, more precisely, the need for exponentially fast concentration of the true regret to its expected value, when using a UCB-type corralling strategy. We now show that, in the absence of such strong concentration guarantees, if only a single copy of each of the base algorithms in Algorithm 9 is run, then linear regret is unavoidable.

Theorem 3.6.3. *There exist instances $\mathcal{A}_1$ and $\mathcal{A}_2$ of UCB-I and a reward distribution, such that, if Algorithm 9 runs a single copy of $\mathcal{A}_1$ and $\mathcal{A}_2$, then $\mathbb{E}[R(T)] \geq \tilde{\Omega}(\Delta_2 T)$.*

Further, for any algorithm $\mathcal{A}_1$ such that $\mathbb{P}\left[R_1(t) \geq \frac{1}{2}\Delta_{1,2}\tau\right] \geq \frac{1}{\tau^c}$, there exists a reward distribution such that if Algorithm 9 runs a single copy of $\mathcal{A}_1$ and $\mathcal{A}_2$, then $\mathbb{E}[R(T)] \geq \tilde{\Omega}((\Delta_{1,2})^c \Delta_2 T)$.

Proof sketch. The idea behind the proof is to show that the algorithm containing the best arm will not play the best arm sufficiently often with probability at least $\Omega(1/t)$ during the first t rounds for sufficiently large $t \leq T$. This in turn will imply that the corralling algorithm will mistake the best algorithm for a sub-optimal one and will only play it after exponentially many rounds (in t) resulting in regret of the order $\tilde{\Omega}(T)$. □

The requirement that the regret of the best algorithm satisfies $\mathbb{P}[R_1(t) \geq \frac{1}{2}\Delta_{1,2}\tau] \geq \frac{1}{\tau^c}$ in Theorem 3.6.3 is equivalent to the condition that the regret of the base algorithms admit only a polynomial concentration. Results in (Salomon and Audibert, 2011) suggest that there cannot be a tighter bound on the tail of the regret for anytime algorithms. It turns out that a more careful construction of the bonuses as done by Cutkosky et al. (2020) will be sufficient for achieving a regret bound of the order $O(\mathbb{E}[R_{i^*}(T)] + \sum_{i \neq i^*} \frac{k_i \log(T)}{\Delta_i})$. These bonuses are inspired by the work of Lattimore (2015).

3.7 Detailed proofs for Section 3.6

Proof of Theorem 3.6.2. For simplicity we assume that $\lceil \log(T) \rceil = \log(T)$. For the rest of the proof we let $t_\ell = T_\ell(t)$ to simplify notation. Further, since $\bar{R}_{\ell_s} = \bar{R}_\ell, \forall s \in [\log(T)]$, we use $\bar{R}_\ell$ as the upper bound on the regret for all algorithms in $\mathbb{A}_\ell$. Let $\psi_\ell(t) = 2\sqrt{\frac{2\log(t)}{t_\ell}} + \frac{\sqrt{2\bar{R}_\ell(t_\ell)}}{t_\ell}$. The proof follows the standard ideas behind analyses of UCB type algorithms. If at time t algorithm $\ell \neq 1$ is selected then one of the following must hold true:

$$\mu_{1,1} \geq \widehat{\mu}_{\bar{1}}(t_1) + \frac{\sqrt{2\bar{R}_1(t_1)} + \sqrt{2t_1 \log(t)}}{t_1}, \tag{3.3}$$

$$\widehat{\mu}_{med_\ell}(t_\ell) > \mu_{1,1} + \sqrt{\frac{2\log(t)}{t_\ell}}, \tag{3.4}$$

$$\Delta_\ell < 2\sqrt{\frac{2\log(t)}{t_\ell}} + \frac{\sqrt{2\bar{R}_\ell(t_\ell)}}{t_\ell}. \tag{3.5}$$

The above conditions can be derived by considering the case when the UCB for $\mathcal{A}_1$ is smaller than the UCB for $\mathcal{A}_\ell$ and every algorithm has been selected a sufficient number of times. Suppose that the three conditions above are false at the same time. Then we have

$$\begin{aligned}\widehat{\mu}_{\bar{1}}(t_1) + \frac{\sqrt{2\bar{R}_1(t_1)} + \sqrt{2t_1 \log(t)}}{t_1} &> \mu_{1,1} = \mu_{1,\ell} + \Delta_\ell \\ &\geq \Delta_\ell + \widehat{\mu}_{med_\ell}(t_\ell) - \sqrt{\frac{2\log(t)}{t_\ell}} \\ &\geq \widehat{\mu}_{med_\ell}(t_\ell) + \frac{\sqrt{2\bar{R}_\ell(t_\ell)} + \sqrt{2t_\ell \log(t)}}{t_\ell},\end{aligned}$$

which contradicts the assumption that algorithm $\mathcal{A}_\ell$ was selected. With slight abuse of notation we use $[k_\ell]$ to denote the set of arms belonging to algorithm $\mathcal{A}_\ell$. Next we bound the expected number of times each sub-optimal algorithm is played up to time T. Let δ be an upper bound on the probability of the event that $\widehat{\mu}_{\bar{1}}(s)$ exceeds the

UCB for $\mathcal{A}_1$.

$$\begin{aligned}
\mathbb{E}[T_\ell] &= \sum_{t=1}^{T} \mathbb{E}[\chi_{a_{i_t,j_t}\in[k_\ell]}] \leq \psi_\ell^{-1}(\Delta_\ell) + \sum_{t>\psi_\ell^{-1}(\Delta_\ell)} \mathbb{P}\left[\text{Equation 3.3 or Equation 3.4 hold}\right]\\
&\leq \psi_\ell^{-1}(\Delta_\ell) + \sum_{t>\psi_\ell^{-1}(\Delta_\ell)} \mathbb{P}\left[\exists s \in [t] : \mu_{1,1} \geq \widehat{\mu}_{\bar{1}}(s) + \frac{\sqrt{2\bar{R}_1(s)} + \sqrt{2s\log(t)}}{s}\right]\\
&+ \sum_{t>\psi_\ell^{-1}(\Delta_\ell)} \mathbb{P}\left[\exists s \in [t] : \widehat{\mu}_{med_\ell}(s) > \mu_{1,1} + \sqrt{\frac{2\log(t)}{s}}\right]\\
&\leq \sum_{t>\psi_\ell^{-1}(\Delta_\ell)} t\delta + \sum_{t>\psi_\ell^{-1}(\Delta_\ell)} \frac{1}{t} + \psi^{-1}(\Delta_\ell),
\end{aligned}$$

where the last inequality follows from the definition of δ and the fact that $\widehat{\mu}_{med_\ell}(s) \leq \widehat{\mu}_{med_\ell,1}(s)$ (empirical mean of arm 1 for algorithm $\mathcal{A}_{med_\ell}$ at time s) and the standard argument in the analysis of UCB-I. Setting $\delta = \frac{1}{T^2}$ finishes the bound on the number of suboptimal algorithm pulls. Next we consider bounding the regret incurred only by playing the median algorithms $\mathcal{A}_{med_\ell}$

$$\begin{aligned}
t\mu_{1,1} - \mathbb{E}\left[\sum_{s=1}^{t} r_s(a_{i_s,j_s})\right] &= t\mu_{1,1} - \mathbb{E}\left[\sum_\ell t_\ell\mu_{\ell,1} + \sum_\ell\sum_i T_{med_\ell,i}(t_\ell)\mu_{\ell,i} - \sum_\ell t_\ell\mu_{\ell,1}\right]\\
&= \mathbb{E}\left[\sum_{\ell\neq 1} t_\ell\Delta_\ell\right] + \sum_{\ell\neq 1}\mathbb{E}[R_{med_\ell}(t_\ell)] + \mathbb{E}\left[R_{med_1}(t)\right]\\
&\leq \sum_{\ell\neq 1}\Delta_\ell\psi^{-1}(\Delta_\ell) + 2\log(T) + \sum_{\ell\neq 1}\mathbb{E}[\sqrt{\alpha k_\ell t_\ell\log(t)}] + \mathbb{E}\left[R_{med_1}(t)\right]\\
&\leq \sum_{\ell\neq 1}\Delta_\ell\psi^{-1}(\Delta_\ell) + 2\log(T) + \sum_{\ell\neq 1}\sqrt{\alpha k_\ell\mathbb{E}[t_\ell]\log(t)} + \mathbb{E}\left[R_{med_1}(t)\right]\\
&\leq \sum_{\ell\neq 1}\Delta_\ell\psi^{-1}(\Delta_\ell) + 2\log(T) + \sum_{\ell\neq 1}\sqrt{\alpha k_\ell\psi^{-1}(\Delta_\ell)\log(t)}\\
&+ \mathbb{E}\left[R_{med_1}(t)\right] + \sqrt{\alpha k_\ell}\log(t)\,.
\end{aligned}$$

Now for the assumed regret bound on the algorithms, we have $\psi_\ell(t) = 2\sqrt{\frac{2\log(t)}{t_\ell}} + \sqrt{2\frac{\alpha k_\ell\log(t_\ell)}{t_\ell}}$. This implies that $\psi_\ell^{-1}(\Delta_\ell) \leq \frac{\alpha' k_\ell\log(t)}{\Delta_\ell^2}$, for some other constant α'. To get the instance independent bound we first notice that by Jensen's inequality we have

$$\begin{aligned}
\sum_\ell\sqrt{\alpha' k_\ell\mathbb{E}[t_\ell]\log(t)} &\leq K\sqrt{\frac{1}{K}\sum_\ell\alpha'\mathbb{E}[t_\ell]k_\ell\log(t)}\\
&\leq \sqrt{\alpha' Kt\log(t)\max_\ell(k_\ell)}.
\end{aligned}$$

Next we can bound $\mathbb{E}\left[\sum_{\ell\neq 1} t_\ell \Delta_\ell\right]$ in the following way

$$\mathbb{E}\left[\sum_{\ell\neq 1} t_\ell\Delta_\ell\right] \leq \sum_\ell \Delta_\ell \sqrt{\mathbb{E}[t_\ell]}\sqrt{\mathbb{E}[t_\ell]} = \sum_\ell \sqrt{\Delta_\ell^2 \mathbb{E}[t_\ell]}\sqrt{\mathbb{E}[t_\ell]}$$
$$= \sum_\ell \sqrt{\alpha' k_\ell \mathbb{E}[t_\ell]\log(t)} \leq \sqrt{\alpha' K t \log(t) \max_\ell(k_\ell)}$$

The theorem now follows. □

Proof of Theorem 3.6.3. Consider an instance of Algorithm 9, except that it runs a single copy of each base learner $\mathcal{A}_i$. Let $\mathcal{A}_1$ be a UCB algorithm with two arms with means $\mu_1 > \mu_2$, respectively. The arm with mean μ_1 is set according to a Bernoulli random variable, and the arm with mean μ_2 is deterministic. Let algorithm $\mathcal{A}_2$ have a single deterministic arm with mean μ_3, such that $\mu_1 > \mu_3$ and $\mu_3 > \mu_2$. Let $\Delta = \mu_1 - \mu_3$. We now follow the lower bounding technique of Audibert et al. (2009).

Consider the event that in the first q pulls of arm $a_1^{\mathcal{A}_1}$, we have $r_t(a_{1,1}) = 0$, i.e. $\mathscr{E} = \{r_1(a_{1,1}) = 0, r_2(a_{1,1}) = 0, \ldots, r_q(a_{1,1}) = 0\}$. This event occurs with probability $(1-\mu_1)^q$. Notice that on event $\mathscr{E}$, the upper confidence bound for μ_1 as per $\mathcal{A}_1$ is $\sqrt{\frac{\alpha\log(T_1(t))}{q}}$ during time t. This implies that for $a_{1,1}$ to be pulled again we need $\sqrt{\frac{\alpha\log(T_1(t))}{q}} > \mu_2$ and hence for the first $\exp(q\mu_2^2/\alpha)$ rounds in which $\mathcal{A}_1$ is selected by the corralling algorithm, $a_{1,1}$ is only pulled q times. Further, on $\mathscr{E}$, the upper confidence bound for $\mathcal{A}_1$ as per the corralling algorithm is of the form $\sqrt{\frac{2\beta\log(t)}{T_1(t)}}$. This implies that for $\mathcal{A}_1$ to be selected again we need $\mu_2 + \sqrt{\frac{2\beta\log(t)}{T_1(t)}} > \mu_3$. Let $\tilde{\Delta} = \mu_3 - \mu_2$. Then, the above implies that in the first $t \leq \exp\left(T_1(t)\tilde{\Delta}^2/(2\beta)\right)$ rounds, $\mathcal{A}_1$ is pulled at most $T_1(t)$ times. Combining with the bound for the number of pulls of $a_{1,1}$ we arrive at the fact that on $\mathscr{E}$, $a_{1,1}$ can not be pulled more than q times in the first $\exp\left(\frac{\tilde{\Delta}^2\exp\left(q\mu_2^2/\alpha\right)}{2\beta}\right)$ rounds. Let q be large enough so that $q \leq \frac{1}{2}\exp\left(\frac{\tilde{\Delta}^2\exp\left(q\mu_2^2/\alpha\right)}{2\beta}\right)$. Then, for large enough T, we have that the pseudo-regret of the corralling algorithm is $\widehat{R}(T) \geq \frac{1}{2}\Delta\exp\left(\frac{\tilde{\Delta}^2\exp\left(q\mu_2^2/\alpha\right)}{2\beta}\right)$. Taking $q = \log\left(\frac{2\beta}{\tilde{\Delta}^2}\log\left(\tau\frac{\alpha}{\mu_2^2}\right)\right)$, we get

$$\mathbb{P}\left[\widehat{R}(T) \geq \frac{1}{2}\Delta\tau\right] \geq \mathbb{P}\left[\mathscr{E}\right] = (1-\mu_1)^q = \frac{1}{\exp(q)^{\log(1/(1-\mu_1))}} = \left(\frac{\tilde{\Delta}^2}{2\beta\log(\tau)}\right)^{\frac{\alpha}{\mu_2^2}\log(1/(1-\mu_1))}.$$

Let $\gamma = \frac{\alpha}{\mu_2^2}\log(1/(1-\mu_1))$. We can now bound the expected pseudo-regret of the algorithm by integrating over $2 \le \tau \le T$, to get

$$\begin{aligned}\mathbb{E}[\widehat{R}(T)] &\geq \frac{1}{2}\Delta\int_2^T \left(\frac{\tilde{\Delta}^2}{2\beta\log(\tau)}\right)^{\frac{\alpha}{\mu_2^2}\log(1/(1-\mu_1))} d\tau = \frac{1}{2}\Delta\Big(\frac{\tilde{\Delta}^2}{2\beta}\Big)^{\gamma}\int_2^T \left(\frac{1}{\log(\tau)}\right)^{\gamma} d\tau \\ &\geq \frac{1}{2}\Delta\Big(\frac{\tilde{\Delta}^2}{2\beta}\Big)^{\gamma}\frac{T-2}{\left(\log\left(\frac{T+2}{2}\right)\right)^{\gamma}},\end{aligned}$$

where the last inequality follows from the Hermite-Hadamart inequality.

It is important to note that the above reasoning will fail if γ is a function of T. This might occur if in the UCB for $\mathcal{A}_1$ we have $\alpha = \log(T)$. In such a case the lower bounds become meaningless as $\frac{1}{\log((T+2)/2)}^{\gamma} \leq o(1/T)$. Further, it should actually be possible to avoid boosting in this case as the tail bound of the regret will now be upper bounded as $\mathbb{P}[R_1(t) \geq \Delta\tau] \leq \frac{1}{T\tau^c}$.

General Approach if Regret has a Polynomial Tail. Assume that, in general, the best algorithm has the following regret tail:

$$\mathbb{P}\left[R_1(t) \geq \frac{1}{2}\Delta_{1,1}\tau\right] \geq \frac{1}{\tau^c},$$

for some constant c. Results in Salomon and Audibert (2011) suggest that for stochastic bandit algorithms which enjoy anytime regret bounds we can not have a much tighter high probability regret bound. Let $\mathscr{E}_{T_1(t)} = \{R_1(T_1(t)) \geq T_1(t)(\mu_1 - \frac{1}{\sqrt{2}}\mu_3)\}$. After $T_1(t)$ pulls of $\mathcal{A}_1$ the reward plus the UCB for $\mathcal{A}_1$ is at most $\frac{\sum_{s=1}^{T_1(t)} r_s(a_{1,j_s})}{T_1(t)} + \sqrt{\frac{\alpha k_1 \log(t)}{T_1(t)}}$, and on $\mathscr{E}_{T_1(t)}$, we have $\frac{\sum_{s=1}^{T_1(t)} r_s(a_{1,j_s})}{T_1(t)} \leq \frac{1}{\sqrt{2}}\mu_3$. This implies that in the first t rounds, $\mathcal{A}_1$ could not have been pulled more than

$$\frac{\alpha k_1 \log(t)}{\left(\mu_3 - \frac{\sum_{s=1}^{T_1(t)} r_s(a_{1,j_s})}{T_1(t)}\right)^2} \leq \frac{2\alpha k_1 \log(t)}{\mu_3^2}.$$

Setting $T_1(t) = \frac{2\alpha k_1 \log(T)}{\mu_3^2}$, we have that $\mathscr{E}_{T_1(t)}$ occurs with probability at least $\left(\frac{\Delta_{1,2}\mu_3^2}{4\alpha k_1 \log(T)}\right)^c$ and hence the expected regret of the corralling algorithm is at least

$$\left(\frac{\Delta_{1,2}\mu_3^2}{4\alpha k_1 \log(T)}\right)^c \Delta\left(T - \frac{2\alpha k_1 \log(T)}{\mu_3^2}\right).$$

□

3.8 Corralling using Tsallis-INF

In this section, we consider an alternative approach, based on the work of Agarwal et al. (2016), which avoids running multiple copies of base algorithms. Since the approach is based on the OMD framework, which is naturally suited to losses instead of rewards, for the rest of the section we switch to losses.

We design a corralling algorithm that maintains a probability distribution $w \in \Delta^{K-1}$ over the base algorithms, $\{\mathcal{A}_i\}_{i=1}^K$. At each round, the corralling algorithm samples $i_t \sim p_t$. Next, $\mathcal{A}_{i_t}$ plays a_{i_t,j_t} and the corralling algorithm observes the loss $\ell_t(a_{i_t,j_t})$. The corralling algorithm updates its distribution over the base algorithms using the observed loss and provides an unbiased estimate $\widehat{\ell}_t(a_{i,j_t})$ of $\ell_t(a_{i,j_t})$ to algorithm $\mathcal{A}_i$: the feedback provided to $\mathcal{A}_i$ is $\widehat{\ell}_t(a_{i_t,j_t}) = \frac{\ell_t(a_{i_t,j_t})}{p_{t,i_t}}$, and for all $a_{i,j_t} \neq a_{i_t,j_t}$, $\widehat{\ell}_t(a_{i,j_t}) = 0$. Notice that $\widehat{\ell}_t \in \mathbb{R}^K$, as opposed to $\ell_t \in [0,1]^{\prod_i k_i}$. Essentially, the loss fed to $\mathcal{A}_i$, with probability $w_{t,i}$, is the true loss rescaled by the probability $w_{t,i}$ to observe the loss, and is equal to 0 with probability $1 - w_{t,i}$.

The change of environment induced by the rescaling of the observed losses is analyzed in Agarwal et al. (2016). Following Agarwal et al. (2016), we denote the environment of the original losses $(\ell_t)_t$ as $\mathcal{E}$ and that of the rescaled losses $(\widehat{\ell}_t)_t$ as $\mathcal{E}'$. Therefore, in environment $\mathcal{E}$, algorithm $\mathcal{A}_i$ observes $\ell_t(a_{i_t,j_t})$ and in environment $\mathcal{E}'$, $\mathcal{A}_i$ observes $\widehat{\ell}_t(a_{i_t,j_t})$. A few important remarks are in order. As in (Agarwal et al., 2016), we need to assume that the base algorithms admit a ***stability property*** under the change of environment. In particular, if $p_{s,i} \geq \frac{1}{\rho_t}$ for all $s \leq t$ and some $\rho_t \in \mathbb{R}$, then $\mathbb{E}[R_i(t)]$ under environment $\mathcal{E}'$ is bounded by $\mathbb{E}[\sqrt{\rho_t} R_i(t)]$. For completeness, we provide the definition of stability by Agarwal et al. (2016).

Definition 3.8.1. Let $\gamma \in (0,1]$ and let $R\colon \mathbb{N} \to \mathbb{R}_+$ be a non-decreasing function. An algorithm $\mathcal{A}$ with action space A is $(\gamma, R(\cdot))$-stable with respect to an environment $\mathcal{E}$ if its regret under $\mathcal{E}$ is $R(T)$ and its regret under $\mathcal{E}'$ induced by the importance weighting is $\max_{a \in \mathrm{A}} \mathbb{E}\left[\sum_{t=1}^{T} \widehat{\ell}_t(a_{i_t,j_t}) - \ell_t(a)\right] \leq \mathbb{E}[(\rho_T)^{\gamma} R(T)]$.

We show that UCB-I (Auer et al., 2002a) satisfies the stability property above with $\gamma = \frac{1}{2}$. The techniques used in the proof are also applicable to other UCB-type algorithms. Other algorithms for stochastic bandits like Thompson sampling and OMD/FTRL variants have been shown to be 1/2-stable in (Agarwal et al., 2016).

The corralling algorithm of Agarwal et al. (2016) is based on Online Mirror Descent (OMD), where a key idea is to increase the step size whenever the probability of selecting some algorithm $\mathcal{A}_i$ becomes smaller than some threshold. This induces a negative regret term which, coupled with a careful choice of step size (dependent on regret upper bounds of the base algorithms), provides regret bounds that scale as a function of the regret of the best base algorithm.

Unfortunately, the analysis of the corralling algorithm always leads to at least a regret bound of $\tilde{\Omega}(\sqrt{T})$ and also requires knowledge of the regret bound of the best algorithm. Since our goal is to obtain instance-dependent regret bounds, we cannot appeal to this type of OMD approach. Instead, we draw inspiration from the recent work of Zimmert and Seldin (2021), who use a Follow-the-Regularized-Leader (FTRL) type of algorithm to design an algorithm that is simultaneously optimal for both stochastic and adversarially generated losses, without requiring knowledge of instance-dependent parameters such as the sub-optimality gaps to the loss of the best arm. The overall intuition for our algorithm is as follows. We use the FTRL-type algorithm proposed by Zimmert and Seldin (2021) until the probability to sample some arm falls below a threshold. Next, we run an OMD step with an increasing step size schedule which contributes a negative regret term. After the OMD step, we resume the normal step size schedule and updates from the FTRL algorithm. After carefully choosing the initial step size rate, which can be done in an instance-independent way, the accumulated negative regret terms are enough to compensate for the increased regret due to the change of environment.

3.8.1 Algorithm and the main result

We now describe our corralling algorithm in more detail. The potential function Ψ_t used in all of the updates is defined by $\Psi_t(w) = -4\sum_{i\in[K]} \frac{1}{\eta_{t,i}}\left(\sqrt{w_i} - \frac{1}{2}w_i\right)$, where $\eta_t = \left[\eta_{t,1}, \eta_{t,2}, \ldots, \eta_{t,K}\right]$ is the step-size schedule during time t. The algorithm proceeds in epochs and begins by running each base algorithm for $\log(T) + 1$ rounds. Each epoch is twice as large as the preceding, so that the number of epochs is bounded by $\log_2(T)$, and the step size schedule remains non-increasing throughout the epochs, except when an OMD step is taken. The algorithm also maintains a set of thresholds, $\rho_1, \rho_2, \ldots, \rho_n$, where $n = O(\log(T))$. These thresholds are used to determine if the algorithm executes an OMD step, while increasing the step size:

$$\begin{aligned} w_{t+1} &= \operatorname*{argmin}_{w\in\Delta^{K-1}} \langle \widehat{\ell}_t, w\rangle + D_{\Psi_t}(w, w_t), \\ \eta_{t+1,i} &= \beta\eta_{t,i} \text{ (for } i : p_{t,i} \leq 1/\rho_{s_i}), \\ w_{t+2} &= \operatorname*{argmin}_{w\in\Delta^{K-1}} \langle \widehat{\ell}_{t+1}, w\rangle + D_{\Psi_{t+1}}(w, w_{t+1}), \rho_{s_i} = 2\rho_{s_i} \end{aligned} \tag{3.6}$$

or the algorithm takes an FTRL step

$$w_{t+1} = \operatorname*{argmin}_{w\in\Delta^{K-1}} \langle \widehat{L}_t, w\rangle + \Psi_{t+1}(w), \tag{3.7}$$

where $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t$, unless otherwise specified by the algorithm. At every iteration the algorithm also mixes w_t with the uniform distribution to ensure that the variance of the loss estimators is bounded

$$p_t = \left(1 - \frac{1}{TK}\right) w_t + \frac{1}{TK} Unif(\Delta^{K-1}). \tag{3.8}$$

We note that the algorithm can only increase the step size during the OMD step. For technical reasons, we require an FTRL step after each OMD step. Further, we require that the second step of each epoch be an OMD step if there exists at least one $p_{t,i} \leq \frac{1}{\rho_1}$. The algorithm also can enter an OMD step during an epoch if at least one $w_{t,i}$ becomes smaller than a threshold $\frac{1}{\rho_{s_i}}$ which has not been exceeded so far.

Algorithm 10: Corralling with Tsallis-INF

```
Input: Mult. constant $\beta$, thresholds $\{\rho_i\}_{i=1}^{n}$, initial step size $\eta$, epochs $\{\tau_i\}_{i=1}^{m}$,
    algorithms $\{\mathcal{A}_i\}_{i=1}^{K}$.
Output: Algorithm selection sequence $(i_t)_{t=1}^{T}$.
Initialize $t = 1$, $w_1 = Unif(\Delta^{K-1})$, $\eta_1 = \eta$
Initialize current threshold list $\theta \in [n]^K$ to $\mathbf{1}$
while $t \leq K \log(T) + K$ do
    for $i \in [K]$ do
        $\mathcal{A}_i$ plays $a_{i,j_t}$, $\widehat{L}_{1,i} += \ell_t(a_{i,j_t}), t += 1$
    end for
end while
$t = 2, w_2 = \nabla\Phi_2(-\widehat{L}_1), 1/\eta_{t+1}^2 = 1/\eta_t^2 + 1$
while $j \leq m$ do
    for $t \in \tau_j$ do
        $\mathcal{R}_t = \emptyset, \widehat{\ell}_t =$ PLAY-ROUND$(w_t)$
        if $t$ is first round of $\tau_j$ and $\exists p_{t,i} \leq \frac{1}{\rho_1}$ then
            for $i\colon p_{t,i} \leq \frac{1}{\rho_1}$ do
                $\theta_i = \min\{s \in [n]\colon w_{t,i} > \frac{1}{\rho_s}\}$, $\mathcal{R}_t = \mathcal{R}_t \bigcup \{i\}$.
            end for
            $(w_{t+3}, \widehat{L}_{t+2}) =$ NRS$(w_t, \widehat{\ell}_t, \eta_t, \mathcal{R}_t, \widehat{L}_{t-1})$, $t = t + 2$, $\widehat{\ell}_t =$ PLAY-ROUND$(w_t)$
        end if
        if $\exists i\colon p_{t,i} \leq \frac{1}{\rho_{\theta_i}}$ and prior step was not NRS then
            for $i\colon w_{t,i} \leq \frac{1}{\rho_{\theta_i}}$ do
                $\theta_i += 1$, $\mathcal{R}_t = \mathcal{R}_t \bigcup \{i\}$.
            end for
            $(w_{t+3}, \widehat{L}_{t+2}) =$ NRS$(w_t, \widehat{\ell}_t, \eta_t, \mathcal{R}_t, \widehat{L}_{t-1})$, $t = t + 2$, $\widehat{\ell}_t =$ PLAY-ROUND$(w_t)$
        else
            $1/\eta_{t+1}^2 = 1/\eta_t^2 + 1$, $w_{t+1} = \nabla\Phi_{t+1}(-\widehat{L}_t)$
        end if
    end for
end while
```

We set the probability thresholds so that $\rho_1 = O(1)$, $\rho_j = 2\rho_{j-1}$ and $\frac{1}{\rho_n} \geq \frac{1}{T}$, so that $n \leq \log_2(T)$. In the beginning of each epoch, except for the first epoch, we check if $p_{t,i} < \frac{1}{\rho_1}$. If it is, we increase the step size as $\eta_{t+1,i} = \beta\eta_{t,i}$ and run the OMD step. The pseudocode for the algorithm is given in Algorithm 10. The routines `OMD-STEP` and `PLAY-ROUND` can be found in Algorithm 12 and Algorithm 13 respectively. `OMD-STEP` essentially does the update described in Equation 3.6 and `PLAY-ROUND` samples and plays an algorithm, after which constructs an unbiased estimator of the losses and

Algorithm 11: NEG-REG-STEP(NRS)

```
Input: Prior iterate $w_t$, loss $\widehat{\ell}_t$, step size $\eta_t$, set of rescaled step-sizes $\mathcal{R}_t$, cumulative
    loss $\widehat{L}_{t-1}$
Output: Plays two rounds of the game and returns distribution $w_{t+3}$ and cumulative
    loss $\widehat{L}_{t+2}$
$(w_{t+1}, \widehat{L}_t) = $ OMD-STEP$(w_t, \widehat{\ell}_t, \eta_t, \mathcal{R}_t, \widehat{L}_{t-1})$
$\widehat{\ell}_{t+1} = $ PLAY-ROUND$(w_{t+1})$, $\widehat{L}_{t+1} = \widehat{L}_t + \widehat{\ell}_{t+1}$
for all $i$ such that $w_{t,i} \leq \frac{1}{\rho_1}$ do
    $\eta_{t+2,i} = \beta\eta_{t,i}$, $\mathcal{R}_t = \mathcal{R}_t \cup \{i\}$ and restart $\mathcal{A}_i$ with updated environment $\theta_i = \frac{1}{2w_{t,i}}$
end for
$w_{t+2} = \nabla\Phi_{t+2}(-\widehat{L}_{t+1})$
$\widehat{\ell}_{t+2} = $ PLAY-ROUND$(w_{t+2})$
$\widehat{L}_{t+2} = \widehat{L}_{t+1} + \widehat{\ell}_{t+2}, \eta_{t+3} = \eta_{t+2}, t = t+2$
$w_{t+1} = \nabla\Phi_{t+1}(-\widehat{L}_t), t = t+1$
```

Algorithm 12: OMD-STEP

```
Input: Previous iterate $w_t$, current loss $\widehat{\ell}_t$, step size $\eta_t$, set of rescaled step-sizes $\mathcal{R}_t$,
    cumulative loss $\widehat{L}_{t-1}$
Output: New iterate $w_{t+1}$, cumulative loss $\widehat{L}_t$
$\nabla\Psi_t(\tilde{w}_{t+1}) = \nabla\Psi_t(w_t) - \widehat{\ell}_t$
$w_{t+1} = \mathrm{argmin}_{w\in\Delta^{K-1}} D_{\Phi_t}(w, \tilde{w}_{t+1})$.
$\mathrm{e} = \sum_{i\in\mathcal{R}_t} \mathrm{e}_i$
$\tilde{L}_{t-1} = (\mathbf{1}_k - \mathrm{e}) \odot (\widehat{L}_{t-1} - (\nu_{t-2} + \nu_{t-1})\mathbf{1}_k) + \frac{1}{\beta}\mathrm{e} \odot ((\widehat{L}_{t-1} - (\nu_{t-2} + \nu_{t-1})\mathbf{1}_k))$ //
    $\nu_{t-2}$ and $\nu_{t-1}$ are the Lagrange multipliers from the previous two FTRL steps.
$\widehat{L}_t = \tilde{L}_{t-1} + \widehat{\ell}_t$
```

Algorithm 13: PLAY-ROUND

```
Input: Sampling distribution $w_t$
Output: Loss vector $\widehat{\ell}_t$
Sample algorithm $i_t$ according to $p_t = \left(1 - \frac{1}{TK}\right) w_t + \frac{1}{TK} Unif(\Delta^{K-1})$.
Algorithm $i_t$ plays action $a_{i_t,j_t}$. Observe loss $\ell_t(a_{i_t,j_t})$ and construct unbiased
    estimator $\widehat{\ell}_t = \frac{\ell_t(a_{i_t,j_t})}{p_{t,i_t}} \mathrm{e}_{i_t}$ of $\ell_t$.
Give feedback to $i$-th algorithm as $\widehat{\ell}_t(a_{i,j_t})$, where $a_{i,j_t}$ was action provided by $\mathcal{A}_i$
```

feeds these back to all of the sub-algorithms. We show the following regret bound for the corralling algorithm.

Theorem 3.8.1. *Let $\bar{R}_i(\cdot)$ be a function upper bounding the expected regret, $\mathbb{E}[R_i(\cdot)]$, of $\mathcal{A}_i$ for all $i \in [K]$. For $\beta = e^{1/\log(T)^2}$ and for η such that for all $i \in [K]$, $\eta_{1,i} \leq$*

$\min_{t\in[T]} \frac{\left(1-\exp\left(-\frac{1}{\log(T)^2}\right)\right)\sqrt{t}}{50\bar{R}_i(t)}$, *the expected regret of Algorithm 10 is bounded as follows:*

$$\mathbb{E}\left[R(T)\right] \leq O\left(\sum_{i\neq i^*} \frac{\log\left(T\right)}{\eta_{1,i}^2\Delta_i} + \mathbb{E}[R_{i^*}(T)]\right).$$

Proof sketch. The proof combines ideas both from (Zimmert and Seldin, 2021) and (Agarwal et al., 2016). To achieve gap-dependent regret bounds we use the self-bounding trick proposed by Wei and Luo (2018) and Zimmert and Seldin (2021). We now quickly go over the self-bounding trick for the MAB problem with K arms. First one rewrites the expected regret as $\mathbb{E}[R(T)] = \sum_{t=1}^{T}\sum_{i\neq i^*} p_{t,i}\Delta_i$. Next, using a careful analysis of FTRL yields an upper bound on the regret of the form $\mathbb{E}[R(T)] \leq \sum_{t=1}^{T}\sum_{i\neq i^*}\frac{1}{\sqrt{t}}\sqrt{p_{t,i}}$. Subtracting 1/2 of $\mathbb{E}[R(T)]$ from both sides we have that

$$\mathbb{E}[R(T)] \leq 2\sum_{i\neq i^*}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sqrt{p_{t,i}} - \Delta_i w_{t,i}.$$

The RHS of the above is bounded by noticing that $\frac{1}{\sqrt{t}}\sqrt{p_{t,i}} - \Delta_i w_{t,i} \leq O(\frac{1}{t\Delta_i})$, for all $t > \Omega(1/\Delta^2)$, which follows by maximizing the LHS with respect to $w_{t,i}$. For $t < O(1/\Delta^2)$ we can bound the sum $\sum_{t=1}^{\lceil 1/\Delta^2\rceil}\frac{1}{\sqrt{t}}\sqrt{p_{t,i}} < O(1/\Delta_i)$ through a simple integration argument. These imply that $\mathbb{E}[R(T)] \leq O\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i}\right)$.

The second part of the proof is integrating the negative regret step (OMD step) into the FTRL framework. This is done by the following lemma.

Lemma 3.8.2. *Let $\widehat{w}_{t+2}$ be defined as in Equation 3.6. Let ν_{t+1} be the constant such that $\nabla\Phi_{t+1}(-\widehat{L}_t) = \nabla\Psi^*_{t+1}(-\widehat{L}_t + \nu_t\mathbf{1}_k)$. Let $\widehat{L}_{t+1} = (\mathbf{1}_k - \mathrm{e})\odot(\widehat{L}_t - (\nu_{t-1}+\nu_t)\mathbf{1}_k) + \frac{1}{\beta}\mathrm{e}\odot((\widehat{L}_t - (\nu_{t-1}+\nu_t)\mathbf{1}_k)) + \widehat{\ell}_{t+1}$ and $\eta_{t+2} = \eta_{t+1}$. Then $(\widehat{L}_{t+1})_i \geq 0$ for all $i \in [K]$ and $\widehat{w}_{t+2} = w_{t+2} = \nabla\Phi_{t+2}(-\widehat{L}_{t+1})$.*

Finally we use the negative regret to balance the added variance from the importance weighted estimators of the losses we send as feedback. □

To parse the bound above, suppose $\{\mathcal{A}_i\}_{i\in[K]}$ are standard stochastic bandit algorithms such as UCB-I. In Theorem 3.8.4, we show that UCB-I is indeed $\frac{1}{2}$-

stable as long as we are allowed to rescale and introduce an additive factor to the confidence bounds. In this case, a worst-case upper bound on the regret of any $\mathcal{A}_i$ is $\mathbb{E}[R_i(t)] \leq c\sqrt{k_i \log{(t)}\, t}$ for all $t \in [T]$ and some universal constant c. We note that the min-max regret bound for the stochastic multi-armed bandit problem is $\Theta(\sqrt{KT})$ and most known any-time algorithms solving the problem achieve this bound up to poly-logarithmic factors. Further we note that $\left(1 - \exp\left(-\frac{1}{\log(T)^2}\right)\right) > \frac{1}{e\log(T)^2}$. This suggests that the bound in Theorem 3.8.1 on the regret of the corralling algorithm is at most $O\Big(\sum_{i\neq i^*} \frac{k_i \log(T)^5}{\Delta_i} + \mathbb{E}[R_{i^*}(T)]\Big)$. In particular, if we instantiate $\mathbb{E}[R_{i^*}(T)]$ to the instance-dependent bound of $O\bigg(\sum_{j\neq 1} \frac{\log(T)}{\Delta_{i^*,j}}\bigg)$, the regret of Algorithm 10 is bounded by $O\bigg(\sum_{i\neq i^*} \frac{k_i \log(T)^5}{\Delta_i} + \sum_{j\neq 1} \frac{\log(T)}{\Delta_{i^*,j}}\bigg)$. In general we cannot exactly compare the current bound with that of UCB-C (Algorithm 9), as the regret bound in Theorem 3.8.1 has worse scaling in the time horizon on the gap-dependent terms, compared to the regret bound in Theorem 3.6.2, but has no additional scaling in front of the $\mathbb{E}[R_{i^*}(T)]$ term. In practice we observe that Algorithm 10 outperforms Algorithm 9.

Since essentially all stochastic multi-armed bandit algorithms enjoy a regret bound, in time horizon, of the order $\tilde{O}(\sqrt{T})$, we are guaranteed that $1/\eta_{t,i}^2$ scales only poly-logarithmically with the time horizon. What happens, however, if algorithm $\mathcal{A}_i$ has a worst case regret bound of the order $\omega(\sqrt{T})$? For the next part of the discussion, we only focus on time horizon dependence. As a simple example, suppose that $\mathcal{A}_i$ has worst case regret of $T^{2/3}$ and that $\mathcal{A}_{i^*}$ has a worst case regret of $\sqrt{T}$. In this case, Theorem 3.8.1 tells us that we should set $\eta_{1,i} = \tilde{O}(1/T^{1/6})$ and hence the regret bound scales at least as $\Omega(T^{1/3}/\Delta_i + \mathbb{E}[R_{i^*}(T)])$. In general, if the worst case regret bound of $\mathcal{A}_i$ is in the order of T^α we have a regret bound scaling at least as $T^{2\alpha-1}/\Delta_i$. This is not unique to Algorithm 10 and a similar scaling of the regret would occur in the bound for Algorithm 9 due to the scaling of confidence intervals.

Corralling in an adversarial environment. Because Algorithm 10 is based on a best of both worlds algorithm, we can further handle the case when the losses/rewards are generated adversarially or whenever the best overall arm is shared across multiple algorithms, similarly to the settings studied by Agarwal et al. (2016); Pacchiano et al. (2020b).

Theorem 3.8.3. *Let $\bar{R}_{i^*}(\cdot)$ be a function upper bounding the expected regret of $\mathcal{A}_{i^*}$, $\mathbb{E}[R_{i^*}(\cdot)]$. For any $\eta_{1,i^*} \leq \min_{t\in[T]} \frac{\left(1-\exp\left(-\frac{1}{\log(T)^2}\right)\right)\sqrt{t}}{50\bar{R}_{i^*}(t)}$ and $\beta = e^{1/\log(T)^2}$ it holds that the expected regret of Algorithm 10 is bounded as follows:*

$$\mathbb{E}\left[R(T)\right] \leq O\left(\max_{w\in\Delta^{K-1}} \sqrt{T}\sum_{i=1}^{K}\frac{\sqrt{w_i}}{\eta_{1,i}} + \mathbb{E}[R_{i^*}(T)]\right).$$

The bound in Theorem 3.8.3 essentially evaluates to $O(\max(\sqrt{TK}, \max_{i\in[K]}\bar{R}_i(T)) + \mathbb{E}[R_{i^*}(T)])$ and its proof is left to Section 3.9. Unfortunately, this is not quite enough to recover the results in (Agarwal et al., 2016; Pacchiano et al., 2020b). This is attributed to the fact that we use the $\frac{1}{2}$-Tsallis entropy as the regularizer instead of the log-barrier function. It is possible to improve the above bound for algorithms with stability $\gamma < 1/2$, however, because model selection is not the primary focus of this work, we will not present such results here.

A few remarks are in order. First, when the rewards obey the stochastically constrained adversarial setting i.e., there exists a gap Δ_i at every round between the best action and every other action during all rounds $t \in [T]$, then the regret for corralling bandit algorithms with worst case regret bounds of the order $\tilde{O}(\sqrt{T})$ in time horizon is at most $\tilde{O}(\sum_{i\neq i^*}\frac{\log(T)^5}{\Delta_i} + R_{i^*}(T))$. On the other hand, if there is no gap in the rewards then a worst case regret bound is still $\tilde{O}(\max\{\sqrt{KT}, \max_i \bar{R}_i(T)\} + R_{i^*}(T))$. This implies that Algorithm 10 can be used as a model selection tool when we are not sure what environment we are playing against. For example, if we are not sure if we should use a contextual bandit algorithm, a linear bandit algorithm or a stochastic multi-armed bandit algorithm, we can corral all of them and Algorithm 10 will perform almost as well as the algorithm for the best environment. Further, if we are in a distributed setting where we have access to multiple algorithms of the same type but not the arms they are playing, we can do almost as well as an algorithm which plays on all the arms simultaneously. We believe that our algorithm will have numerous other applications outside of the scope of the above examples.

3.8.2 Stability of UCB and UCB-like algorithms under a change of environment

In this section we discuss how the regret bounds for UCB and similar algorithms change whenever the variance of the stochastic losses is rescaled by Algorithm 10. Assume that the UCB algorithm plays against stochastic rewards bounded in $[0, 1]$. We begin by noting that after every call to `OMD-STEP` (Algorithm 12) the UCB algorithm should be restarted with a change in the environment which reflects that the variance of the losses has now been rescaled. Let the UCB algorithm of interest be $\mathcal{A}_i$. If the OMD step occurred at time t' and it was the case that $\frac{1}{\rho_{s-1}} \geq p_{t',i} > \frac{1}{\rho_s}$, then we know that the rescaled rewards will be in $[0, \rho_s]$ until the next time the UCB algorithm is restarted. This suggests that the confidence bound for arm j at time t should become $\sqrt{\frac{\rho_s^2 \log(t)}{T_{i,j}(t)}}$. However, we note that the second moment of the rescaled rewards is only $\frac{\ell_t(a_{i,j_t})^2}{p_{t,i}}$. A slightly more careful analysis using Bernstein's inequality for martingales (e.g. Lemma 10 Bartlett et al. (2008)) allows us to show the following.

Theorem 3.8.4. *Suppose that during epoch τ of size $\mathcal{T}$ UCB-I is restarted and its environment was changed by ρ_s so that the upper confidence bound is changed to $\sqrt{\frac{4\rho_s \log(t)}{T_{i,j}(t)}} + \frac{4\rho_s \log(t)}{3T_{i,j}(t)}$ for arm j at time t. Then the expected regret of the algorithm is bounded by*

$$\mathbb{E}[R_i(\mathcal{T})] \leq \sqrt{8\rho_s k_i \mathcal{T} \log(\mathcal{T})}$$

Proof of Theorem 3.8.4. Let the reward of arm j at time t be $r_{t,j}$ and the rescaled reward be $\widehat{r}_{t,j}$. Without loss of generality assume that the arm with highest reward is $j = 1$. Denote the mean of arm j as μ_j and denote the mean of the best arm as

μ^*. During this run of UCB we know that each $|\widehat{r}_{t,j}| \leq \rho_s$. Further if we denote the probability with which the algorithm is sampled at time t as $p_{t,i}$ we have $\mathbb{E}[\widehat{r}_{t,j} - \mu_j | p_{1:t-1,i}] = 0$ and hence $r_{t,j} - \mu_j$ is a martingale difference. Further notice that the conditional second moment of $r_{t,j}$ is $\mathbb{E}[\widehat{r}^2_{t,j} | p_{1:t-1,i}] = \mathbb{E}[w_{t,i}\frac{r^2_{t,j}}{w^2_{t,i}} + 0 | p_{1:t-1,i}] \leq \rho$. Let $Y_t = (\widehat{r}_{\tau,j} - \mu_j)$. Bernstein's inequality for martingales (Bartlett et al. (2008)[Lemma 10]) now implies that $\mathbb{P}\left[\sum_{t=1}^{\mathcal{T}} Y_t > \sqrt{2\mathcal{T}\rho\log(1/\delta)} + \frac{2}{3}\rho\log(1/\delta)\right] \leq \delta$. This implies that the confidence bound should be changed to

$$\sqrt{\frac{4\rho_s \log(t)}{T_{i,j}(t)}} + \frac{4\rho_s \log(t)}{3T_{i,j}(t)}.$$

Following the standard proof of UCB we can now conclude that a suboptimal arm can be pulled at most $T_{i,j}(t)$ times up to time t where

$$2\Delta_j \geq \sqrt{\frac{4\rho_s \log(t)}{T_{i,j}(t)}} + \frac{4\rho_s \log(t)}{3T_{i,j}(t)}.$$

This implies that

$$\mathbb{E}\left[T_{i,j}(t)\right] \leq \frac{8\rho_s \log(t)}{\Delta_j^2}.$$

Next we bound the regret of the algorithm up to time t as follows:

$$\begin{aligned}
\mathbb{E}[R_i(t)] &\leq \sum_{j\neq 1} \Delta_j \mathbb{E}\left[T_{i,j}(t)\right] = \sum_{j \neq j^*} \sqrt{\mathbb{E}\left[T_{i,j}(t)\right]}\sqrt{\Delta_j^2 \mathbb{E}\left[T_{i,j}(t)\right]} \\
&\leq \sum_{j \neq j^*} \sqrt{\mathbb{E}\left[T_{i,j}(t)\right]}\sqrt{8\rho_s \log(t)} \leq k_i \sqrt{\frac{1}{k_i}\sum_j \mathbb{E}[T_{i,j}(t)]} = \sqrt{8\rho_s k_i t \log(t)}.
\end{aligned}$$

□

In general the argument can be repeated for other UCB-type algorithms (e.g. Successive Elimination) and hinges on the fact that the rescaled rewards $\widehat{r}_{t,j}$ have second moment bounded by ρ since with probability $p_{t,i}$ we have $\widehat{r}^2_{t,j} = \frac{r^2_{t,j}}{p^2_{t,i}}$ and with probability $1 - p_{t,i}$ it equals $\widehat{r}^2_{t,j} = 0$. We are not sure if similar arguments can be carried out for more delicate versions of UCB, like KL-UCB and leave it as future work to check.

3.9 Detailed proofs from Section 3.8

3.9.1 Proof of Theorem 3.8.1

3.9.1.1 Potential function and auxiliary lemmas

First we recall the definition of conjugate of a convex function f, denoted as f^*

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

In our algorithm, we are going to use the following potential at time t

$$\begin{aligned} \Psi_t(w) &= -4 \sum_{i=1}^{K} \frac{\sqrt{w_i} - \frac{1}{2} w_i}{\eta_{t,i}} \\ \nabla \Psi_t(w)_i &= -2 \frac{\frac{1}{\sqrt{w_i}} - 1}{\eta_{t,i}} \\ \nabla^2 \Psi_t(w)_{i,i} &= \frac{1}{w_i^{3/2} \eta_{t,i}}, \nabla^2 \Psi_t(w)_{i,j} = 0 \\ \nabla \Psi_t^*(Y)_i &= \frac{1}{\left(-\frac{\eta_{t,i}}{2} Y_i + 1\right)^2} \\ \Phi_t(Y) &= \max_{w \in \Delta^{K-1}} \langle Y, w \rangle - \Psi_t(w) = \left(\Psi_t + \mathrm{I}_{\Delta^{K-1}}\right)^* (Y). \end{aligned} \tag{3.9}$$

Further for a function f we use $D_f(x, y)$ to denote the Bregman divergence between x and y induced by f equal to

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle = f(x) + f^*(\nabla f(y)) - \langle \nabla f(y), x \rangle,$$

where the second inequality follows by the Fenchel duality equality $f^*(\nabla f(y)) + f(y) = \langle \nabla f(y), y \rangle$. We now present a couple of auxiliary lemmas useful for analyzing the OMD and FTRL updates.

Lemma 3.9.1. *For any $x, y \in \Delta^{K-1}$ it holds*

$$D_{\Psi_t}(x, y) = D_{\Phi_t}(\nabla \Phi_t^*(y), \nabla \Phi_t^*(x)).$$

Proof. Since $\Psi_t + \mathrm{I}_{\Delta^{K-1}}$ is a convex, closed function on Δ^{K-1} it holds that $\Psi_t + \mathrm{I}_{\Delta^{K-1}} = ((\Psi_t + \mathrm{I}_{\Delta^{K-1}})^*)^*$ (see for e.g. (Brezis, 2010) Theorem 1.11). Further, $\Phi_t^*(x) =$

$((\Psi_t + \mathrm{I}_{\Delta^{K-1}})^*)^*(x) = \Psi_t(x)$. The above implies

$$D_{\Psi_t}(x, y) = D_{\Phi_t^*}(x, y) = D_{\Phi_t}(\nabla\Phi_t^*(y), \nabla\Phi_t^*(x)).$$

□

Lemma 3.9.2. *For any positive $\widehat{L}_t$ and w_{t+1} generated according to update 3.7 we have*

$$w_{t+1} = \nabla\Phi_{t+1}(-\widehat{L}_t) = \nabla\Psi_{t+1}^*(-\widehat{L}_t + \nu_{t+1}\mathbf{1}),$$

for some scalar ν_t. Further $(\widehat{L}_t - \nu_{t+1}\mathbf{1})_i > 0$ for all $i \in [K]$.

Proof. The proof is contained in Section 4.3 in Zimmert and Seldin (2021). □

Lemma 3.9.3 (Lemma 16 Zimmert and Seldin (2021))**.** *Let $w \in \Delta^{K-1}$ and $\tilde{w} = \nabla\Psi_t^*(\nabla\Psi_t(w) - \ell)$. If $\eta_{t,i} \leq \frac{1}{4}$, then for all $\ell > -1$ it holds that $\tilde{w}_i^{3/2} \leq 2w_i^{3/2}$.*

3.9.1.2 Regret bound

We begin by studying the instantaneous regret of the FTRL update. The bound follows the one in Zimmert and Seldin (2021). Let $u = e_{i^*}$ be the unit vector corresponding to the optimal algorithm $\mathcal{A}_{i^*}$. To help with notation we will treat the algorithms as sampled from w_t instead of p_t. This is WLOG as the extra regret incurred due to the uniform exploration mixed in p_t is only $O(1)$. First we decompose the regret into a stability term and a penalty term:

$$\begin{aligned}\langle\widehat{\ell}_t, w_t - u\rangle = \langle\widehat{\ell}_t, w_t\rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1}) \text{ \textit{(Stability)}}\\ - \Phi_t(-\widehat{L}_t) + \Phi_t(-\widehat{L}_{t-1}) - \langle\widehat{\ell}_t, u\rangle \text{ \textit{(Penalty)}}.\end{aligned}$$

The bound on the stability term follows from Lemma 11 in Zimmert and Seldin (2021), however, we will show this carefully, since parts of the proof will be needed to bound other terms. Recall the definition of $\Phi_t(Y) = \max_{w\in\Delta^{K-1}}\langle Y, w\rangle - \Psi_t(w)$. Since w is in the simplex we have $\Phi_t(Y + \alpha\mathbf{1}_k) = \max_{w\in\Delta^{K-1}}\langle Y, w\rangle + \langle \alpha\mathbf{1}, w\rangle - \Psi_t(w) = \Phi_t(Y) + \alpha$. We also note that from Lemma 3.9.2 it follows that we can write $\nabla\Psi_t(w_t) = -\widehat{L}_{t-1} + \nu_t\mathbf{1}$. Combining the two facts we have

$$\begin{aligned}
&\langle \ell_t, w_t\rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1}) = \langle \ell_t, w_t\rangle + \Phi_t(\nabla\Psi_t(w_t) - \widehat{\ell}_t - \nu_t\mathbf{1}) - \Phi_t(\nabla\Psi_t(w_t) - \nu_t\mathbf{1})\\
&= \langle \ell_t - \alpha\mathbf{1}_k, w_t\rangle + \Phi_t(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k) - \Phi_t(\nabla\Psi_t(w_t))\\
&\leq \langle \ell_t - \alpha\mathbf{1}_k, w_t\rangle + \Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k) - \Psi_t^*(\nabla\Psi_t(w_t))\\
&= D_{\Psi_t^*}(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k, \nabla\Psi_t(w_t))\\
&\leq \max_{z\in[\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k, \nabla\Psi_t(w_t)]} \frac{1}{2}\|\widehat{\ell}_t - \alpha\mathbf{1}\|^2_{\nabla^2\Psi_t*(z)}\\
&= \max_{w\in[w_t, \nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]} \frac{1}{2}\|\widehat{\ell}_t - \alpha\mathbf{1}\|^2_{\nabla^2\Psi_t^{-1}(w)},
\end{aligned}$$

where the first inequality holds since $\Psi_t^* \geq \Phi_t$ and $\Psi_t^*(\nabla\Psi(w_t)) = \langle\nabla\Psi(w_t), w_t\rangle - \Psi_t(w_t) = \Phi_t(\nabla\Psi(w_t))$ and the second inequality follows since by Taylor's theorem there exists a z on the line segment between $\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k$ and $\Psi_t(w_t)$ such that $D_{\Psi_t^*}(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k, \nabla\Psi_t(w_t)) = \frac{1}{2}\|\widehat{\ell}_t - \alpha\mathbf{1}\|^2_{\nabla^2\Psi_t*(z)}$.

Lemma 3.9.4. *Let $w_t \in \Delta^{K-1}$ and let $i_t \sim w_t$. Let $\widehat{\ell}_{t,i_t} = \frac{\ell_{t,i_t}}{w_{t,i_t}}$ and $\widehat{\ell}_{t,i} = 0$ for all $i \neq i_t$. It holds that*

$$\mathbb{E}\left[\max_{w\in[w_t,\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]}\|\widehat{\ell}_t\|^2_{\nabla^2\Psi_t^{-1}(w)}\right] \leq \sum_{i=1}^{K}\frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]}$$

$$\mathbb{E}\left[\max_{w\in[w_t,\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]}\|\widehat{\ell}_t - \chi_{(i_t=j)}\ell_{t,j}\mathbf{1}\|^2_{\nabla^2\Psi_t^{-1}(w)}\right] \leq \sum_{i\neq j}\frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]} + \frac{\eta_{t,i}+\eta_{t,j}}{2}\mathbb{E}[w_{t,i}].$$

Proof. First notice that:

$$\begin{aligned}
&\mathbb{E}\left[\max_{w\in[w_t,\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]}\|\widehat{\ell}_t - \alpha\mathbf{1}_k\|^2_{\nabla^2\Psi_t^{-1}(w)}\right]\\
\leq&\mathbb{E}\left[\sum_{i=1}^{K}\max_{w_i\in[w_{t,i},\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)_i]}\frac{\eta_{t,i}}{2}w_i^{3/2}(\widehat{\ell}_{t,i} - \alpha)^2\right]
\end{aligned}$$

From the definition of $\nabla\Psi^*(Y)_i$ (Equation 3.9) we know that $\nabla\Psi^*(Y)_i$ is increasing on $(-\infty, 0]$ and hence for $\alpha = 0$ we have $w_{t,i} \geq \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t)_i$. This implies the maximum of each of the terms is attained at $w_i = w_{t,i}$. Thus

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \max_{w_i \in [w_{t,i}, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k)_i]} \frac{\eta_{t,i}}{2} w_i^{3/2} (\widehat{\ell}_{t,i})^2\right] \\
=&\mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_{t,i}}{2} w_{t,i}^{3/2} \chi_{(i_t=i)} \frac{\ell_{t,i}^2}{w_{t,i}^2}\right] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_{t,i}}{2} w_{t,i}^{3/2} \frac{\ell_{t,i}^2}{w_{t,i}}\right] \leq \sum_{i=1}^{K} \frac{\eta_{t,i}}{2} \sqrt{\mathbb{E}[w_{t,i}]}.
\end{aligned}
$$

When $\alpha = \chi_{(i_t=j)}\ell_{t,j}$ we consider several cases. First if $i_t \neq j$ the same bound as above holds. Next if $i_t = j$ for all $i \neq j$ we have $\widehat{\ell}_{t,i} - \alpha = -\alpha = -\ell_{t,j} \geq -1$ and for $\nabla\Phi_t^*(\nabla\Phi_t(w_t) - \widehat{\ell}_t + \ell_{t,j}) = \nabla\Phi_t^*(\nabla\Phi_t(w_t) + \ell_{t,j}) \leq 2^{2/3} w_{t,i}$ by Lemma 3.9.3. This implies that in this case the maximum in the terms is bounded by $2w_{t,i}^{3/2}\ell_{t,j}^2$. Finally if $i_t = j$ for the j-th term we again use the fact that $w_{t,j} \geq \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \ell_{t,j})_j$ since $-\widehat{\ell}_{t,j} + \ell_{t,j} \leq 0$. Combining all of the above we have

$$
\begin{aligned}
&\mathbb{E}\left[\max_{w \in [w_t, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k)]} \|\widehat{\ell}_t - \chi_{(i_t=j)}\ell_{t,j}\mathbf{1}\|^2_{\nabla^2\Psi_t^{-1}(w)}\right] \leq \sum_{i \neq j} \frac{\eta_{t,i}}{2} \sqrt{\mathbb{E}[w_{t,i}]} \\
+&\mathbb{E}\left[\chi_{(i_t=j)} \left(\frac{\eta_{t,j}}{2} \left(\frac{\ell_{t,j}}{w_{t,j}} - \ell_{t,j}\right)^2 w_{t,j}^{3/2} + \sum_{i \neq j} \ell_{t,j}^2 \frac{\eta_{t,i}}{2} w_{t,i}^{3/2}\right)\right] \\
=&\sum_{i \neq j} \frac{\eta_{t,i}}{2} \sqrt{\mathbb{E}[w_{t,i}]} + \mathbb{E}\left[\frac{\eta_{t,j}}{2} \left(\ell_{t,j}(1 - w_{t,j})\right)^2 w_{t,j}^{1/2} + \sum_{i \neq j} \ell_{t,j}^2 \frac{\eta_{t,i}}{2} w_{t,i}^{3/2} w_{t,j}\right] \\
\leq&\sum_{i \neq j} \frac{\eta_{t,i} + \eta_{t,j}}{2} \left(\sqrt{\mathbb{E}[w_{t,i}]} + \mathbb{E}[w_{t,i}]\right).
\end{aligned}
$$

□

Now the stability term is bounded by Lemma 3.9.4. Next we proceed to bound the penalty term in a slightly different way. Direct computation yields

$$
\begin{aligned}
D_{\Phi_t}(-\widehat{L}_{t-1}, \nabla\Phi_t^*(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi_t^*(u)) &= -\Phi_t(-\widehat{L}_t) + \Phi_t(-\widehat{L}_{t-1}) - \langle -\widehat{L}_{t-1} + \widehat{L}_t, u\rangle \\
&\quad + \Phi_t(\nabla\Phi_t^*(u)) - \Phi_t(\nabla\Phi_t^*(u)) \\
&= -\Phi_t(-\widehat{L}_t) + \Phi_t(-\widehat{L}_{t-1}) - \langle \widehat{\ell}_t, u\rangle.
\end{aligned} \tag{3.10}
$$

Using the next lemma and telescoping will result in a bound for the sum of the penalty terms

Lemma 3.9.5. *Let $u = e_{i^*}$ be the optimal algorithm. For any w_{t+1} such that $w_{t+1} = \nabla\Phi_{t+1}(-\widehat{L}_t)$ and $\eta_{t+1} \leq \eta_t$ it holds that*

$$D_{\Phi_{t+1}}(-\widehat{L}_t, \nabla\Phi^*_{t+1}(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi^*_t(u)) \leq 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}}\right)\left(\sqrt{w_{t+1,i}} - \frac{1}{2}w_{t+1,i}\right).$$

Proof.

$$\begin{aligned}
&D_{\Phi_{t+1}}(-\widehat{L}_t, \nabla\Phi^*_{t+1}(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi^*_t(u)) \\
&= \Phi_{t+1}(-\widehat{L}_t) - \Phi_t(-\widehat{L}_t) + \Phi^*_{t+1}(u) - \Phi^*_t(u) - \langle u, \widehat{L}_t - \widehat{L}_t\rangle \\
&= \Phi_{t+1}(-\widehat{L}_t) - \Phi_t(-\widehat{L}_t) + \Psi_{t+1}(u) - \Psi_t(u) \\
&= \Phi_{t+1}(-\widehat{L}_t) - \Phi_t(-\widehat{L}_t) - 2\left(\frac{1}{\eta_{t+1,i^*}} - \frac{1}{\eta_{t,i^*}}\right) \\
&= \langle w_{t+1}, -\widehat{L}_t\rangle - \Psi_{t+1}(w_{t+1}) - \Phi_t(-\widehat{L}_t) \\
&- 2\left(\frac{1}{\eta_{t+1,i^*}} - \frac{1}{\eta_{t,i^*}}\right) \\
&\leq \langle w_{t+1}, -\widehat{L}_t\rangle - \Psi_{t+1}(w_{t+1}) - \langle w_{t+1}, -\widehat{L}_t\rangle + \Psi_t(w_{t+1}) \\
&- 2\left(\frac{1}{\eta_{t+1,i^*}} - \frac{1}{\eta_{t,i^*}}\right) \\
&\leq 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}}\right)\left(\sqrt{w_{t+1,i}} - \frac{1}{2}w_{t+1,i}\right).
\end{aligned}$$

The first equality holds by Fenchel duality and the definition of Bregman divergence. The second equality holds by the fact that on the simplex $\Phi^*_t(\cdot) = \Psi_t(\cdot)$. The third equality holds because $\Psi_t(u) = -4(\sqrt{1} - \frac{1}{2})$. The fourth equality holds because w_{t+1} is the maximizer of $\langle -\widehat{L}_t, w\rangle + \Psi_{t+1}(w)$ and this is exactly how $\Phi_{t+1}(-\widehat{L}_t)$ is defined. The first inequality holds because

$$\begin{aligned}
-\Phi_t(-\widehat{L}_t) &= \max_{w\in\Delta^{K-1}} \langle -\widehat{L}_t, w\rangle + \Psi_t(w) \\
&\leq \langle -\widehat{L}_t, w_{t+1}\rangle + \Psi_t(w_{t+1}).
\end{aligned}$$

The final inequality holds because $\Psi_t(w_{t+1}) - \Psi_{t+1}(w_{t+1}) = 4\sum_i(1/\eta_{t+1,i} - 1/\eta_{t,i})(\sqrt{w_{t+1,i}} - w_{t+1,i}/2)$ and the fact that $\sqrt{w_{t+1,i^*}} - \frac{1}{2}w_{t+1,i^*} \leq \frac{1}{2}$. □

Next we focus on the OMD update. By the 3-point rule for Bregman divergence we can write

$$\begin{aligned}
\langle \widehat{\ell}_t, w_t - u\rangle &= \langle \nabla\Psi_t(w_t) - \nabla\Psi_t(\tilde{w}_{t+1}), w_t - u\rangle \\
&= D_{\Psi_t}(u, w_t) - D_{\Psi_t}(u, \tilde{w}_{t+1}) + D_{\Psi_t}(w_t, \tilde{w}_{t+1}) \\
&\leq D_{\Psi_t}(u, w_t) - D_{\Psi_t}(u, \widehat{w}_{t+1}) + D_{\Psi_t}(w_t, \tilde{w}_{t+1}), \\
\langle \widehat{\ell}_{t+1}, \widehat{w}_{t+1} - u\rangle &\leq D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_{t+1}}(u, \widehat{w}_{t+2}) + D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2}),
\end{aligned}$$

where the first inequality follows from the fact that $D_{\Psi_t}(u, \tilde{w}_{t+1}) \leq D_{\Psi_t}(u, \widehat{w}_{t+1})$ as $\widehat{w}_{t+1}$ is the projection of $\tilde{w}_{t+1}$ with respect to the Bregman divergence onto Δ^{K-1}.

We now explain how to control each of the terms. First we begin by matching $D_{\Psi_{t+1}}(u, \widehat{w}_{t+1})$ with $-D_{\Psi_t}(u, \widehat{w}_{t+1})$.

$$\begin{aligned}
D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_t}(u, \widehat{w}_{t+1}) &= \Psi_{t+1}(u) - \Psi_t(u) + \Psi_t(\widehat{w}_{t+1}) - \Psi_{t+1}(\widehat{w}_{t+1}) \\
&\quad + \langle \nabla\Psi_t(\widehat{w}_{t+1}), u - \widehat{w}_{t+1}\rangle - \langle \nabla\Psi_{t+1}(\widehat{w}_{t+1}), u - \widehat{w}_{t+1}\rangle \\
&= -2\left(\frac{1}{\eta_{t+1,i^*}} - \frac{1}{\eta_{t,i^*}}\right) \\
&\quad -4\sum_i \left(\sqrt{\widehat{w}_{t+1,i}} - \frac{1}{2}\widehat{w}_{t+1,i}\right)\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t+1,i}}\right) \\
&\quad -2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 1\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right) \\
&\quad +2\sum_i \widehat{w}_{t+1,i}\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i}}} - 1\right)\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t+1,i}}\right), \\
&= 2\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right) \\
&\quad -2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 1\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right) \\
&\quad -2\sum_i \sqrt{\widehat{w}_{t+1,i}}\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t+1,i}}\right),
\end{aligned}$$

where we have set $u = e_{i^*}$. Since the step size schedule is non-decreasing during OMD updates, we have that the above is bounded by

$$
\begin{aligned}
D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_t}(u, \widehat{w}_{t+1}) &\leq 2\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right) - 2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 1\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right) \\
&\leq -2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 2\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right).
\end{aligned} \tag{3.11}
$$

Next we explain how to control the terms $D_{\Psi_t}(w_t, \tilde{w}_{t+1})$ and $D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2})$. These can be thought of as the stability terms in the FTRL update.

Lemma 3.9.6. *For iterates generated by the OMD step in Equation 3.6 and any j it holds that*

$$
\begin{aligned}
\mathbb{E}[D_{\Psi_t}(w_t, \tilde{w}_{t+1})] &\leq \sum_{i=1}^{K} \frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]}, \\
\mathbb{E}[D_{\Psi_t}(w_t, \tilde{w}_{t+1})] &\leq \sum_{i\neq j} \frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]} + \frac{\eta_{t,i}+\eta_{t,j}}{2}\mathbb{E}[w_{t,i}], \\
\mathbb{E}[D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2})] &\leq \sum_{i=1}^{K} \frac{\eta_{t+1,i}}{2}\sqrt{\mathbb{E}[\widehat{w}_{t+1,i}]}, \\
\mathbb{E}[D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2})] &\leq \sum_{i\neq j} \frac{\eta_{t+1,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]} + \frac{\eta_{t+1,i}+\eta_{t+1,j}}{2}\mathbb{E}[w_{t,i}],
\end{aligned}
$$

where $\tilde{w}_{t+1}$ is any iterate such that $\widehat{w}_{t+1} = \operatorname{argmin}_{w\in\Delta^{K-1}} D_{\Psi_t}(w, \tilde{w}_{t+1})$.

Proof. We show the first two inequalities. The second couple of inequalities follow similarly. First we notice that we have

$$
\widehat{w}_{t+1} = \operatorname*{argmin}_{w\in\Delta^{K-1}} \langle w, \widehat{\ell}_t\rangle + D_{\Psi_t}(w, w_{t+1}) = \operatorname*{argmin}_{w\in\Delta^{K-1}} \langle w, \widehat{\ell}_t - \alpha\mathbf{1}_k\rangle + D_{\Psi_t}(w, w_{t+1}),
$$

for any α. This implies that $\widehat{w}_{t+1} = \operatorname{argmin}_{w\in\Delta^{K-1}} D_{\Psi_t}(w, \tilde{w}_{t+1})$ for $\tilde{w}_{t+1} = \operatorname{argmin}_{w\in\mathbb{R}^K} \langle w, \widehat{\ell}_t - \alpha\mathbf{1}_k\rangle + D_{\Psi_t}(w, w_{t+1})$. We can now write

$$
\begin{aligned}
D_{\Psi_t^*}(\nabla\Psi_t(\tilde{w}_{t+1}), \nabla\Psi_t(w_t)) &= D_{\Psi_t^*}(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k, \nabla\Psi_t(w_t)) \\
&\leq \max_{w\in[w_t, \nabla\Psi_t^*(\nabla\Psi_t(w_t) - \widehat{\ell}_t + \alpha\mathbf{1}_k)]} \|\widehat{\ell}_t - \alpha\mathbf{1}_k\|^2_{\nabla^2\Psi_t^{-1}(w)}.
\end{aligned}
$$

The proof is finished by Lemma 3.9.4. □

Finally we explain how to control $D_{\Psi_t}(u, w_t)$ and $D_{\Psi_{t+1}}(u, \widehat{w}_{t+2})$. First by Lemma 3.9.1 it holds that

$$D_{\Psi_t}(u, w_t) = D_{\Phi_t}(-L_{t-1}, \nabla\Phi_t^*(u)).$$

This term can now be combined with the term $-D_{\Phi_{t-1}}(-L_{t-1}, \nabla\Phi_{t-1}^*(u))$ coming from the prior FTRL update and both terms can be controlled through Lemma 3.9.5. To control $-D_{\Psi_{t+1}}(u, \widehat{w}_{t+2})$ we show that $-D_{\Psi_{t+1}}(u, \widehat{w}_{t+2}) = -D_{\Phi_{t+1}}(-\widehat{L}_{t+1}, \nabla\Phi_{t+1}^*(u))$. This is done by showing that if $\widehat{w}_{t+1}$ and $\widehat{w}_{t+2}$ are defined as in Equation 3.6 we can equivalently write $\widehat{w}_{t+2}$ as an FTRL step coming from a slightly different loss.

Lemma 3.9.7. *Let $\widehat{w}_{t+2}$ be defined as in Equation 3.6. Let ν_{t+1} be the constant such that $\nabla\Phi_{t+1}(-\widehat{L}_t) = \nabla\Psi_{t+1}^*(-\widehat{L}_t + \nu_t\mathbf{1}_k)$. Let $\widehat{L}_{t+1} = (\mathbf{1}_k - \mathrm{e}) \odot (\widehat{L}_t - (\nu_{t-1} + \nu_t)\mathbf{1}_k) + \frac{1}{\beta}\mathrm{e} \odot ((\widehat{L}_t - (\nu_{t-1} + \nu_t)\mathbf{1}_k)) + \widehat{\ell}_{t+1}$ and $\eta_{t+2} = \eta_{t+1}$. Then $(\widehat{L}_{t+1})_i \geq 0$ for all $i \in [K]$ and $\widehat{w}_{t+2} = w_{t+2} = \nabla\Phi_{t+2}(-\widehat{L}_{t+1})$.*

Proof. By the definition of the update we have

$$\begin{aligned}\widehat{w}_{t+1} &= \nabla\Phi_t(\nabla\Psi_t(w_t) - \widehat{\ell}_t) = \nabla\Phi_t(-\widehat{L}_t + \nu_{t-1}\mathbf{1}_k)\\ &= \nabla\Psi_t^*(-\widehat{L}_t + (\nu_{t-1} + \nu_t)\mathbf{1}_k),\\ \widehat{w}_{t+2} &= \nabla\Phi_{t+1}(\nabla\Psi_{t+1}(\widehat{w}_{t+1}) - \widehat{\ell}_{t+1}),\end{aligned}$$

where in the first equality we have used the fact that $\nabla\Psi_t(w_t) = -L_{t-1} + \nu_{t-1}\mathbf{1}_k$. For any i such that the OMD update increased the step size, i.e. $\eta_{t+1,i} = \beta\eta_{t,i}$ it holds from the definition of $\nabla\Psi_{t+1}(\cdot)$ that $\nabla\Psi_{t+1}(w)_i = \frac{1}{\beta}\nabla\Psi_t(w)_i$. Since $\nabla\Psi_t^*$ inverts $\nabla\Psi_t$ coordinate wise, we can write

$$\nabla\Psi_{t+1}(\widehat{w}_{t+1})_i = \frac{1}{\beta}\nabla\Psi_t(\widehat{w}_{t+1})_i = \frac{1}{\beta}(-\widehat{L}_t + (\nu_{t-1} + \nu_t)\mathbf{1}_k)_i.$$

If we let e be the the sum of all e_i's such that $\eta_{t+1,i} = \beta\eta_{t,i}$ we can write

$$\widehat{w}_{t+2} = \nabla\Phi_{t+1}\left((\mathbf{1}_k - e) \odot (-\widehat{L}_t + (\nu_{t-1} + \nu_t)\mathbf{1}_k) + \frac{1}{\beta}e \odot ((-\widehat{L}_t + (\nu_{t-1} + \nu_t)\mathbf{1}_k)) - \widehat{\ell}_{t+1}\right).$$

The fact that $\widehat{L}_{t+1,i} \geq 0$ for any i follows since any coordinate $\nabla\Psi_t(\widehat{w}_{t+1})_i \leq 0$ which implies that any coordinate of $(-\widehat{L}_t + (\nu_{t-1} + \nu_t)\mathbf{1}_k)_i \leq 0$. □

We can finally couple $-D_{\Phi_{t+1}}(-\widehat{L}_{t+1}, \nabla\Phi^*_{t+1}(u))$ with the term from the next FTRL step which is $D_{\Phi_{t+2}}(-\widehat{L}_{t+1}, \nabla\Phi^*_{t+2}(u))$ and use Lemma 3.9.5 to bound the sum of this two terms. Putting everything together we arrive at the following regret guarantee.

Theorem 3.9.8. *The regret bound for Algorithm 10 for any step size schedule which is non-increasing on the FTRL steps and any T_0 satisfies*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t - u\rangle\right] \leq{}& \sum_{t=T_0+1}^{T}\sum_{i\neq i^*}\mathbb{E}[\frac{3}{2}\eta_{t,i}\sqrt{w_{t,i}} + \frac{\eta_{t,i}+\eta_{t,i^*}}{2}w_{t,i}] + \sum_{t=1}^{T_0}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] \\ &+ \sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\ &+ \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{OMD}} 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right].\end{aligned}$$

Proof. Let $\mathcal{T}_{FTRL}$ be the set of all rounds in which the FTRL step is taken except for all rounds immediately before the OMD step and immediately after the OMD step. Let $\mathcal{T}_{OMD}$ be the set of all round immediately before the OMD step. The regret is bounded as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t - u\rangle\right] &= \sum_{t\in\mathcal{T}_{FTRL}}\mathbb{E}\left[\langle\widehat{\ell}_t, w_t - u\rangle\right] + \sum_{t\in[T]\setminus\mathcal{T}_{FTRL}}\mathbb{E}\left[\langle\widehat{\ell}_t, w_t - u\rangle\right] \\ &= \sum_{t\in[T]\setminus\mathcal{T}_{FTRL}}\mathbb{E}\left[\langle\widehat{\ell}_t, w_t - u\rangle\right] + \sum_{t\in\mathcal{T}_{FTRL}}\mathbb{E}\left[\langle\widehat{\ell}_t, w_t\rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1})\right. \\ &\quad \left. + D_{\Phi_t}(-\widehat{L}_{t-1}, \nabla\Phi^*_t(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi^*_t(u))\right].\end{aligned}$$

For any T_0, by the stability bound in Lemma 3.9.4 we have

$$\begin{aligned}\sum_{t\in\mathcal{T}_{FTRL}}\mathbb{E}\left[\langle\widehat{\ell}_t, w_t\rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1})\right] \leq{}& \sum_{t\in\mathcal{T}_{FTRL}\bigcap\{[T_0]\}}\sum_{i=1}^{K}\frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]} \\ &+ \sum_{t\in\mathcal{T}_{FTRL}\setminus\{[T_0]\}}\sum_{i\neq i^*}\mathbb{E}[\frac{\eta_{t,i}}{2}(\sqrt{w_{t,i}} + w_{t,i})].\end{aligned}$$

Next we consider the penalty term

$$\begin{aligned}&\sum_{t\in\mathcal{T}_{FTRL}}\mathbb{E}\left[D_{\Phi_t}(-\widehat{L}_{t-1},\nabla\Phi_t^*(u))-D_{\Phi_t}(-\widehat{L}_t,\nabla\Phi_t^*(u))\right]=\mathbb{E}\left[D_{\Phi_1}(0,\nabla\Phi_1^*(u))\right]\\&+\sum_{t+1\in\mathcal{T}_{FTRL}}\mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t,\nabla\Phi_t^*(u))-D_{\Phi_t}(-\widehat{L}_t,\nabla\Phi_t^*(u))\right]\\&-\mathbb{E}\left[\sum_{t\in\mathcal{T}_{OMD}}D_{\Phi_{t-1}}(-\widehat{L}_{t-1},\nabla\Phi_{t-1}^*(u))\right]\\&+\mathbb{E}\left[\sum_{t\in\mathcal{T}_{OMD}}D_{\Phi_{t+2}}(-\widehat{L}_{t+1},\nabla\Phi_{t+2}^*(u))\right]-\mathbb{E}\left[D_{\Phi_T}(-\widehat{L}_T,\nabla\Phi_T^*(u))\right].\end{aligned}$$

We are now going to complete the penalty term by considering the extra terms which do not bring negative regret from $\sum_{t\in[T]\setminus\mathcal{T}_{FTRL}}\mathbb{E}[\langle\widehat{\ell}_t,w_t-u\rangle]$.

$$\begin{aligned}&\sum_{t\in[T]\setminus\mathcal{T}_{FTRL}}\mathbb{E}[\langle\widehat{\ell}_t,w_t-u\rangle]\le\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Psi_t}(u,w_t)-D_{\Psi_t}(u,\widehat{w}_{t+1})+D_{\Psi_t}(w_t,\tilde{w}_{t+1})\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Psi_{t+1}}(u,\widehat{w}_{t+1})-D_{\Psi_{t+1}}(u,\widehat{w}_{t+2})+D_{\Psi_{t+1}}(\widehat{w}_{t+1},\tilde{w}_{t+2})\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[\langle\widehat{\ell}_{t+2},w_{t+2}\rangle+\Phi_{t+2}(-\widehat{L}_{t+2})-\Phi_{t+2}(-\widehat{L}_{t+1})\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1},\nabla\Phi_{t+2}^*(u))-D_{\Phi_{t+2}}(-\widehat{L}_{t+2},\nabla\Phi_{t+2}^*(u))\right]\\&=\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[\langle\widehat{\ell}_{t+2},w_{t+2}\rangle+\Phi_{t+2}(-\widehat{L}_{t+2})-\Phi_{t+2}(-\widehat{L}_{t+1})+D_{\Psi_t}(w_t,\tilde{w}_{t+1})+D_{\Psi_{t+1}}(\widehat{w}_{t+1},\tilde{w}_{t+2})\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Psi_{t+1}}(u,\widehat{w}_{t+1})-D_{\Psi_t}(u,\widehat{w}_{t+1})\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Phi_t}(-\widehat{L}_{t-1},\nabla\Phi_t^*(u))-D_{\Phi_{t+2}}(-\widehat{L}_{t+2},\nabla\Phi_{t+2}^*(u))\right]\\&+\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1},\nabla\Phi_{t+2}^*(u))-D_{\Psi_{t+1}}(u,\widehat{w}_{t+2})\right],\end{aligned}$$

where in the first inequality we have used the 3-point rule for Bregman divergence and the definition of the set τ_{FTRL}. For any T_0 the term

$$\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[\langle\widehat{\ell}_{t+2},w_{t+2}\rangle+\Phi_{t+2}(-\widehat{L}_{t+2})-\Phi_{t+2}(-\widehat{L}_{t+1})+D_{\Psi_t}(w_t,\tilde{w}_{t+1})+D_{\Psi_{t+1}}(\widehat{w}_{t+1},\tilde{w}_{t+2})\right]$$

is bounded by Lemma 3.9.4 and Lemma 3.9.6 as follows

$$\begin{aligned}&\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[\langle\widehat{\ell}_{t+2},w_{t+2}\rangle+\Phi_{t+2}(-\widehat{L}_{t+2})-\Phi_{t+2}(-\widehat{L}_{t+1})+D_{\Psi_t}(w_t,\tilde{w}_{t+1})+D_{\Psi_{t+1}}(\widehat{w}_{t+1},\tilde{w}_{t+2})\right]\\&\le\sum_{t\in\mathcal{T}_{OMD}\setminus\{[T_0]\}}\sum_{i\neq i^*}\mathbb{E}[\eta_{t+2,i}\sqrt{w_{t+2,i}}+\frac{\eta_{t+2,i}+\eta_{t+2,i^*}}{2}w_{t+2,i}]+\sum_{t\in\mathcal{T}_{OMD}\bigcap\{[T_0]\}}\sum_{i=1}^{K}\frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t+2,i}]},\end{aligned}$$

where we have used the $i \neq i^*$ bound from the above lemmas for all terms past T_0 and the bound which includes all $i \in [K]$ for the first T_0 terms. The term $\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Psi_{t+1}}(u,\widehat{w}_{t+1}) - D_{\Psi_t}(u,\widehat{w}_{t+1})\right]$ is bounded from Equation 3.11 as follows

$$\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Psi_{t+1}}(u,\widehat{w}_{t+1}) - D_{\Psi_t}(u,\widehat{w}_{t+1})\right] \leq \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}}-2\right)\left(\frac{1}{\eta_{t,i^*}}-\frac{1}{\eta_{t+1,i^*}}\right)\right].$$

By Lemma 3.8.2 and Lemma 3.9.1

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1},\nabla\Phi^*_{t+2}(u)) - D_{\Psi_{t+1}}(u,\widehat{w}_{t+2})\right] \\
&= \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1},\nabla\Phi^*_{t+2}(u)) - D_{\Phi_{t+1}}(-\widehat{L}_{t+1},\nabla\Phi^*_{t+1}(u))\right].
\end{aligned}$$

Combining all of the above we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t-u\rangle\right] \leq{}& \sum_{t=T_0+1}^{T}\sum_{i\neq i^*}\mathbb{E}[\frac{3}{2}\eta_{t,i}\sqrt{w_{t,i}}+\frac{\eta_{t,i}+\eta_{t,i^*}}{2}w_{t,i}] + \sum_{t=1}^{T_0}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] \\
&+ \sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}}-2\right)\left(\frac{1}{\eta_{t,i^*}}-\frac{1}{\eta_{t+1,i^*}}\right)\right] \\
&+ \sum_{t\in[T]\setminus\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t,\nabla\Phi^*_{t+1}(u)) - D_{\Phi_t}(-\widehat{L}_t,\nabla\Phi^*_t(u))\right] \\
&+ \mathbb{E}[D_{\Phi_1}(0,\nabla\Phi^*_1(u))] - \mathbb{E}[D_{\Phi_T}(-\widehat{L}_T,\nabla\Phi^*_T(u))].
\end{aligned} \tag{3.12}$$

Using Lemma 3.9.5 we have that

$$\begin{aligned}
&\sum_{t\in[T]\setminus\mathcal{T}_{OMD}}\mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t,\nabla\Phi^*_{t+1}(u)) - D_{\Phi_t}(-\widehat{L}_t,\nabla\Phi^*_t(u))\right] \leq \\
&\sum_{t\in[T]\setminus\mathcal{T}_{OMD}}\mathbb{E}\left[4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t+1,i}}-\frac{1}{\eta_{t,i}}\right)\left(\sqrt{w_{t+1,i}}-\frac{1}{2}w_{t+1,i}\right)\right].
\end{aligned}$$

By definition of w_1 we have $D_{\Phi_1}(0,\nabla\Phi^*_1(u)) = \Psi_1(u)-\Psi_1(w_1)$. Plugging back into Equation 3.12 we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t-u\rangle\right] \leq{}& \sum_{t=T_0+1}^{T}\sum_{i\neq i^*}\mathbb{E}[\frac{3}{2}\eta_{t,i}\sqrt{w_{t,i}}+\frac{\eta_{t,i}+\eta_{t,i^*}}{2}w_{t,i}] + \sum_{t=1}^{T_0}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] \\
&+ \sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}}-3\right)\left(\frac{1}{\eta_{t,i^*}}-\frac{1}{\eta_{t+1,i^*}}\right)\right] \\
&+ \mathbb{E}\left[\Psi_1(u)-\Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{OMD}} 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t,i}}-\frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right].
\end{aligned}$$

□

The algorithm begins by running each algorithm for $\log(T)+1$ rounds. We set the probability thresholds so that $\rho_1 = 36$, $\rho_j = 2\rho_{j-1}$ and $\frac{1}{\rho_n} \geq \frac{1}{KT}$, because we mix each w_t with the uniform distribution weighted by $1/KT$. This implies $n \leq \log_2(T)$. The algorithm now proceeds in epochs. The sizes of the epochs are as follows. The first epoch was of size $K\log(T) + K$, each epoch after doubles the size of the preceding one so that the number of epochs is bounded by $\log(T)$. In the beginning of each epoch, except for the first epoch we check if $w_{t,i} < \frac{1}{\rho_1}$. If it is we increase the step size $\eta_{t+1,i} = \beta\eta_{t,i}$ and run the OMD step. Let the τ-th epoch have size s_τ. Let $\frac{1}{\rho_\tau}$ be the largest threshold which was not exceeded during epoch τ. We require that each of the algorithms have the following expected regret bound under the unbiased rescaling of the losses $\bar{R}_i(t)$: $\mathbb{E}[\bar{R}_i(\sum_{\tau=1}^S s_\tau)] \leq \sum_{\tau=1}^S \mathbb{E}[\sqrt{\rho_\tau}R(s_\tau)]$. This can be ensured by restarting the algorithms in the beginning of the epochs if at the beginning of epoch τ it happens that $w_{t,i} > \frac{1}{\rho_{\tau-1}}$. Let ℓ_t be the loss over all possible actions. Let i_t be the algorithm selected by the corralling algorithm at time t. Let a^* be the best overall action.

Lemma 3.9.9. *Let $\bar{R}_{i^*}(\cdot)$ be a function upper bounding the expected regret of $\mathcal{A}_{i^*}$, $\mathbb{E}[R_{i^*}(\cdot)]$. For any η such that $\eta_{1,i} \leq \min_{t\in[T]} \frac{\left(1-\exp\left(-\frac{1}{\log(T)^2}\right)\right)\sqrt{t}}{50\bar{R}_i(t)}, \forall i \in [K]$ it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_{i_t,j_t}) - \ell_t(a^*)\right] \leq \sum_{t=T_0+1}^{T}\sum_{i\neq i^*} \mathbb{E}[\frac{3}{2}(\eta_{t,i}+\eta_{t,i^*})(\sqrt{w_{t,i}}+w_{t,i})] + \sum_{t=1}^{T_0}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right]$$

$$+\mathbb{E}\left[\Psi_1(u)-\Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{OMD}} 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t,i}}-\frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right] + 1 + 36\mathbb{E}[R_{i^*}(T)].$$

Proof. First we note that $\mathbb{E}[\widehat{\ell}_t(i^*)] = \mathbb{E}\left[w_{i^*,t}\frac{\ell_t(a_{i^*,j_t})}{w_{i^*,t}}\right] = \mathbb{E}[\ell_t(a_{i^*,j_t})]$. Using Theorem 3.9.8 we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\ell_t(a_{i_t,j_t}) - \ell_t(a^*)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(a_{i^*,j_t}) - \ell_t(a^*)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\langle \widehat{\ell}_t, \bar{w}_t - u \rangle\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\widehat{\ell}_t(i^*) - \ell_t(a^*)\right] + \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \mathbb{E}[\frac{3}{2}(\eta_{t,i} + \eta_{t,i^*})(\sqrt{w_{t,i}} + w_{t,i})] \\
&+ \sum_{t=1}^{T_0} \sum_{i=1}^{K} \mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] + \sum_{t \in \mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\
&+ \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t \in [T] \setminus \mathcal{T}_{OMD}} 4 \sum_{i \neq i^*} \left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right] + 1.
\end{aligned}
$$

Let us focus on $\sum_{t=1}^{T} \mathbb{E}\left[\widehat{\ell}_t(i^*) - \ell_t(a^*)\right] - 2\sum_{t \in \mathcal{T}_{OMD}} \mathbb{E}\left[\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right]$. By our assumption on $\mathcal{A}_{i^*}$ it holds that

$$
\sum_{t=1}^{T} \mathbb{E}\left[\widehat{\ell}_t(i^*) - \ell_t(a^*)\right] \leq \mathbb{E}\left[R_{i^*}(\sum_{\tau=1}^{\log(T)} s_\tau)\right] \leq \sum_{\tau=1}^{\log(T)} \mathbb{E}[\sqrt{\rho_\tau} R_{i^*}(s_\tau)].
$$

We now claim that during epoch τ there is a t in that epoch such that also $t \in \mathcal{T}_{OMD}$ and for which $w_{t,i^*} \leq \frac{1}{\rho_{\tau-1}}$. We consider two cases, first if OMD was invoked because at least one of the probability thresholds ρ_s was passed by a w_{t_s,i^*}, we must have $\rho_s \leq \rho_\tau$. Also by definition of ρ_τ as the largest threshold not passed by any w_{t,i^*} there exists at least one $t' \geq t_s$ for which $\frac{1}{\rho_{\tau-1}} \geq w_{t',i^*} > \frac{1}{\rho_\tau}$. This implies that we have subtracted at least $2\mathbb{E}\left[\left(\frac{1}{\sqrt{\widehat{w}_{t'+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t',i^*}} - \frac{1}{\eta_{t'+1,i^*}}\right)\right] \geq 2\mathbb{E}\left[\left(\sqrt{\rho_{\tau-1}} - 3\right)\left(\frac{1}{\eta_{t',i^*}} - \frac{1}{\eta_{t'+1,i^*}}\right)\right]$. In the second case we have that for all t in epoch τ it holds that $\frac{1}{\rho_{\tau-1}} \geq w_{t,i^*} > \frac{1}{\rho_\tau}$ or $w_{t,i^*} > \frac{1}{\rho_1}$. In the second case we only incur regret $\mathbb{E}[R_1(t)]$ scaled by 36 and in the first case the OMD played in the beginning of the epoch has resulted in at least $-2\mathbb{E}\left[\left(\sqrt{\rho_{\tau-1}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right]$ negative contribution, where t indexes the beginning of the epoch. We set $\beta = e^{1/\log(T)^2}$ and now evaluate the difference $\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}} \geq \left(1 - \frac{1}{\beta}\right)\frac{\sqrt{t}}{25\eta_{1,i^*}}$. Where we have used the fact that $\eta_{t,i^*} \leq \frac{\eta_{1,i^*}\beta^{\log_2(T)^2}}{\sqrt{t}} \leq \frac{25\eta_{1,i^*}}{\sqrt{t}}$. This follows by noting that there are $\log_2(T)$ epochs and during each epoch one can call the OMD step only $\log_2(T)$ times. Let $\beta' = \left(1 - \frac{1}{\beta}\right)$. Thus if t_τ is the beginning of epoch τ we subtract at least $\frac{\beta'\sqrt{t_\tau \rho_{\tau-1}}}{25\eta_{1,i^*}}$. Notice that the length of each epoch s_τ does not exceed $2t_\tau$, thus we have

$$\sum_{\tau=1}^{\log(T)} \mathbb{E}[\sqrt{\rho_\tau} R_{i^*}(s_\tau)] \leq \sum_{\tau=1}^{\log(T)} \mathbb{E}[\sqrt{\rho_\tau} R_{i^*}(2t_\tau)],$$

and so as long as we set $\eta_{1,i^*} \leq \frac{\beta' \sqrt{2t_\tau}}{50\bar{R}_{i^*}(2t_\tau)}$, where $\mathbb{E}[R_{i^*}(2t_\tau)] \leq \bar{R}_{i^*}(2t_\tau)$ we have

$$\begin{aligned} \sum_{\tau=1}^{\log(T)} \mathbb{E}[\sqrt{\rho_\tau} R_{i^*}(2t_\tau)] - 2 \sum_{t \in \mathcal{T}_{OMD}} \mathbb{E}\left[\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\ \leq \sum_{\tau=1}^{\log(T)} \mathbb{E}[\sqrt{\rho_\tau} R_{i^*}(2t_\tau) - \sqrt{\rho_\tau} R_{i^*}(2t_\tau)] \leq 0. \end{aligned}$$

□

We can now use the self-bounding trick of the regret as in Zimmert and Seldin (2021) to finish the proof. Let μ^* denote the reward of the best arm. First note that we can write

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_{i_t,j_t}) - \ell_t(a^*)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \chi_{i_t \neq i^*}(\ell_t(a_{i_t,j_t}) - \mu^*)\right] + \mathbb{E}\left[R_{i^*}(T_{i^*}(T))\right] \\ &\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}\chi_{i_t \neq i^*}(\ell_t(a_{i_t,j_t}) - \mu^*)\right] \\ &\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i \neq i^*} w_{t,i}\Delta_i\right]. \end{aligned}$$

Theorem 3.9.10 (Theorem 3.8.1)**.** *Let $\bar{R}_{i^*}(\cdot)$ be a function upper bounding the expected regret of $\mathcal{A}_{i^*}$, $\mathbb{E}[R_{i^*}(\cdot)]$. For any η such that $\eta_{1,i} \leq \min_{t \in [T]} \frac{\left(1 - \exp\left(-\frac{1}{\log(T)^2}\right)\right)\sqrt{t}}{50\bar{R}_i(t)}, \forall i \in [K]$ and $\beta = e^{1/\log(T)^2}$ it holds that the expected regret of Algorithm 10 is bounded as*

$$\begin{aligned} \mathbb{E}\left[R(T)\right] \leq \sum_{i \neq i^*} \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{\Delta_i}\left(\log\left(\frac{T\Delta_i - 15\eta_{1,i}}{T_0\Delta_i - 15\eta_{1,i}}\right) + \log\left(225\eta_{1,i}^2\Delta_i/\Delta_1\right)\right) \\ + \sum_{i \in [K]} \frac{8}{\eta_{1,i}\sqrt{K}} + 2 + 72R_{i^*}(T), \end{aligned}$$

where $T_0 = \max_{i \neq i^*} \frac{225\eta_{1,i}^2}{\Delta_i}$.

Proof of Theorem 3.8.1. By Lemma 3.9.9 we have that the overall regret is bounded by

$$\begin{aligned}
\mathbb{E}[R(T)] \quad \leq \quad & \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \mathbb{E}[\frac{3}{2}(\eta_{t,i} + \eta_{t,i^*})(\sqrt{w_{t,i}} + w_{t,i})] + \sum_{t=1}^{T_0} \sum_{i=1}^{K} \mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] \\
& + \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t \in [T] \setminus \mathcal{T}_{OMD}} 4 \sum_{i \neq i^*} \left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right] \\
& + 1 + 36\mathbb{E}[R_{i^*}(T)] \\
\leq \quad & \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \mathbb{E}[\frac{75\eta_{1,i}}{2\sqrt{t}}(\sqrt{w_{t,i}} + w_{t,i})] + \sum_{t=1}^{T_0} \sum_{i=1}^{K} \mathbb{E}\left[\frac{25\eta_{1,i}}{2\sqrt{t}}\sqrt{w_{t,i}}\right] \\
& + \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \frac{10}{\eta_{1,i}\sqrt{t}}\left(\sqrt{w_{t,i}}\right)\right] \\
& + 1 + 36\mathbb{E}[R_{i^*}(T)] \\
\leq \quad & \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \mathbb{E}[\frac{75\eta_{1,i}}{2\sqrt{t}}(\sqrt{w_{t,i}} + w_{t,i})] + \sum_{t=1}^{T_0} \sum_{i=1}^{K} \mathbb{E}\left[\frac{25\eta_{1,i}}{2\sqrt{t}}\sqrt{w_{t,i}}\right] \\
& + \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \frac{10}{\eta_{1,i}\sqrt{t}}\left(\sqrt{w_{t,i}}\right)\right] \\
& + 1 + 36\mathbb{E}[R_{i^*}(T)] + \mathbb{E}[R(T)] - \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} w_{t,i}\Delta_i\right] \\
\leq \quad & \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \mathbb{E}[\frac{75\eta_{1,i}}{\sqrt{t}}(\sqrt{w_{t,i}} + w_{t,i})] + \sum_{t=1}^{T_0} \sum_{i=1}^{K} \mathbb{E}\left[\frac{25\eta_{1,i}}{\sqrt{t}}\sqrt{w_{t,i}}\right] \\
& + 2\mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \frac{20}{\eta_{1,i}\sqrt{t}}\left(\sqrt{w_{t,i}}\right)\right] \\
& + 2 + 3672\mathbb{E}[R_{i^*}(T)] - \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} w_{t,i}\Delta_i\right].
\end{aligned}$$

In the first inequality we used the fact that for any i we have $\eta_{t,i} \leq 25\eta_{1,i}/\sqrt{t}$, in the second inequality we have used the self bounding property derived before the statement of the theorem and in the third inequality we again used the bound on the expected regret $\mathbb{E}[R(T)]$ from the first inequality. We are now going to use the fact that for any $w > 0$ it holds that $2\alpha\sqrt{w} - \beta w \leq \frac{\alpha^2}{\beta}$. For $t \leq T_0$ we have

$$\sum_{t=1}^{T_0} \sum_{i \neq i^*} \left(20\frac{\sqrt{w_{t,i}}}{\sqrt{t}}\left(\frac{1}{\eta_{1,i}} + \eta_{1,i}\right) - \Delta_i w_{t,i}\right) \leq \sum_{t=1}^{T_0} \sum_{i \neq i^*} \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{t\Delta_i}.$$

For $t > T_0$ we have

$$\sum_{T_0+1}^{T} \sum_{i \neq i^*} \left(\frac{\sqrt{w_{t,i}}}{\sqrt{t}} \left(\frac{20}{\eta_{1,i}} + 75\eta_{1,i} \right) - \left(\Delta_i - \frac{15\eta_{1,i}}{\sqrt{t}} \right) w_{t,i} \right) \leq \sum_{t=T_0+1}^{T} \sum_{i \neq i^*} \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{t\Delta_i - 15\eta_{1,i}\sqrt{t}}$$
$$\leq \sum_{i \neq i^*} \int_{T_0}^{T} \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{t\Delta_i - 15\eta_{1,i}\sqrt{t}} dt$$
$$= \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{\Delta_i} \log \left(\frac{15\eta_{1,i} - T\Delta_i}{15\eta_{1,i} - T_0\Delta_i} \right).$$

We now choose $T_0 = \max_{i \neq i^*} \frac{225\eta_{1,i}^2}{\Delta_i}$. To bound $\mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right]$ we have set w_1 to be the uniform distribution over the K algorithms and recall that $\Psi_1(w) = -4\sum_i \frac{\sqrt{w_i} - \frac{1}{2}w_i}{\eta_{1,i}}$. This implies $\Psi_1(u) - \Psi_1(w_1) \leq \sum_{i \in [K]} \frac{4}{\eta_{1,i}\sqrt{K}}$. Putting everything together we have

$$\mathbb{E}\left[R(T)\right] \leq \sum_{i \neq i^*} \frac{1500(1/\eta_{1,i} + \eta_{1,i})^2}{\Delta_i} \left(\log \left(\frac{T\Delta_i - 15\eta_{1,i}}{T_0\Delta_i - 15\eta_{1,i}} \right) + \log \left(225\eta_{1,i}^2/\Delta_i \right) \right)$$
$$+ \sum_{i \in [K]} \frac{8}{\eta_{1,i}\sqrt{K}} + 2 + 72R_{i^*}(T)$$

□

3.9.2 Proof of Theorem 3.8.3

We now consider the setting in which the best overall arm does not maintain a gap at every round. Following the proof of Theorem 3.9.8 we are able to show the following.

Theorem 3.9.11. *The regret bound for Algorithm 10 for any step size schedule which is non-increasing on the FTRL steps satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \widehat{\ell}_t, w_t - u \rangle \right] \leq 4 \max_{w \in \Delta^{K-1}} \sqrt{T} \sum_{i=1}^{K} \left(\eta_{1,i} + \frac{1}{\eta_{1,i}} \right) \sqrt{w_i}$$
$$+ \sum_{t \in \mathcal{T}_{OMD}} \mathbb{E}\left[-2 \left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3 \right) \left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}} \right) \right].$$

Proof. From the proof of Theorem 3.9.8 we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \widehat{\ell}_t, w_t - u \rangle \right] = \sum_{t \in [T] \setminus \mathcal{T}_{FTRL}} \mathbb{E}\left[\langle \widehat{\ell}_t, w_t - u \rangle \right] + \sum_{t \in \mathcal{T}_{FTRL}} \mathbb{E}\Big[\langle \widehat{\ell}_t, w_t \rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1})$$
$$+ D_{\Phi_t}(-\widehat{L}_{t-1}, \nabla\Phi_t^*(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi_t^*(u)) \Big].$$

Lemma 3.9.4 implies

$$\sum_{t\in\mathcal{T}_{FTRL}} \mathbb{E}\left[\langle \widehat{\ell}_t, w_t\rangle + \Phi_t(-\widehat{L}_t) - \Phi_t(-\widehat{L}_{t-1})\right] \le \sum_{t\in\mathcal{T}_{FTRL}} \sum_{i=1}^{K} \frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t,i}]}.$$

As before the penalty term is decomposed as follows

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{FTRL}} \mathbb{E}\left[D_{\Phi_t}(-\widehat{L}_{t-1}, \nabla\Phi_t^*(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi_t^*(u))\right] = \mathbb{E}\left[D_{\Phi_1}(0, \nabla\Phi_1^*(u))\right]\\
&+ \sum_{t+1\in\mathcal{T}_{FTRL}} \mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t, \nabla\Phi_t^*(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi_t^*(u))\right]\\
&-\mathbb{E}\left[\sum_{t\in\mathcal{T}_{OMD}} D_{\Phi_{t-1}}(-\widehat{L}_{t-1}, \nabla\Phi_{t-1}^*(u))\right]\\
&+\mathbb{E}\left[\sum_{t\in\mathcal{T}_{OMD}} D_{\Phi_{t+2}}(-\widehat{L}_{t+1}, \nabla\Phi_{t+2}^*(u))\right] - \mathbb{E}\left[D_{\Phi_T}(-\widehat{L}_T, \nabla\Phi_T^*(u))\right].
\end{aligned}$$

Next the term $\sum_{t\in[T]\setminus\mathcal{T}_{FTRL}} \mathbb{E}[\langle \widehat{\ell}_t, w_t - u\rangle]$ is again decomposed as in the proof of Theorem 3.9.8

$$\begin{aligned}
&\sum_{t\in[T]\setminus\mathcal{T}_{FTRL}} \mathbb{E}[\langle \widehat{\ell}_t, w_t - u\rangle]\\
&\le \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[\langle \widehat{\ell}_{t+2}, w_{t+2}\rangle + \Phi_{t+2}(-\widehat{L}_{t+2}) - \Phi_{t+2}(-\widehat{L}_{t+1}) + D_{\Psi_t}(w_t, \tilde{w}_{t+1}) + D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2})\right]\\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_t}(u, \widehat{w}_{t+1})\right]\\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_t}(-\widehat{L}_{t-1}, \nabla\Phi_t^*(u)) - D_{\Phi_{t+2}}(-\widehat{L}_{t+2}, \nabla\Phi_{t+2}^*(u))\right]\\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1}, \nabla\Phi_{t+2}^*(u)) - D_{\Psi_{t+1}}(u, \widehat{w}_{t+2})\right].
\end{aligned}$$

Using Lemma 3.9.4 and Lemma 3.9.6 we bound the first term of the above inequality as

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[\langle \widehat{\ell}_{t+2}, w_{t+2}\rangle + \Phi_{t+2}(-\widehat{L}_{t+2}) - \Phi_{t+2}(-\widehat{L}_{t+1}) + D_{\Psi_t}(w_t, \tilde{w}_{t+1}) + D_{\Psi_{t+1}}(\widehat{w}_{t+1}, \tilde{w}_{t+2})\right]\\
&\le \sum_{t\in\mathcal{T}_{OMD}} \sum_{i=1}^{K} \frac{\eta_{t,i}}{2}\sqrt{\mathbb{E}[w_{t+2,i}]}
\end{aligned}$$

The term $\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_t}(u, \widehat{w}_{t+1})\right]$ is bounded from Equation 3.11 as follows

$$\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\Psi_t}(u, \widehat{w}_{t+1})\right] \le \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 2\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right].$$

By Lemma 3.8.2 and Lemma 3.9.1

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1}, \nabla\Phi^*_{t+2}) - D_{\Psi_{t+1}}(u, \widehat{w}_{t+2})\right] \\
=&\sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+2}}(-\widehat{L}_{t+1}, \nabla\Phi^*_{t+2}) - D_{\Phi_{t+1}}(-\widehat{L}_{t+1}, \nabla\Phi^*_{t+2})\right].
\end{aligned}$$

Combining all of the above we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t - u\rangle\right] \leq \sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] + \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 2\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\
+ \sum_{t\in[T]\setminus\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t, \nabla\Phi^*_t(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi^*_t(u))\right] \\
+ \mathbb{E}[D_{\Phi_1}(0, \nabla\Phi^*_1(u))] - \mathbb{E}[D_{\Phi_T}(-\widehat{L}_T, \nabla\Phi^*_T(u))].
\end{aligned} \tag{3.13}$$

The last two terms are bounded in the same way as in the proof of Theorem 3.9.8

$$\begin{aligned}
&\sum_{t\in[T]\setminus\mathcal{T}_{OMD}} \mathbb{E}\left[D_{\Phi_{t+1}}(-\widehat{L}_t, \nabla\Phi^*_t(u)) - D_{\Phi_t}(-\widehat{L}_t, \nabla\Phi^*_t(u))\right] \\
&\qquad + \mathbb{E}[D_{\Phi_1}(0, \nabla\Phi^*_1(u))] - \mathbb{E}[D_{\Phi_T}(-\widehat{L}_T, \nabla\Phi^*_T(u))] \\
\leq&\mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] + \mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{OMD}} 4\sum_{i\neq i^*}\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right]
\end{aligned}$$

Plugging back into Equation 3.13 we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t - u\rangle\right] \leq& \sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\eta_{t,i}}{2}\sqrt{w_{t,i}}\right] + \mathbb{E}\left[\Psi_1(u) - \Psi_1(w_1)\right] \\
&+ 4\mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{T}_{OMD}}\sum_{i=1}^{K}\left(\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}\right)\left(\sqrt{w_{t,i}}\right)\right] \\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\
\leq& \sum_{t=1}^{T} 4\sum_{i=1}^{K}\left(\eta_{1,i} + \frac{1}{\eta_{1,i}}\right)\sqrt{\frac{w_{t,i}}{t}} \\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right] \\
\leq& 4\max_{w\in\Delta^{K-1}}\sqrt{T}\sum_{i=1}^{K}\left(\eta_{1,i} + \frac{1}{\eta_{1,i}}\right)\sqrt{w_i} \\
&+ \sum_{t\in\mathcal{T}_{OMD}} \mathbb{E}\left[-2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}} - 3\right)\left(\frac{1}{\eta_{t,i^*}} - \frac{1}{\eta_{t+1,i^*}}\right)\right],
\end{aligned}$$

where the last inequality follows from the fact that the maximizer of the function $\sum_{i=1}^{K}\sqrt{\frac{w_i}{t}}\alpha_i$ over the simplex, for $\alpha_i \geq 0$ is the same for all $t \in [T]$. $\square$

Following the proof of Lemma 3.9.9 and replacing the bound on $\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{\ell}_t, w_t - u\rangle\right]$ from Theorem 3.9.8 with the one from Theorem 3.9.11 finishes the proof.

3.10 Model selection with Tsallis-Inf and proof of Theorem 3.2.1

Recall the model selection problem for linear bandits from Section 3.2. We assume that there are K base learners $\{\mathcal{A}_i\}_{i=1}^{K}$ such that the regret of $\mathcal{A}_i$, for $i \geq i^*$, is bounded by $\tilde{O}(d_i^{\alpha}\sqrt{T})$. That is, whenever the model is correctly specified, the i-th algorithm admits a meaningful regret guarantee. In the setting of Foster et al. (2019), $\mathcal{A}_i$ can be instantiated as LinUCB and in that case $\alpha = 1/2$. Further, in the setting of infinite arms, $\mathcal{A}_i$ can be instantiated as OFUL (Abbasi-Yadkori et al., 2011), in which case $\alpha = 1$. Both $\alpha = 1/2$ and $\alpha = 1$ govern the min-max optimal rates in the respective settings. Our algorithm is now a simple modification of Algorithm 10. At every time-step t, we update $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t + \mathbf{d}$, where $\mathbf{d}_i = \frac{d_i^{2\alpha}}{\sqrt{T}}$. Intuitively, our modification creates a gap between the losses of $\mathcal{A}_{i^*}$ and any $\mathcal{A}_i$ for $i > i^*$ of the order $d_i^{2\alpha}$. On the other hand for any $i < i^*$, perturbing the loss can result in at most additional $d_{i^*}^{2\alpha}\sqrt{T}$ regret. With the above observations, the bound guaranteed by Theorem 3.8.1 implies that the modified algorithm should incur at most $\tilde{O}(d_{i^*}^{2\alpha}\sqrt{T})$ regret. We arrive at the statement of Theorem 3.2.1.

Theorem 3.10.1 (Theorem 3.2.1)**.** *Assume that every base learner $\mathcal{A}_i$, $i \geq i^*$, admits a $\tilde{O}(d_i^{\alpha}\sqrt{T})$ regret. Then, there exists a corralling strategy with expected regret bounded by $\tilde{O}(d_{i^*}^{2\alpha}\sqrt{T} + K\sqrt{T})$. Moreover, under the additional assumption that the following holds for any $i < i^*$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$*

$$\mathbb{E}[\langle\beta_i, \phi_i(x, a)\rangle] - \min_{a\in\mathcal{A}}\mathbb{E}[\langle\beta^*, \phi_{i^*}(x, a)\rangle] \geq 2\frac{d_{i^*}^{2\alpha} - d_i^{2\alpha}}{\sqrt{T}},$$

the expected regret of the same strategy is bounded as $\tilde{O}(d_{i^*}^{\alpha}\sqrt{T} + K\sqrt{T})$.

Typically, we have $K = O(\log(T))$ and thus Theorem 3.2.1 guarantees a regret of at most $\tilde{O}(d_{i^*}^{2\alpha}\sqrt{T})$. Furthermore, under a gap-assumption, which implies that the value of the smallest loss for the optimal embedding i^* is sufficiently smaller compared to the value of any sub-optimal embedding $i < i^*$, we can actually achieve a corralling regret of the order $R_{i^*}(T)$. In particular, for the setting of Foster et al. (2019), our strategy yields the desired $\tilde{O}(\sqrt{d_{i^*}T})$ regret bound. Notice that the regret guarantees are only meaningful as long as $d_{i^*} = o(T^{1/(2\alpha)})$. In such a case, the second assumption on the gap is that the gap is lower bounded by $o(1)$. This is a completely problem-dependent assumption and in general we expect that it cannot be satisfied.

3.10.1 Proof of Theorem 3.2.1

Since the losses might not be bounded in $[0,1]$ as $d_K = \Theta(T)$ we need to slightly modify the bound for the Stability term in Lemma 3.9.4 and the term $D_{\Psi_t}(w_t, \tilde{w}_{t+1})$ in Lemma 3.9.6. Recall that we need to bound the term $\mathbb{E}\left[\max_{w\in[w_t,\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]}\|\widehat{\ell}_t\|^2_{\nabla^2\Psi_t^{-1}(w)}\right]$. The argument is the same as in 3.9.4 up to

$$\mathbb{E}\left[\max_{w\in[w_t,\nabla\Psi_t^*(\nabla\Psi_t(w_t)-\widehat{\ell}_t+\alpha\mathbf{1}_k)]}\|\widehat{\ell}_t\|^2_{\nabla^2\Psi_t^{-1}(w)}\right] \leq \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_{t,i}}{2}w_{t,i}^{3/2}(\widehat{\ell}_{t,i})^2\right].$$

Let $\ell_{t,i} = \langle\beta_{i_t}, \phi_{i_t}(x_t, a_{i_t,j_t}\rangle + \xi_t)$, then we have

$$\mathbb{E}\left[\frac{\eta_{t,i}}{2}w_{t,i}^{3/2}(\widehat{\ell}_{t,i})^2\right] \leq \mathbb{E}\left[\eta_{t,i}\chi_{(i_t=i)}w_{t,i}^{3/2}\frac{\ell_{t,i}^2}{w_{t,i}^2} + \eta_{t,i}w_{t,i}^{3/2}\frac{d_i^{4\alpha}}{T}\right] \leq \mathbb{E}\left[\eta_{t,i}w_{t,i}\frac{d_i^{4\alpha}}{T}\right] + 2\mathbb{E}\left[\eta_{t,i}\sqrt{w_{t,i}}\right],$$

where in the last inequality we have used the fact that $w_{t,i} \geq w_{t,i}^{3/2}$ together with the our assumption that ξ_t is zero-mean with variance proxy 1. Following the proof of

Lemma 3.9.9 with the bound on the stability term we can bound

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_{i_t,j_t})-\ell_t(a^*)\right]=\sum_{t=1}^{T}\mathbb{E}\left[\ell_t(a_{i^*,j_t})-\ell_t(a^*)\right]+\sum_{t=1}^{T}\mathbb{E}\left[\langle\widehat{\ell}_t+\mathbf{d},w_t-u\rangle\right]\\
&-\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{d},w_t-u\rangle\right]+1\\
&\leq\sum_{t=1}^{T}\mathbb{E}\left[\widehat{\ell}_t(i^*)-\ell_t(a^*)\right]+2\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\left[\eta_{t,i}\sqrt{w_{t,i}}+\eta_{t,i}w_{t,i}\frac{d_i^{4\alpha}}{T}\right]\\
&+4\sum_{t=1}^{T}\sum_{i=1}^{K}\left(\frac{1}{\eta_{t,i}}-\frac{1}{\eta_{t-1,i}}\right)\sqrt{w_{t,i}}\\
&-\sum_{t=1}^{T}\mathbb{E}[\langle\mathbf{d},w_t-u\rangle]-\sum_{t\in\mathcal{T}_{OMD}}\mathbb{E}\left[2\left(\frac{1}{\sqrt{\widehat{w}_{t+1,i^*}}}-3\right)\left(\frac{1}{\eta_{t,i^*}}-\frac{1}{\eta_{t+1,i^*}}\right)\right]+\sqrt{K}+1\\
&\leq 4\sum_{t=1}^{T}\sum_{i=1}^{K}\sqrt{\frac{w_{t,i}}{t}}\left(\eta_{1,i}+\frac{1}{\eta_{t,i}}\right)+2\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\left[\eta_{t,i}w_{t,i}\frac{d_i^{4\alpha}}{T}\right]\\
&-\sum_{t=1}^{T}\mathbb{E}[\langle\mathbf{d},w_t-u\rangle]+36\mathbb{E}[R_{i^*}(T)].
\end{aligned}
$$

For a fixed t we have

$$
-\langle\mathbf{d},w_t-u\rangle=\frac{d_{i^*}^{2\alpha}}{\sqrt{T}}(1-w_{t,i^*})-\sum_{i\neq i^*}w_{t,i}\frac{d_i^{2\alpha}}{\sqrt{T}}=\sum_{i<i^*}w_{t,i}\frac{d_{i^*}^{2\alpha}-d_i^{2\alpha}}{\sqrt{T}}-\sum_{i>i^*}w_{t,i}\frac{d_i^{2\alpha}-d_{i^*}^{2\alpha}}{\sqrt{T}}.
$$

First we consider the terms $i>i^*$. Assume WLOG that $d_K^{2\alpha}\leq T/4$, as otherwise the learning guarantees are trivial. For these terms we have

$$
\sqrt{\frac{w_{t,i}}{t}}\frac{1}{\eta_{1,i}}+w_{t,i}\left(\frac{\eta_{1,i}}{\sqrt{t}}\frac{d_i^{4\alpha}}{T}-\frac{d_i^{2\alpha}}{\sqrt{T}}\right)\leq\sqrt{\frac{w_{t,i}}{t}}\frac{1}{\eta_{1,i}}-w_{t,i}\frac{d_i^{2\alpha}}{2\sqrt{T}}\leq\frac{\sqrt{T}}{td_i^{2\alpha}\eta_{1,i}^2}.
$$

Since $\eta_{1,i}=\tilde{\Theta}(1/d_i^{\alpha})$ we have that the above is further bounded by $\tilde{O}(\sqrt{T}/t)$.

Next we consider the terms for $i<i^*$ given by $w_{t,i}\frac{d_{i^*}^{2\alpha}-d_i^{2\alpha}}{\sqrt{T}}$. Here we use our assumption that the regret $\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_{i_t,j_t})-\ell_t(a^*)\right]\geq\mathbb{E}\left[w_{t,i}\Delta_i\right]$, where $\Delta_i=\mathbb{E}[\langle\beta_i,\phi_i(x,a)\rangle]-\min_{a\in\mathcal{A}}\mathbb{E}[\langle\beta^*,\phi_{i^*}(x,a)\rangle]$. Using the self-bounding trick we can cancel out the terms $w_{t,i}\frac{d_{i^*}^{2\alpha}-d_i^{2\alpha}}{\sqrt{T}}$ as soon as $\Delta_i\geq 2w_{t,i}\frac{d_{i^*}^{2\alpha}-d_i^{2\alpha}}{\sqrt{T}}$, which holds by the gap assumption in the theorem. All other terms in the regret bound are bounded by $\tilde{O}(d_{i^*}^{\alpha}\sqrt{T})$. Thus we have shown that the regret of the corralling algorithm is bounded as

$$
\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_{i_t,j_t})-\ell_t(a^*)\right]\leq\tilde{O}\left(\mathbb{E}[R_{i^*}(T)]+K\sqrt{T}\right).
$$

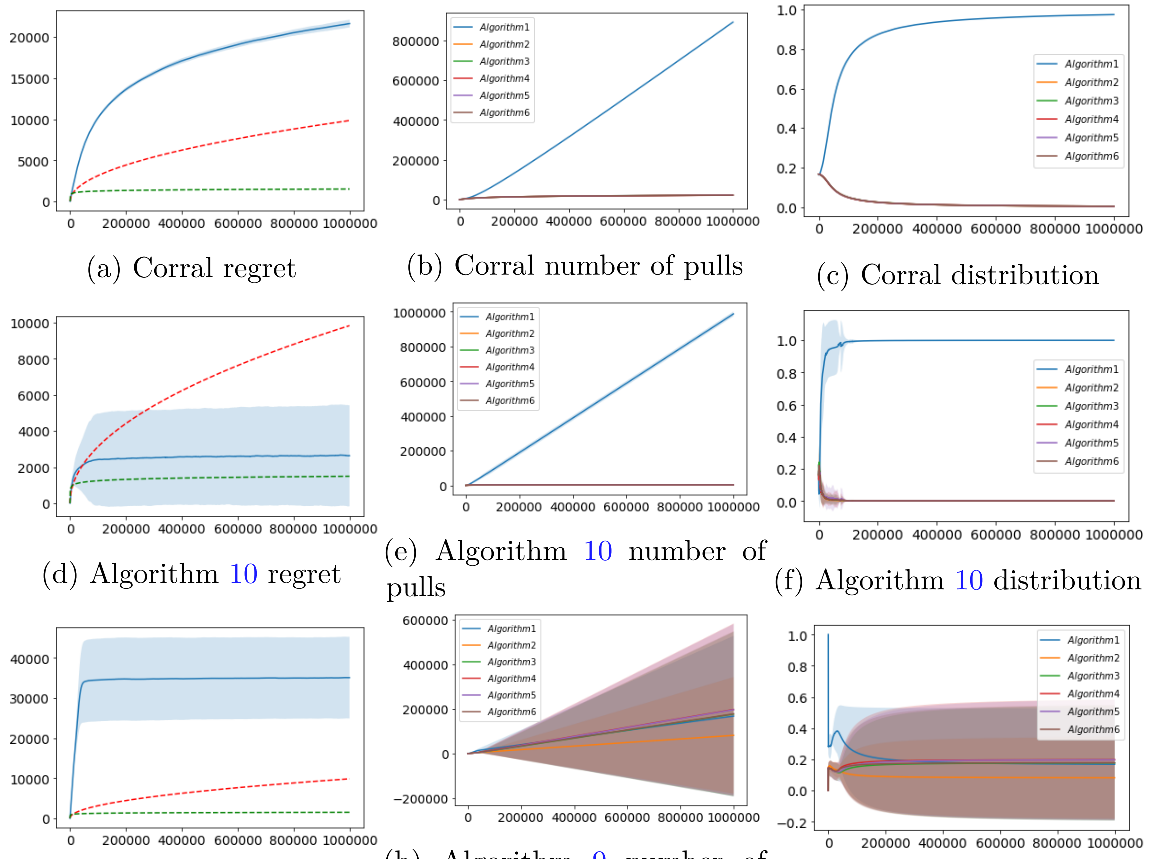
3.11 Empirical results

In this section, we further examine the empirical properties of our algorithms via experiments on synthetically generated datasets. We compare Algorithm 9 and Algorithm 10 to the Corral algorithm (Agarwal et al., 2016)[Algorithm 1], which is also used in (Pacchiano et al., 2020b). We note that Pacchiano et al. (2020b) also use Exp3.P as a corralling algorithm. Recent work (Lee et al., 2020a) suggests that Corral exhibits similar high probability regret guarantees as Exp3.P and that Corral would completely outperform Exp3.P. One of the corralling algorithms in (Pacchiano et al., 2020b) is precisely the Corral algorithm. The second algorithm considered in (Pacchiano et al., 2020b) is the Exp3.P (Auer et al., 2002b) algorithm. We do not compare against Exp3.P as we already expect that the performance will be worse than the Corral algorithm. Our intuition is guided by the fact that the Corral algorithm already comes equipped with high-probability regret guarantees, which match the once provided by Exp3.P, without the need for additional exploration (Lee et al., 2020a).

Experimental setup. The algorithms that we corral are UCB-I, Thompson sampling (TS), and FTRL with $\frac{1}{2}$-Tsallis entropy reguralizer (Tsallis-INF). We chose these algorithms as they all come with regret guarantees for the stochastic multi-armed problem and they broadly represent three different classes of algorithms, i.e, algorithms based on the optimism in the face of uncertainty principle, algorithms based on posterior sampling, and algorithms based on online mirror descent. When implementing Algorithm 10 and Corral, we make an important deviation from what theory prescribes: we ***never*** restart the corralled algorithms and run them with their default parameters. Even though, there are no theoretical guarantees for this modification of the corralling algorithms, we will see that the regret bounds remain meaningful in practice. In all of the experiments we corral two instances of UCB-I, TS, and FTRL for a total of six algorithms. The best algorithm plays over 10 arms. Every other algorithm plays over 5 arms. Intuitively, the higher the number of arms implies higher complexity of the best algorithm which would lead to higher regret and a harder corralling problem. The rewards for each algorithm are Bernoulli random variables setup according to the following parameters: BASE_REWARD, IN_GAP, OUT_GAP, and LOW_REWARD. The best overall arm has expected reward BASE_REWARD + IN_GAP + OUT_GAP. Every other arm of Algorithm 1 has expected reward equal to LOW_REWARD. For all other algorithms the best arm has reward BASE_REWARD + IN_GAP and other arms have reward BASE_REWARD. In all of the experiments we set BASE_REWARD $= 0.5$, IN_GAP $= 0.01$, LOW_REWARD $= 0.2$. While a small IN_GAP implies a large regret for the algorithms containing sub-optimal arms, it also reduces the likelihood that said algorithms would have small average reward. Combined with setting LOW_REWARD $= 0.2$, this will make the average reward of $\mathcal{A}_1$ look small in the initial number of rounds, compared to the average reward of $\mathcal{A}_i, i > 1$ and hence makes the corralling problem harder. We run two set of experiments, an easy set for which OUT_GAP $= 0.19$, which translates to gaps $\Delta_i = 0.2$ in our regret bounds, and a hard set for which OUT_GAP $= 0.01$ which implies $\Delta_i = 0.02$. Finally time horizon is set to $T = 10^6$.

Large gap experiments. Figure 3-1 reports the regret (top) and number of plays of each algorithm found in our experiments when $\Delta_i = 0.2$. The plots represent the average regret, in blue, and the average number of pulls of each algorithm (color according to the legend) over 75 runs of each experiment. The standard deviation is represented by the shaded blue region. The algorithm that contains the optimal arm is $\mathcal{A}_1$ and is an instance of UCB-I. The red dotted line in the top plots is given by $4\sqrt{KT} + \mathbb{E}[R_1(T)]$, and the green dotted line is given by $4\sum_{i\neq 1} \frac{k_i \log(T)}{\Delta_i} + \mathbb{E}[R_1(T)]$. These lines serve as a reference across experiments and we believe they are more accurate upper bounds for the regret of the proposed and existing algorithms. As expected, we see that, in the large gap regime, the Corral algorithm exhibits $\Omega(\sqrt{T})$

regret, while the regret of Algorithm 10 remains bounded in $O(\log(T))$. Algorithm 9 admits two regret phases. In the initial phase, its regret is linear, while in the second phase it is logarithmic. This is typical of UCB strategies in the stochastic MAB problem (Garivier et al., 2018).



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

(f) Algorithm 10 distribution

(g) Algorithm 9 regret

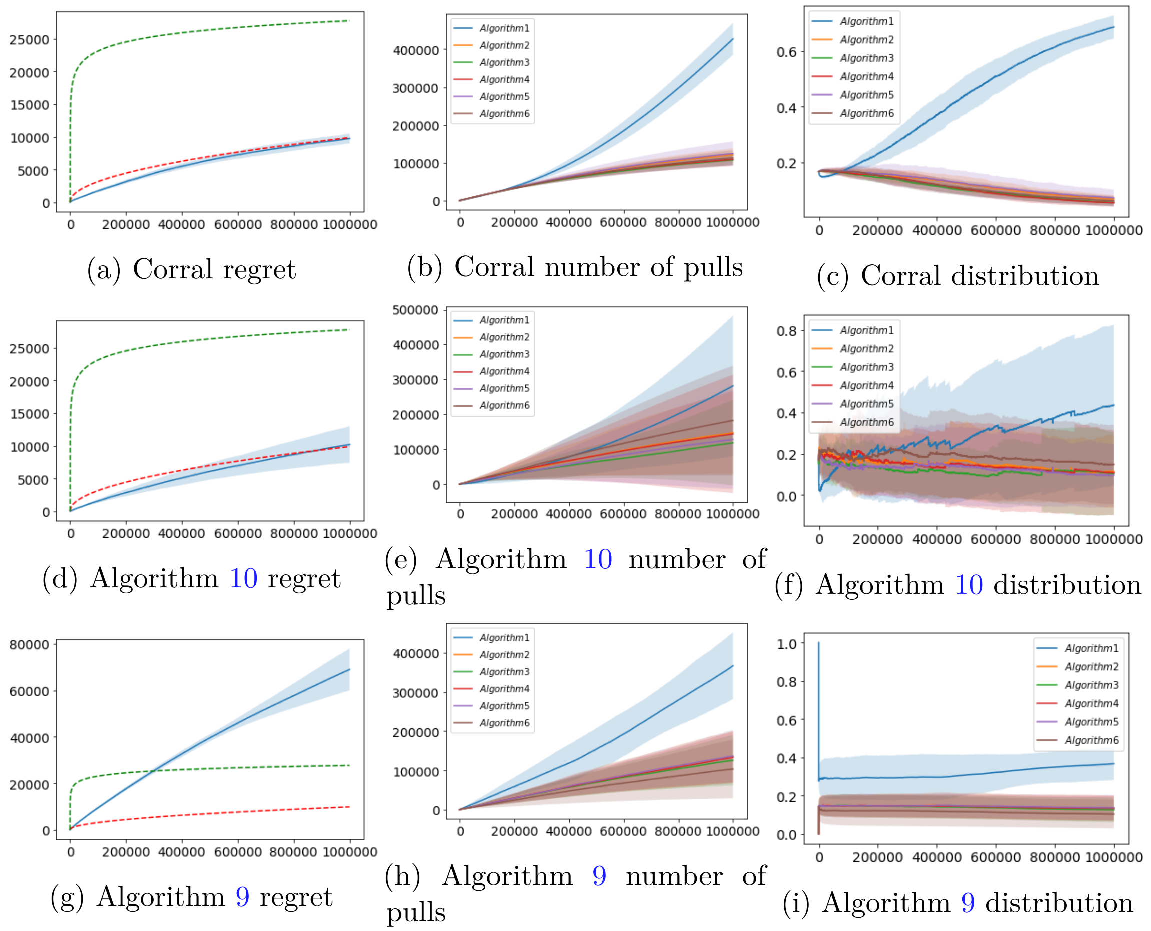
(h) Algorithm 9 number of pulls

(i) Algorithm 9 distribution

Figure 3-1: UCB-I contains best arm,$\Delta_i = 0.2, \mathrm{ALG}_{1:2} = \text{UCB-I}, \mathrm{ALG}_{3:4} = \textsc{Tsallis-INF}, \mathrm{ALG}_{5:6} = \text{TS}$.

Small gap experiments Figure 3-2 reports the results of our experiments for $\Delta_i = 0.02$. The setting of the experiments is the same as in the large gap case. We observe that both Corral and Algorithm 10 behave according to the $O(\sqrt{T})$ bounds. This is expected since, when $\Delta_i = 0.02$, the optimistic bound dominates the $\sqrt{T}$-bound. The result for Algorithm 9 might be somewhat surprising, as its regret exceeds both the green and red lines. We emphasize that this experiment does not contradict

Theorem 3.6.2. Indeed, if we were to plot the green and red lines according to the bounds of Theorem 3.6.2, the regret would remain below both lines.



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

(f) Algorithm 10 distribution

(g) Algorithm 9 regret

(h) Algorithm 9 number of pulls

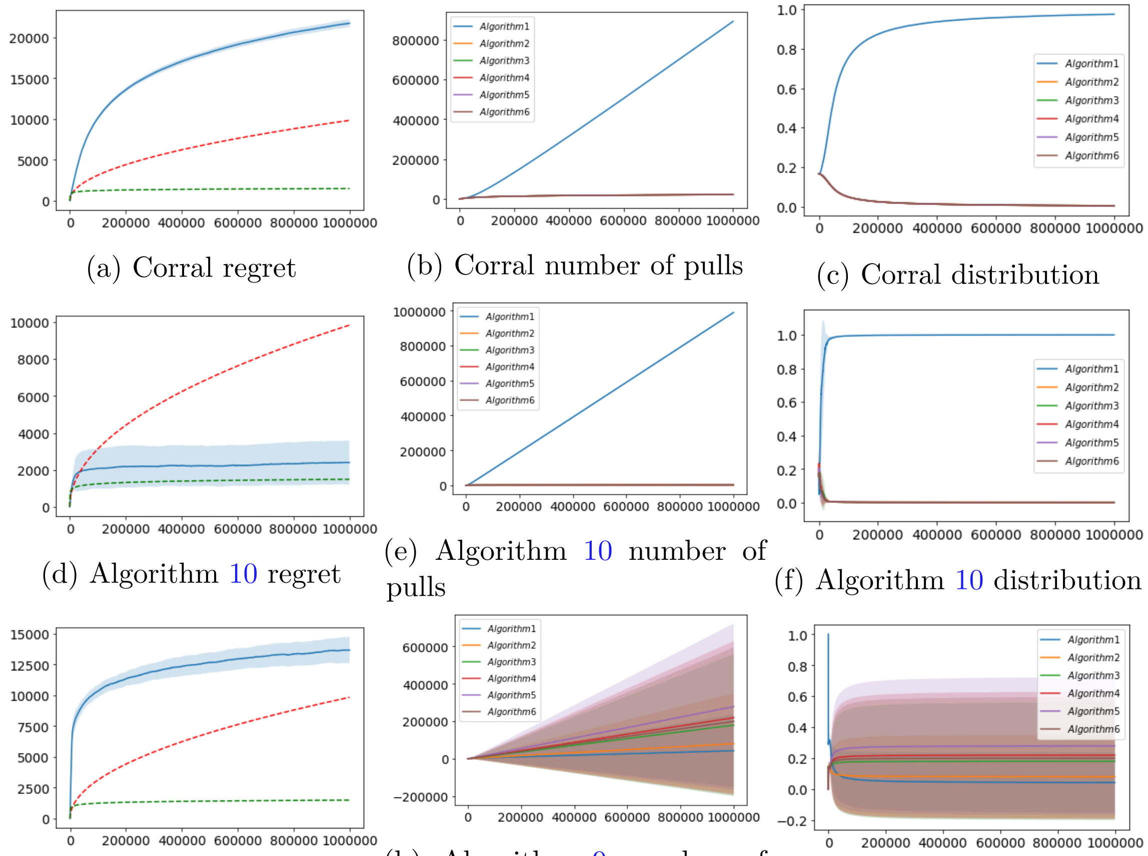
(i) Algorithm 9 distribution

Figure 3-2: UCB-I contains best arm,$\Delta_i = 0.02, \mathrm{ALG}_{1:2} = \text{UCB-I}, \mathrm{ALG}_{3:4} =$ Tsallis-INF$, \mathrm{ALG}_{5:6} = \text{TS}$.

Our experiments suggest that Algorithm 10 is the best corralling algorithm. A tighter analysis would potentially yield optimistic regret bounds in the order of $O\left(\sum_{i \neq i^*} \frac{k_i \log(T)}{\Delta_i} + \mathbb{E}[R_{i^*}(T)]\right)$. Furthermore, we expect that the bounds of Theorem 3.6.2 are tight. In Section 3.11.1 and Section 3.11.2 we show experiments from the same setup as just described, however, the algorithm containing the best arm is 1/2-Tsallis-INF and Thompson Sampling respectively.

3.11.1 Tsallis-INF contains best arm

Experiments can be found in Figure 3-3 for $\Delta_i = 0.2$ and in Figure 3-4 for $\Delta_i = 0.02$.



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

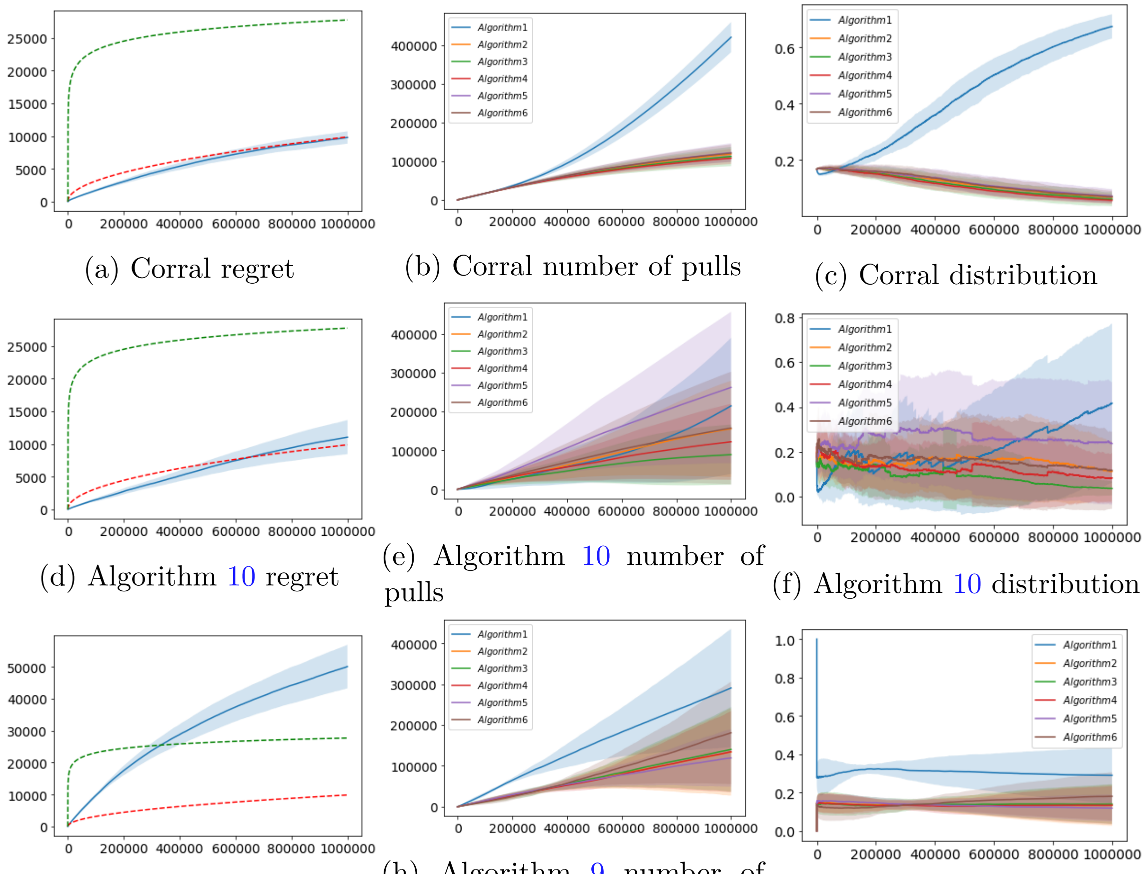
(f) Algorithm 10 distribution

(g) Algorithm 9 regret

(h) Algorithm 9 number of pulls

(i) Algorithm 9 distribution

Figure 3-3: Tsallis-INF contains best arm,$\Delta_i = 0.2, \text{ALG}_{1:2} =$ Tsallis-INF, $\text{ALG}_{3:4} =$ UCB-I, $\text{ALG}_{5:6} =$ TS.



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

(f) Algorithm 10 distribution

(g) Algorithm 9 regret

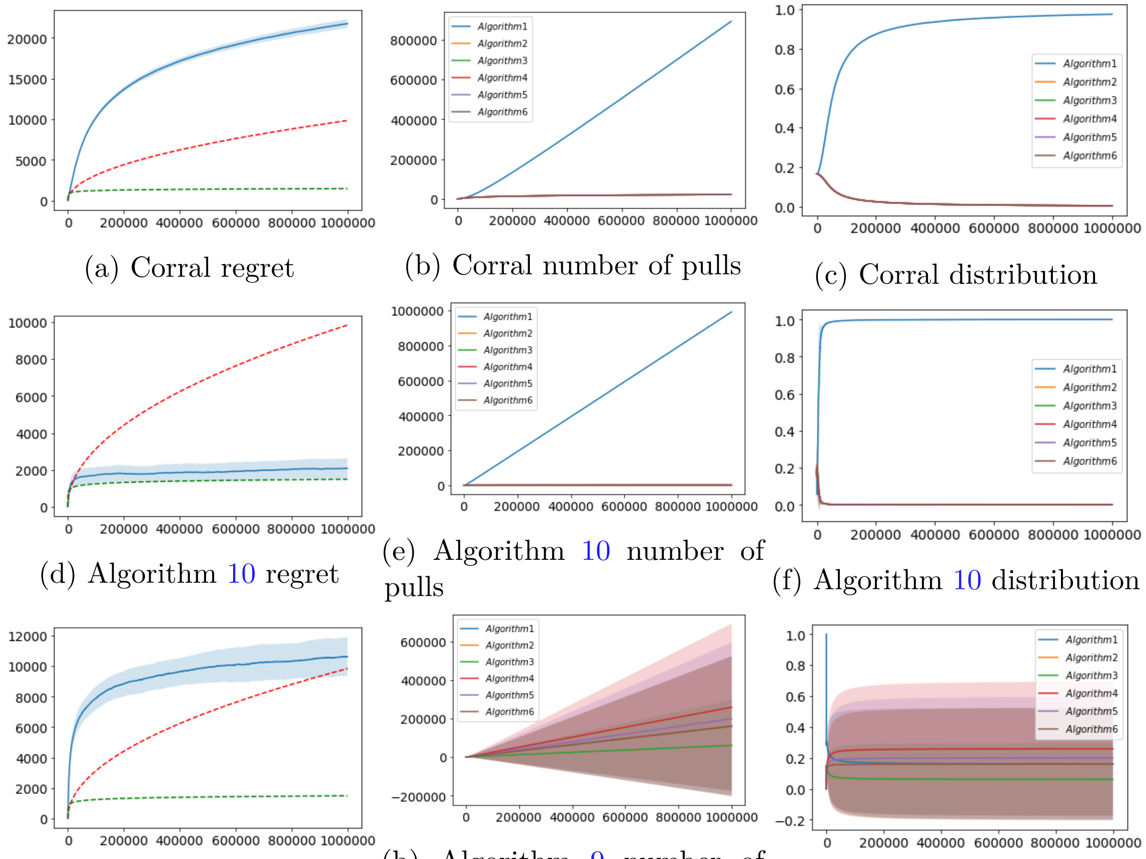
(h) Algorithm 9 number of pulls

(i) Algorithm 9 distribution

Figure 3-4: Tsallis-INF contains best arm, $\Delta_i = 0.02, \mathrm{ALG}_{1:2} = \textsc{Tsallis-INF}, \mathrm{ALG}_{3:4} = \text{UCB-I}, \mathrm{ALG}_{5:6} = \text{TS}$.

3.11.2 Thompson sampling contains best arm

Experiments can be found in Figure 3-5 for $\Delta_i = 0.2$ and in Figure 3-6 for $\Delta_i = 0.02$.



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

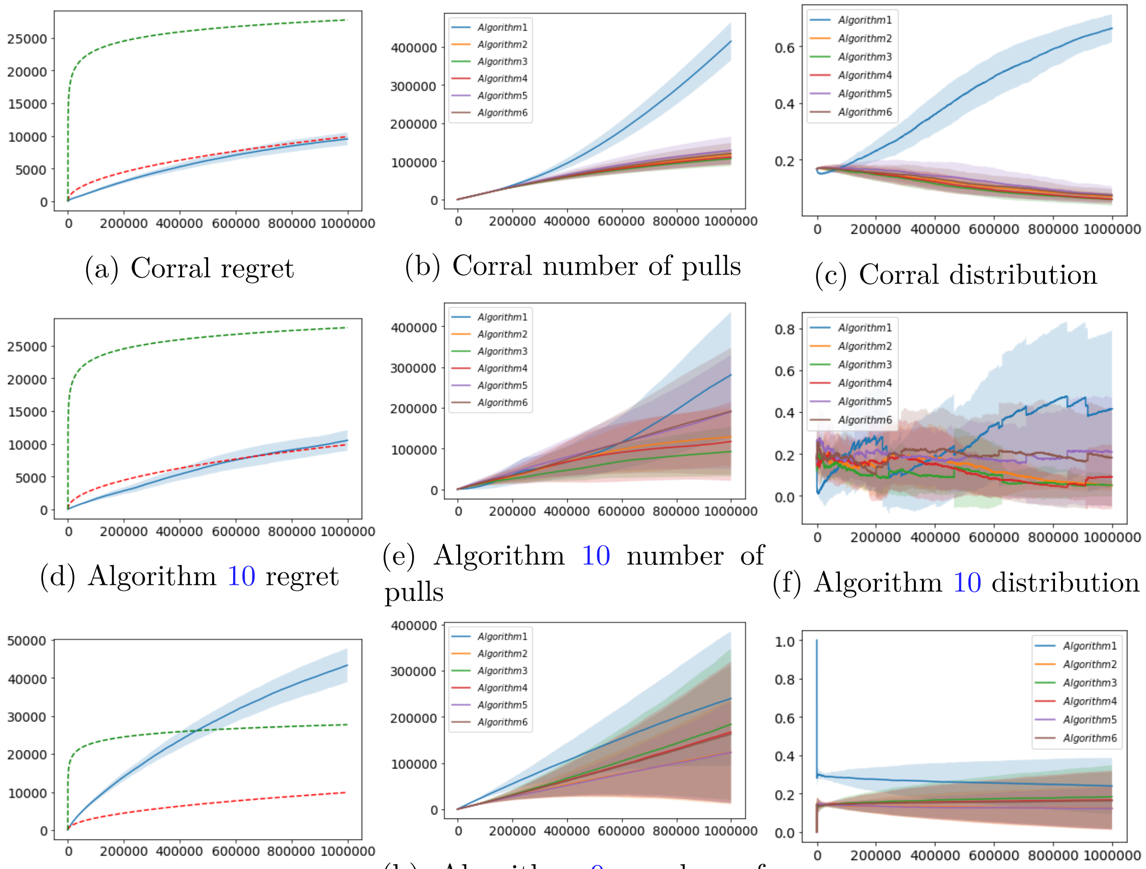
(f) Algorithm 10 distribution

(g) Algorithm 9 regret

(h) Algorithm 9 number of pulls

(i) Algorithm 9 distribution

Figure 3-5: Thompson sampling (TS) contains best arm, $\Delta_i = 0.2, \mathrm{ALG}_{1:2} = \mathrm{TS}, \mathrm{ALG}_{3:4} = \text{UCB-I}, \mathrm{ALG}_{5:6} = \textsc{Tsallis}\text{-INF}$.



(a) Corral regret

(b) Corral number of pulls

(c) Corral distribution

(d) Algorithm 10 regret

(e) Algorithm 10 number of pulls

(f) Algorithm 10 distribution

(g) Algorithm 9 regret

(h) Algorithm 9 number of pulls

(i) Algorithm 9 distribution

Figure 3-6: TS contains best arm, $\Delta_i = 0.02, \text{ALG}_{1:2} = \text{TS}, \text{ALG}_{3:4} = \text{UCB-I}, \text{ALG}_{5:6} = \textsc{Tsallis-INF}$.

Chapter 4

Policy regret in repeated games

We first show that there are online learning settings in which policy regret and external regret are incompatible: any sequence of play that achieves a favorable regret with respect to one definition must do poorly with respect to the other. We then focus on the game-theoretic setting where the adversary is a self-interested agent. In that setting, we show that external regret and policy regret are not in conflict and, in fact, that a wide class of algorithms can ensure a favorable regret with respect to both definitions, so long as the adversary is also using such an algorithm. We also show that the sequence of play of no-policy regret algorithms converges to a ***policy equilibrium***, a new notion of equilibrium that we introduce. Relating this back to external regret, we show that coarse correlated equilibria, which no-external regret players converge to, are a strict subset of policy equilibria. Thus, in game-theoretic settings, every sequence of play with no external regret also admits no policy regret, but the converse does not hold. The main contributions of this chapter are based on Arora et al. (2018). This work was done in collaboration with Dr. Raman Arora, Dr. Michael Dinitz, and Dr. Mehryar Mohri.

4.1 Incompatibility of policy regret and external regret

Arora et al. (2012a) show that there exists an adaptive adversary against which any online learning algorithm admits linear policy regret, even when the external regret may be sublinear, we ask if no policy regret implies no external regret. One could expect this to be the case since policy regret seems to be a ***stronger*** notion than external regret. However, we show that this in fact is ***not*** the case and that the two notions of regret are incompatible: there exist adversaries (or sequence of utilities) on which action sequences with sublinear external regret admit linear policy regret and action sequences with sublinear policy regret incur linear external regret.

Theorem 4.1.1. *There exists a sequence of m-memory bounded utility functions $(u_t)_{t=1}^T$, where $u_t : \mathcal{A} \to \mathbb{R}$, such that for any constant $m \geq 2$ (independent of T), any action sequence with sublinear policy regret will have linear external regret and any action sequence with sublinear external regret will have linear policy regret.*

Proof sketch. The proof of the above theorem constructs a sequence for which no reasonable play can attain sublinear external regret. In particular, the only way the learner can have sublinear external regret is if they choose to have very small utility. To achieve this, the utility functions chosen by the adversary are the following. At time t, if the player chose to play the same action as their past 2 actions then they get utility $\frac{1}{2}$. If the player's past two actions were equal but their current action is different, then they get utility 1, and if their past two actions differ then no matter what their current action is they receive utility 0. It is easy to see that the maximum utility play for this sequence (and the lowest 2-memory bounded policy regret strategy) is choosing the same action at every round. However, such an action sequence admits linear external regret. Moreover, every sublinear external regret strategy must then admit sublinear utility and thus linear policy regret. □

4.2 Policy regret in strategic environments

The incompatibility result rests on the fact that the player is facing a completely malicious adversary. In many realistic environments we can instead think of the adversary as a self-interested agent trying to maximize their own utility, rather than trying to maximize the regret of the player. This more strategic environment is better captured by the game theory setting, in particular a 2-player game where both players are trying to maximize their utility. Even though we have argued that external regret is not a good measure, our next result shows that minimizing policy regret in games can be done if both players choose their strategies according to certain no external regret algorithms. More generally, we adapt a classical notion of stability from the statistical machine learning setting and argue that if the players use no external regret algorithms that are ***stable***, then the players will have no policy regret in expectation. To state the result formally we first need to introduce some notation.

Definition 4.2.1 (Game definition)**.** We consider a 2-player game $\mathcal{G}$, with players 1 and 2. The action set of player i is denoted by $\mathcal{A}_i$, which we think of as being embedded into $\mathbb{R}^{|\mathcal{A}_i|}$ in the obvious way where each action corresponds to a standard basis vector. The corresponding probability simplex is $\Delta\mathcal{A}_i$. The action of player 1 at time t is a_t and of player 2 is b_t. The observed utility for player i at time t is $u_i(a_t, b_t)$ and this is a bi-linear form with corresponding matrix P_i. We assume that the utilities are bounded in $[0, 1]$.

Algorithm of the player. When discussing algorithms, we take the view of player 1. Specifically, at time t, player 1 plays according to an algorithm which can be described as $Alg_t : (\mathcal{A}_1 \times \mathcal{A}_2)^t \rightarrow \Delta\mathcal{A}_1$. We distinguish between two settings: full information, in which the player observes the full utility function at time t (i.e., $u_1(\cdot, b_t)$), and the bandit setting, in which the player only observes $u_1(a_t, b_t)$. In the full information setting, algorithms like multiplicative weight updates (MWU Arora et al. (2012b)) depend only on the past $t-1$ utility functions $(u_1(\cdot, b_\ell))_{\ell=1}^{t-1}$, and thus we can think of Alg_t as a function $f_t : \mathcal{A}_2^t \to \Delta\mathcal{A}_1$. In the bandit setting, though, the output at time t of the algorithm depends both on the previous $t-1$ actions $(a_\ell)_{\ell=1}^{t-1}$ and on the utility functions (i.e., the actions picked by the other player).

But even in the bandit setting, we would like to think of the player's algorithm as a function $f_t : \mathcal{A}_2^t \to \Delta\mathcal{A}_1$. We cannot quite do this, however we ***can*** think of the player's algorithm as a ***distribution*** over such functions. So how do we remove the dependence on $\mathcal{A}_1^t$? Intuitively, if we fix the sequence of actions played by player 2, we want to take the expectation of Alg_t over possible choices of the t actions played by player 1. In order to do this more formally, consider the distribution μ over $\mathcal{A}_1^{t-1} \times \mathcal{A}_2^{t-1}$ generated by simulating the play of the players for t rounds. Then let $\mu_{b_{0:t}}$ be the distribution obtained by conditioning μ on the actions of player 2 being $b_{0:t}$. Now we let $f_t(b_{0:t-1})$ be the distribution obtained by sampling $a_{0:t-1}$ from $\mu_{b_{1:t-1}}$ and using $Alg(a_{0:t-1}, b_{0:t-1})$. When taking expectations over f_t, the expectation is taken with respect to the above distribution. We also refer to the output $p_t = f_t(b_{0:t-1})$ as the strategy of the player at time t. Now that we can refer to algorithms simply as functions (or distributions over functions), we introduce the notion of a stable algorithm.

Definition 4.2.2. Let $f_t : \mathcal{A}_2^t \to \Delta\mathcal{A}_1$ be a sample from Alg_t (as described above), mapping the past t actions in $\mathcal{A}_2$ to a distribution over the action set $\mathcal{A}_1$. Let the distribution returned at time t be $p_t^1 = f_t(b_1, \ldots, b_t)$. We call this algorithm ***on average*** $(m, S(T))$ ***stable*** with respect to the norm $\|\cdot\|$, if for any $b'_{t-m+1}, \ldots, b'_t \in \mathcal{A}_2$ such that $\tilde{p}_t^1 = f_t(b_1, \ldots, b_{t-m}, b'_{t-m+1}, \ldots, b'_t) \in \Delta\mathcal{A}_1$, it holds that $\mathbb{E}[\sum_{t=1}^T \|p_t^1 - \tilde{p}_t^1\|] \leq S(T)$, where the expectation is taken with respect to the randomization in the algorithm.

Even though this definition of stability is given with respect to the game setting, it is not hard to see that it can be extended to the general online learning setting, and in fact this definition is similar in spirit to the one given in Saha et al. (2012). It turns out that most natural no external regret algorithms are stable. In particular we show, in the supplementary, that both Exp3 (Auer et al., 2002b) and MWU are on average $(m, m\sqrt{T})$ stable with respect to ℓ_1 norm for any $m < o(\sqrt{T})$. It is now possible to show that if each of the players are facing stable no external regret algorithms, they will also have bounded policy regret (so the incompatibility from Theorem 4.1.1 cannot occur in this case).

Theorem 4.2.1. *Let $(a_t)_{t=1}^T$ and $(b_t)_{t=1}^T$ be the action sequences of player 1 and 2 and suppose that they are coming from no external regret algorithms modeled by functions f_t and g_t, with regrets $R_1(T)$ and $R_2(T)$ respectively. Assume that the algorithms are on average $(m, S(T))$ stable with respect to the ℓ_2 norm. Then*

$$\mathbb{E}\left[\sum_{t=1}^T u_1(a, g_t(a_{0:t-m}, a, \ldots, a)) - u_1(a_t, g_t(a_{0:t-1}))\right] \leq \|\mathrm{P}_1\| S(T) + R_1(T)$$
$$\mathbb{E}\left[\sum_{t=1}^T u_2(f_t(b_{0:t-m}, b, \ldots, b), b) - u_2(f_t(b_{0:t-1}), b_t)\right] \leq \|\mathrm{P}_2\| S(T) + R_2(T),$$

The above holds for any fixed actions $b \in \mathcal{A}_2$ and $a \in \mathcal{A}_1$. Here the matrix norm $\|\cdot\|$ is the spectral norm. [1]

Proof.

$$\begin{aligned}
&\mathbb{E}\left[u_2(f_t(b_0, \cdots, b_{t-m+1}, b, \cdots, b), b) - \sum_{t=1}^T u_2(f_t(b_0, \cdots, b_{t-2}, b_{t-1}), b_t)\right] \\
&= \mathbb{E}\left[u_2(f_t(b_0, \cdots, b_{t-m+1}, b, \cdots, b), b) - \sum_{t=1}^T u_2(f_t(b_0, \cdots, b_{t-2}, b_{t-1}), b)\right] \\
&+ \mathbb{E}\left[u_2(f_t(b_0, \cdots, b_{t-1}), b) - \sum_{t=1}^T u_2(f_t(b_0, \cdots, b_{t-2}, b_{t-1}), b_t)\right] \\
&\leq \sum_{t=1}^T \|b^\top \mathrm{P}_2\|_2 \|f_t(b_0, \cdots, b_{t-m+1}, b, \cdots, b) - f_t(b_0, \cdots, b_{t-2}, b_{t-1})\|_2 + R_2(T) \\
&\leq \|b\|_1 \|\mathrm{P}_2\| S(T) + R_2(T) = \|\mathrm{P}_2\| S(T) + R_2(T)
\end{aligned}$$

[1] We would like to thank Mengxiao Zhang (USC) for suggesting how to improve on the above theorem and discovering a small error in one of our proofs, which has been corrected.

where the first inequality holds by Cauchy-Schwartz, the second inequality holds by using the m-stability of the algorithm, together with the inequality between l_1 and l_2 norms. □

In Theorem 4.2.1 the quantity $\mathbb{E}\left[\sum_{t=1}^{T} u_1(a, g_t(a_{0:t-m}, a, \ldots, a)) - u_1(a_t, g_t(a_{0:t-1}))\right]$ is precisely the m-memory bounded policy regret with respect to the fixed action $a \in \mathcal{A}_1$. To see this, consider the m-memory bounded reward function $r_t(\cdot) := u_1(\cdot, g_t(a_{0:t-m}, \cdot)) : \mathcal{A}_1^m \to [0, 1]$.

The next result shows that both Exp3 and MWU are on average $(m, m\sqrt{T})$ stable algorithms.

Theorem 4.2.2. *MWU is an on average $(m, m\sqrt{T})$ stable algorithm with respect to ℓ_1, for any $m < o(\sqrt{T})$. Further, Exp3 is an on average $(m, m\sqrt{T})$ stable algorithm with respect to ℓ_1, for any $m < o(\sqrt{T})$.*

Proof sketch. The proof idea is to show that consecutive iterates p_t, p_{t+1} of Exp3 and MWU do not differ by more than $\frac{1}{\sqrt{T}}$ in ℓ_1 norm. This allows us to reason about the drift in iterates induced by two different loss sequences which differed at time $t - m$. □

Combining the above results together, we can show that both players will also have no-policy regret for any $m < o(\sqrt{T})$.

Corollary 4.2.3. *Let $(a_t)_{t=1}^T$ and $(b_t)_{t=1}^T$, be the action sequences of players 1 and 2 respectively and suppose that the sequences are coming from MWU or Exp3. Then for any fixed m, it holds:*

$$\mathbb{E}\left[u_2(f_t(b_{0:t-m+1}, b, \cdots, b), b) - \sum_{t=1}^{T} u_2(f_t(b_{0:t-1}), b_t)\right] \leq O(\|\mathrm{P}_2\| m\sqrt{T}),$$

for any fixed action $b \in \mathcal{A}_2$, where f_t are the functions corresponding to the MWU algorithm used by player 1.

Proof. From Theorem 4.2.2 it follows that MWU and Exp3 are on average $(m, m\sqrt{T})$ stable and have regret at most $O(\sqrt{T})$. □

4.3 Detailed proofs for Section 4.1 and Section 4.2

Proof of Theorem 4.1.1. Define the following reward functions

$$r_t(a_{t-m+1}, .., a_t) = \begin{cases} 1 & a_{t-m+i} = a_{t-m+i+1} = 1 \text{ for } i \in \{1, .., m-2\} \wedge a_{t-1} \neq a_t \\ \frac{1}{2} & a_{t-m+i} = a_{t-m+i+1} = 1 \text{ for } i \in \{1, .., m-1\} \\ 0 & \text{otherwise} \end{cases}.$$

Let $(a_t)_{t=1}^T$ be a sequence with sublinear policy regret. Then this sequence has total reward at least $\frac{T}{2} - o(T)$ and so there are at most $o(T)$ actions in the sequence which are not equal to 1. Let the subsequence consisting of all $a_t = 0$ be indexed by $\mathcal{I}$. Define $\tilde{\mathcal{I}} = \{t, t+1, \cdots, t+m-1 : t \in \mathcal{I}\}$ and consider the subsequence of functions $(r_t)_{t \notin \tilde{\mathcal{I}}}$. This is precisely the sequence of functions which have reward $\frac{1}{2}$ with respect to the sequence of play $(a_t)_{t=1}^T$. Notice that the length of this sequence is at least $T - mo(T) = T - o(T)$. The reward of this sequence is $\sum_{t \notin \tilde{\mathcal{I}}} r_t(a_{t-m+1}, .., a_t) = \sum_{t \notin \tilde{\mathcal{I}}} r_t(1, .., 1) = \frac{T-o(T)}{2}$, however, this subsequence has linear external regret, since $\sum_{t \notin \tilde{\mathcal{I}}} r_t(a_{t-m+1}, .., a_{t-1}, 0) = \sum_{t \notin \tilde{\mathcal{I}}} r_t(1, .., 1, 0) = T - o(T)$. Thus the external regret of $(a_t)_{t=1}^T$ is

$$\begin{aligned}
\sum_{t=1}^{T} [r_t(a_{t-m+1}, .., 0) - r_t(a_{t-m+1}, .., a_t)] &= \sum_{t \notin \tilde{\mathcal{I}}} [r_t(a_{t-m+1}, .., 0) - r_t(a_{t-m+1}, .., a_t)] \\
&\quad + \sum_{t \in \tilde{\mathcal{I}}} [r_t(a_{t-m+1}, .., 0) - r_t(a_{t-m+1}, .., a_t)] \\
&\geq \frac{T - o(T)}{2} + \sum_{t \in \tilde{\mathcal{I}}} [r_t(a_{t-m+1}, .., 0) - r_t(a_{t-m+1}, .., a_t)] \\
&\geq \frac{T}{2} - o(T),
\end{aligned}$$

where the last inequality follows from the fact that the cardinality of $\tilde{\mathcal{I}}$ is at most $o(T)$ and thus $\sum_{t \in \tilde{\mathcal{I}}} [r_t(a_{t-m+1}, .., 0) - r_t(a_{t-m+1}, .., a_t)] \geq -o(T)$.

Assume that $(a_t)_{t=1}^T$ has sublinear external regret. From the above argument, it follows that the reward of the sequence is at most $o(T)$ (otherwise if the sequence has reward $\omega(T)$, we can repeat the previous argument and get a contradiction with the fact the sequence has no-external regret). This implies that the the policy regret of the sequence is $\sum_{t=1}^{T}[r_t(1,1,\cdots,1) - r_t(a_{t-m+1},..,a_t)] = \frac{T}{2} - o(T)$. □

Proof of Theorem 4.2.2. We first show the result for MWU. We think of MWU as Exponentiated Gradient (EG) where the loss vector has i-th entry equal to the negative utility if the player decided to play action i. Let the observed loss vector at time j be $\widehat{l}_j$ and the output distribution be p_j, then the update of EG can be written as $p_{j+1} = \arg\min_{p\in\mathcal{C}}\langle\widehat{l}_j, p\rangle + \frac{1}{\eta}D(p,p_j)$, where C is the simplex of the set of possible actions and D is the KL-divergence. Using Lemma 3 in Saha et al. (2012), with $f(p) = \langle\widehat{l}_j, p\rangle + \frac{1}{\eta}D(p,p_j)$ and the fact that the KL-divergence is 1-strongly convex over the simplex $\mathcal{C}$ with respect to the ℓ_1 norm, we have:

$$\begin{aligned}\frac{1}{2\eta}\|p_j - p_{j+1}\|_1^2 \leq f(p_j) - f(p_{j+1}) &= \langle p_j - p_{j+1}, \widehat{l}_j\rangle - \frac{1}{\eta}D(p_{j+1},p_j)\\ &\leq \|p_j - p_{j+1}\|_1\|\widehat{l}_j\|_\infty \leq \|p_j - p_{j+1}\|_1,\end{aligned}$$

where the second inequality follows from Hölder's inequality and the fact that $D(p_{j+1},p_j) \geq 0$. Thus with step size $\eta \sim \sqrt{\frac{1}{T}}$, we have $\|p_j - p_{j+1}\|_1 \leq \frac{1}{2\sqrt{T}}$. Using triangle inequality, we can get $\|p_{j-1} - p_{j+1}\|_1 \leq \|p_{j-1} - p_j\|_1 + \|p_j - p_{j+1}\|_1 \leq \frac{2}{2\sqrt{T}}$ and induction shows that $\|p_{j-m+1} - p_{j+1}\|_1 \leq \frac{m}{2\sqrt{T}}_1$. Suppose for the last m iterations, a fixed loss function l_a was played instead and the resulting output of the algorithm becomes $\tilde{p}_{j+1}$. Then using the same argument as above we have $\|p_{j-m+1} - \tilde{p}_{j+1}\|_1 \leq \frac{m}{2\sqrt{T}}$ and thus $\|p_{j+1} - \tilde{p}_{j+1}\|_1 \leq \frac{m}{\sqrt{T}}$. Summing over all T rounds concludes the proof.

Now we show the result holds in expectation for Exp3. The update at time t, conditioning on the the draw being i, is given by $p_{t+1}^i = \frac{w_t^i\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right)}{w_t^i\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right)+\sum_{j\neq i}w_t^j}$ and for $j \neq i$, $p_{t+1}^j = \frac{w_t^j}{w_t^i\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right)+\sum_{j\neq i}w_t^j}$, where u_t is the utility vector at time t, w_t is the weight vector at time t i.e. $w_{t+1}^i = w_t^i\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right)$, $w_{t+1}^j = w_t^j$, and k is the number of actions. We have the following bound:

$$|p_{t+1}^i - p_t^i| = \left|\frac{w_t^i \exp\left(\frac{\gamma}{kp_t^i}u_t^i\right)}{w_t^i \exp\left(\frac{\gamma}{kp_t^i}u_t^i\right) + \sum_{j\neq i} w_t^j} - \frac{w_t^i}{\sum_j w_t^j}\right| \leq \left|\frac{w_t^i(\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right) - 1)}{\sum_j w_t^j}\right|$$
$$= p_t^i\left(\exp\left(\frac{\gamma}{kp_t^i}u_t^i\right) - 1\right) \leq p_t^i 2\frac{\gamma}{kp_t^i}u_t^i \leq 2\frac{\gamma}{k},$$

where the first inequality uses the fact that $p_{t+1}^i \geq p_t^i$ and the second inequality uses the choice of γ together with $\exp(x) \leq 2x+1$ for $x \in [0,1]$. Similarly for $j \neq i$, we have:

$$|p_{t+1}^j - p_t^j| = \left|\frac{w_t^j}{w_t^i \exp\left(\frac{\gamma}{kp_t^i}u_t^i\right) + \sum_{j\neq i} w_t^j} - \frac{w_t^j}{\sum_j w_t^j}\right| \leq \left|\frac{w_t^j}{\sum_j w_t^j}\left(1 - \frac{1}{\exp\left(\frac{\gamma u_t^i}{kp_t^i}\right)}\right)\right|$$
$$= p_t^j\left(1 - \exp\left(-\frac{\gamma u_t^i}{kp_t^i}\right)\right) \leq p_t^j\frac{\gamma u_t^i}{kp_t^i} \leq \frac{p_t^j}{kp_t^i}\gamma,$$

where we have used $p_{t+1}^j \leq p_t^j$ and $\exp(-x) \geq 1-x$ for all x. We can now proceed to bound $\mathbb{E}_{i_t}[\|p_{t+1} - p_t\|_1 | i_{1:t-1}]$, where i_t is the random variable denoting the draw at time t:

$$\mathbb{E}_{i_t}\left[\|p_{t+1} - p_t\|_1 | i_{1:t-1}\right] = \sum_i p_t^i \|p_{t+1} - p_t\|_1 = \sum_i p_t^i\left(|p_{t+1}^i - p_t^i| + \sum_{j\neq i}|p_{t+1}^j - p_t^j|\right)$$
$$\leq 2\gamma + \sum_{i,j} p_t^i \frac{p_t^j}{kp_t^i}\gamma = 3\gamma.$$

Setting $\gamma \sim \frac{1}{\sqrt{T}}$ finishes the proof. □

4.4 Policy equilibrium

Recall that unlike external regret, policy regret captures how other players in a game might react if a player decides to deviate from their strategy. The story is similar when considering different notions of equilibria. In particular Nash equlibria, Correlated equilibria and CCEs can be interpreted in the following way: if player i deviates from the equilibrium play, their utility will not increase no matter how they decide to switch, provided that ***all other players continue to play according to the***

equilibrium. This sentiment is a reflection of what no external and no swap regret algorithms guarantee. Equipped with the knowledge that no policy regret sequences are obtainable in the game setting under reasonable play from all parties, it is natural to reason how other players would react if player i deviated and what would be the cost of deviation when taking into account possible reactions.

Let us again consider the 2-player game setup through the view of player 1. The player believes their opponent might be m-memory bounded and decides to proceed by playing according to a no policy regret algorithm. After many rounds of the game, player 1 has computed an empirical distribution of play $\widehat{\sigma}$ over $\mathcal{A} := \mathcal{A}_1 \times \mathcal{A}_2$. The player is familiar with the guarantees of the algorithm and knows that, if instead, they changed to playing any fixed action $a \in \mathcal{A}_1$, then the resulting empirical distribution of play $\widehat{\sigma}_a$, where player 2 has responded accordingly in a memory-bounded way, is such that $\mathbb{E}_{(a,b)\sim\widehat{\sigma}}\left[u_1(a,b)\right] \geq \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a}\left[u_1(a,b)\right] - \epsilon$. This thought experiment suggests that if no policy regret play converges to an equilibrium, then the equilibrium is not only described by the deviations of player 1, but also through the change in player 2's behavior, which is encoded in the distribution $\widehat{\sigma}_a$. Thus, any equilibrium induced by no policy regret play, can be described by tuples of distributions $\{(\sigma, \sigma_a, \sigma_b) : (a,b) \in \mathcal{A}\}$, where σ_a is the distribution corresponding to player 1's deviation to the fixed action $a \in \mathcal{A}_1$ and σ_b captures player 2's deviation to the fixed action $b \in \mathcal{A}_2$. Clearly σ_a and σ_b are not arbitrary but we still need a formal way to describe how they arise.

For convenience, lets restrict the memory of player 2 to be 1. Thus, what player 1 believes is that at each round t of the game, they play an action a_t and player 2 plays a function $f_t : \mathcal{A}_1 \to \mathcal{A}_2$, mapping a_{t-1} to $b_t = f_t(a_{t-1})$. Finally, the observed utility is $u_1(a_t, f_t(a_{t-1}))$. The empirical distribution of play, $\widehat{\sigma}$, from the perspective of player 1, is formed from the observed play $(a_t, f_t(a_{t-1}))_{t=1}^T$. Moreover, the distribution, $\widehat{\sigma}_a$, that would have occurred if player 1 chose to play action a on every round is formed from the play $(a, f_t(a))_{t=1}^T$. In the view of the world of player 1, the actions taken by player 2 are actually functions rather than actions in $\mathcal{A}_2$. This suggests that the equilibrium induced by a no-policy regret play, is a distribution over the functional space defined below.

Definition 4.4.1. Let $\mathcal{F}_1 := \{f : \mathcal{A}_2^{m_1} \to \mathcal{A}_1\}$ and $\mathcal{F}_2 := \{g : \mathcal{A}_1^{m_2} \to \mathcal{A}_2\}$ denote the ***functional spaces of play*** of players 1 and 2, respectively. Denote the product space by $\mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2$.

Note that when $m_1 = m_2 = 0$, $\mathcal{F}$ is in a one-to-one correspondence with $\mathcal{A}$, i.e. when players believe their opponents are oblivious, we recover the action set studied in standard equilibria. For simplicity, for the remainder of the paper we assume that $m_1 = m_2 = 1$. However, all of the definitions and results that follow can be extended to the fully general setting of arbitrary m_1 and m_2; see the supplementary for details.

Let us now investigate how a distribution π over $\mathcal{F}$ can give rise to a tuple of distributions $(\widehat{\sigma}, \widehat{\sigma}_a, \widehat{\sigma}_b)$.

We begin by defining the utility of π such that it equals the utility of a distribution σ over $\mathcal{A}$ i.e., we want $\mathbb{E}_{(f,g)\sim\pi}\left[u_1(f,g)\right] = \mathbb{E}_{(a,b)\sim\sigma}\left[u_1(a,b)\right]$. Since utilities are not defined for functions, we need an interpretation of $\mathbb{E}_{(f,g)\sim\pi}\left[u_1(f,g)\right]$ which makes sense. We notice that π induces a Markov chain with state space $\mathcal{A}$ in the following way.

Definition 4.4.2. Let π be any distribution over $\mathcal{F}$. Then π ***induces a Markov process*** with transition probabilities $\mathbb{P}\left[(a_2,b_2)|(a_1,b_1)\right] = \sum_{(f,g)\in\mathcal{F}_1\times\mathcal{F}_2: f(b_1)=a_2, g(a_1)=b_2} \pi(f,g)$. We associate with this Markov process the transition matrix $\mathrm{M} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$, with $\mathrm{M}_{x_1,x_2} = \mathbb{P}\left[x_2|x_1\right]$ where $x_i = (a_i, b_i)$.

Since every Markov chain with a finite state space has a stationary distribution, we think of utility of π as the utility of a particular stationary distribution σ of M. How we choose σ among all stationary distributions is going to become clear later, but for now we can think about σ as the distribution which maximizes the utilities of both players. Next, we need to construct σ_a and σ_b, which capture the deviation in play, when player 1 switches to action a and player 2 switches to action b. The no-policy regret guarantee can be interpreted as $\mathbb{E}_{(f,g)\sim\pi}\left[u_1(f,g)\right] \geq \mathbb{E}_{(f,g)\sim\pi}\left[u_1(a,g(a))\right]$ i.e., if player 1 chose to switch to a fixed action (or equivalently, the constant function which maps everything to the action $a \in \mathcal{A}_1$), then their utility should not increase. Switching to a fixed action a, changes π to a new distribution π_a over $\mathcal{F}$. This turns out to be a product distribution which also induces a Markov chain.

Definition 4.4.3. Let π be any distribution over $\mathcal{F}$. Let δ_a be the distribution over $\mathcal{F}_1$ putting all mass on the constant function mapping all actions $b \in \mathcal{A}_2$ to the fixed action $a \in \mathcal{A}_1$. Let $\pi_{\mathcal{F}_2}$ be the marginal of π over $\mathcal{F}_2$. The ***distribution resulting from player 1 switching to playing a fixed action*** $a \in \mathcal{A}$, is denoted as $\pi_a = \delta_a \times \pi_{\mathcal{F}_2}$. This distribution induces a Markov chain with transition probabilities $\mathbb{P}\left[(a,b_2)|(a_1,b_1)\right] = \sum_{(f,g):g(a_1)=b_2} \pi(f,g)$ and ***the transition matrix of this Markov process is denoted by*** M_a. The distribution π_b and matrix M_b are defined similarly for player 2.

Since the no policy regret algorithms we work with do not directly induce distributions over the functional space $\mathcal{F}$ but rather only distributions over the action space $\mathcal{A}$, we would like to state all of our utility inequalities in terms of distributions over $\mathcal{A}$. Thus, we would like to check if there is a stationary distribution σ_a of M_a such that $\mathbb{E}_{(f,g)\sim\pi}\left[u_1(a,g(a))\right] = \mathbb{E}_{(a,b)\sim\sigma_a}\left[u_1(a,b)\right]$. This is indeed the case as verified by the following theorem.

Theorem 4.4.1. *Let π be a distribution over the product of function spaces $\mathcal{F}_1 \times \mathcal{F}_2$. There exists a stationary distribution σ_a of the Markov chain M_a for any fixed $a \in \mathcal{A}_1$ such that $\mathbb{E}_{(a,b)\sim\sigma_a}\left[u_1(a,b)\right] = \mathbb{E}_{(f,g)\sim\pi}\left[u_1(a,g(a))\right]$. Similarly, for every fixed action $b \in \mathcal{A}_2$, there exists a stationary distribution σ_b of M_b such that $\mathbb{E}_{(a,b)\sim\sigma_b}\left[u_2(a,b)\right] = \mathbb{E}_{(f,g)\sim\pi}\left[u_2(f(b),b)\right]$.*

Proof. The proof is constructive and we include it below.

Note that, by definition $(\mathrm{M}_a)_{(\tilde{a},\tilde{b}),(\widehat{a},\widehat{b})} = 0$ if $\widehat{a} \neq a$ and $M_{(\tilde{a},\tilde{b}),(\widehat{a},\widehat{b})} = \sum_{f,g:g(\tilde{a})=\widehat{b}} \pi(f,g)$ if $\widehat{a} = a$. Consider the distribution $\tilde{\sigma}$ over $\mathcal{A}$, where $\tilde{\sigma}_{(\tilde{a},\tilde{b})} = 0$ if $\tilde{a} \neq a$ and $\tilde{\sigma}_{(\tilde{a},\tilde{b})} = \sum_{f,g:g(a)=\tilde{b}} \pi(f,g)$ if $\tilde{a} = a$. We now show that $\tilde{\sigma}$ is a stationary distribution of M_a:

$$\begin{aligned}
\left(\tilde{\sigma}^\top \mathrm{M}_a\right)_{(a,\tilde{b})} &= \sum_{(\widehat{a},\widehat{b})} \tilde{\sigma}_{(\widehat{a},\widehat{b})} (\mathrm{M}_a)_{(\widehat{a},\widehat{b}),(a,\tilde{b})} = \sum_{\widehat{b}} \tilde{\sigma}_{(a,\widehat{b})} (\mathrm{M}_a)_{(a,\widehat{b}),(a,\tilde{b})} \\
&= \sum_{\widehat{b}} \left(\sum_{f,g:g(a)=\widehat{b}} \pi(f,g) \right) \left(\sum_{f,g:g(a)=\tilde{b}} \pi(f,g) \right) \\
&= \left(\sum_{f,g:g(a)=\tilde{b}} \pi(f,g) \right) \left(\sum_{\widehat{b}} \sum_{f,g:g(a)=\widehat{b}} \pi(f,g) \right) \\
&= \sum_{f,g:g(a)=\tilde{b}} \pi(f,g) = \tilde{\sigma}_{(a,\tilde{b})}.
\end{aligned}$$

Finally, notice that:

$$\mathbb{E}_{(f,g)\sim\pi} \left[u_1(a, g(a)) \right] = \sum_{b \in \mathcal{A}_2} u_1(a,b) \mathbb{P} \left[g(a) = b \right] = \sum_{b \in \mathcal{A}_2} u_1(a,b) \sum_{f,g:g(a)=b} \pi(f,g).$$

□

With all of this notation we are ready to formally describe what no-policy regret play promises in the game setting in terms of an equilibrium.

Definition 4.4.4. A distribution π over $\mathcal{F}_1 \times \mathcal{F}_2$ is a ***policy equilibrium*** if for all fixed actions $a \in \mathcal{A}_1$ and $b \in \mathcal{A}_2$, which generate Markov chains M_a and M_b respectively, with stationary distributions σ_a and σ_b from Theorem 4.4.1, there exists a stationary distribution σ of the Markov chain M induced by π such that:

$$\begin{aligned}
\mathbb{E}_{(a,b)\sim\sigma} \left[u_1(a,b) \right] &\geq \mathbb{E}_{(a,b)\sim\sigma_a} \left[u_1(a,b) \right] \\
\mathbb{E}_{(a,b)\sim\sigma} \left[u_2(a,b) \right] &\geq \mathbb{E}_{(a,b)\sim\sigma_b} \left[u_2(a,b) \right].
\end{aligned} \tag{4.1}$$

In other words, π is a policy equilibrium if there exists a stationary distribution σ of the Markov chain corresponding to π, such that, when actions are drawn according to σ, no player has incentive to change their action. We present a simple example of a policy equilibrium see Section 4.4.4.

4.4.1 Convergence to the set of policy equilibria

We have tried to formally capture the notion of equilibria in which player 1's deviation would lead to a reaction from player 2 and vice versa in Definition 4.4.4. This definition is inspired by the counter-factual guarantees of no policy regret play and we would like to check that if players' strategies yield sublinear policy regret then the play converges to a policy equilibrium. Since the definition of sublinear policy regret does not include a distribution over functional spaces but only works with empirical distributions of play, we would like to present our result in terms of distributions over the action space $\mathcal{A}$. Thus we begin by defining the set of all product distributions $\sigma \times \sigma_a \times \sigma_b$, induced by policy equilibria π as described in the previous subsection. Here σ_a and σ_b represent the deviation in strategy if player 1 changed to playing the fixed action $a \in \mathcal{A}_1$ and player 2 changed to playing the fixed action $b \in \mathcal{A}_2$ respectively as constructed in Theorem 4.4.1.

Definition 4.4.5. For a policy equilibrium π, let S_π be the set of all stationary distributions which satisfy the equilibrium inequalities (4.1), $S_\pi := \{\sigma \times \sigma_a \times \sigma_b : (a, b) \in \mathcal{A}\}$. Define $S = \bigcup_{\pi \in \Pi} S_\pi$, where Π is the set of all policy equilibria.

Our main result states that the sequence of empirical product distributions formed after T rounds of the game $\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b$ is going to converge to S. Here $\widehat{\sigma}_a$ and $\widehat{\sigma}_b$ denote the distributions of deviation in play, when player 1 switches to the fixed action $a \in \mathcal{A}_1$ and player 2 switches to the fixed action $b \in \mathcal{A}_2$ respectively. We now define these distributions formally.

Definition 4.4.6. Suppose player 1 is playing an algorithm with output at time t given by $f_t : \mathcal{A}_2^t \to \Delta\mathcal{A}_1$ i.e. $p_t^1 = f_t(b_{0:t-1})$. Similarly, suppose player 2 is playing an algorithm with output at time t given by $p_t^2 = g_t(a_{0:t-1})$. The empirical distribution at time T is $\widehat{\sigma} := \frac{1}{T}\sum_{t=1}^{T} p_t$, where $p_t = p_t^1 \times p_t^2$ is the product distribution over $\mathcal{A}$ at time t. Further let $(p_a^2)_t = g_t(a_{0:t-m}, a, \ldots, a)$ denote the distribution at time t, provided that player 1 switched their strategy to the constant action $a \in \mathcal{A}_1$. Let δ_a denote the distribution over $\mathcal{A}_1$ which puts all the probability mass on action a. Let $(p_a)_t = \delta_a \times (p_a^2)_t$ be the product distribution over $\mathcal{A}$, corresponding to the change of play at time t. Denote by $\widehat{\sigma}_a = \frac{1}{T}\sum_{t=1}^{T}(p_a)_t$ the empirical distribution corresponding to the change of play. The distribution $\widehat{\sigma}_b$ is defined similarly.

Suppose that f_t and g_t are no-policy regret algorithms, then our main result states that the sequence $(\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b)_T$ converges to the set S.

Theorem 4.4.2. *If the algorithms played by player* 1 *in the form of* f_t *and player* 2 *in the form of* g_t *give sub-linear policy regret sequences, then the sequence of product distributions* $(\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b)_{T=1}^{\infty}$ *converges weakly to the set* S.

In particular if both players are playing MWU or Exp3, we know that they will have sublinear policy regret. Not surprisingly, we can show something slightly stronger as well. Let $\tilde{\sigma}$, $\tilde{\sigma}_a$ and $\tilde{\sigma}_b$ denote the empirical distributions of observed play corresponding to $\widehat{\sigma}$, $\widehat{\sigma}_a$ and $\widehat{\sigma}_b$, i.e. $\tilde{\sigma} = \frac{1}{T}\delta_t$, where δ_t denotes the Dirac distribution, putting all weight on the played actions at time t. Then these empirical distributions also converge to S almost surely.

Corollary 4.4.3. *The sequence of product distributions* $(\tilde{\sigma} \times \tilde{\sigma}_a \times \tilde{\sigma}_b)_{T=1}^{\infty}$ *converges weakly to the set* S *almost surely.*

We would like to emphasize that the convergence guarantee of Theorem 4.4.2 does not rely on there being a unique stationary distribution of the empirical Markov chains $\widehat{\mathrm{M}}$, $\widehat{\mathrm{M}}_a$ and $\widehat{\mathrm{M}}_b$ or their respective limits $\mathrm{M}, \mathrm{M}_a, \mathrm{M}_b$. Indeed, Theorem 4.4.2 shows that any limit point of $\{(\widehat{\sigma}, \widehat{\sigma}_a, \widehat{\sigma}_b)_T\}_{T=1}^{\infty}$ satisfies the conditions of Definition 4.4.4. The proof does not require that any of the respective Markov chains have a unique stationary distribution, but rather requires only that $\widehat{\sigma}$ has sublinear policy regret. We would also like to remark that $\{(\widehat{\sigma}, \widehat{\sigma}_a, \widehat{\sigma}_b)_T\}_{T=1}^{\infty}$ need not have a unique limit and our convergence result only guarantees that the sequence is going to the set S. This is standard when showing that any type of no regret play converges to an equilibrium, see for example Stoltz and Lugosi (2007).

4.4.2 Proof sketch for Theorem 4.4.2

The proof of Theorem 4.4.2 has three main steps. The first step defines the natural empirical Markov chains $\widehat{\mathrm{M}}$, $\widehat{\mathrm{M}}_a$ and $\widehat{\mathrm{M}}_b$ from the empirical play $(p_t)_{t=1}^{t}$ and shows that the empirical distributions $\widehat{\sigma}$, $\widehat{\sigma}_a$ and $\widehat{\sigma}_b$ are stationary distributions of the respective Markov chains. The next step is to show that the empirical Markov chains converge to Markov chains M, M_a and M_b induced by some distribution π over $\mathcal{F}$. The final step is to show that π is a policy equilibrium.

We begin with the definition of the empirical Markov chains.

Definition 4.4.7. Let the empirical Markov chain be $\widehat{\mathrm{M}}$, with $\widehat{\mathrm{M}}_{i,j} = \frac{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)p_t(x_j)}{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)}$ if $\frac{1}{T}\sum_{t=1}^{T} p_t(x_i) \neq 0$ and 0 otherwise, where p_t is defined in 4.4.6. For any fixed $a \in \mathcal{A}_1$, let the empirical Markov chain corresponding to the deviation in play of player 1 be $\widehat{\mathrm{M}}_a$, with $(\widehat{\mathrm{M}}_a)_{i,j} = \frac{\frac{1}{T}\sum_{t=1}^{T}(p_a)_t(x_i)(p_a)_t(x_j)}{\frac{1}{T}\sum_{t=1}^{T}(p_a)_t(x_i)}$, if $\frac{1}{T}\sum_{t=1}^{T}(p_a)_t(x_i) \neq 0$ and 0 otherwise, where $(p_a)_t$ is defined in 4.4.6. The Markov chain $\widehat{\mathrm{M}}_b$ is defined similarly for any $b \in \mathcal{A}_2$.

The intuition behind constructing these Markov chains is as follows – if we were only provided with the observed empirical play $(x_t)_{t=1}^{T} = (a_t, b_t)_{t=1}^{T}$ and someone told us that the x_t's were coming from a Markov chain, we could try to build an estimator of the Markov chain by approximating each of the transition probabilities. In particular the estimator of transition from state i to state j is given by $\tilde{\mathrm{M}}_{i,j} = \frac{\frac{1}{T}\sum_{t=1}^{T}\delta_{t-1}(x_i)\delta_t(x_j)}{\frac{1}{T}\sum_{t=1}^{T}\delta_t(x_i)}$, where $\delta_t(x_i) = 1$ if x_i occurred at time t and 0 otherwise. When the players are playing according to a no-regret algorithm i.e. at time t, x_t is sampled from p_t, it is possible to show that $\tilde{\mathrm{M}}_{i,j}$ concentrates to $\widehat{\mathrm{M}}_{i,j}$ (see section 4.5.1). Not only does $\widehat{\mathrm{M}}$ arise naturally, but it turns out that the empirical distribution $\widehat{\sigma}$ defined in 4.4.6 is also a stationary distribution of $\widehat{\mathrm{M}}$.

Lemma 4.4.4. *The distribution of play* $\widehat{\sigma} = \frac{1}{T}\sum_{t=1}^{T} p_t$ *is a stationary distribution of* $\widehat{\mathrm{M}}$. *Similarly the distributions* $\tilde{\sigma}, \widehat{\sigma}_a, \tilde{\sigma}_a, \widehat{\sigma}_b$ *and* $\tilde{\sigma}_b$ *are stationary distributions of the Markov chains* $\tilde{\mathrm{M}}, \widehat{\mathrm{M}}_a, \tilde{\mathrm{M}}_a, \widehat{\mathrm{M}}_b$ *and* $\tilde{\mathrm{M}}_b$ *respectively.*

Proof. We show the result for $\widehat{\sigma}$ and $\widehat{\mathrm{M}}$. The rest of the results can then be derived in the same way.

$$(\widehat{\sigma}^{\top}\widehat{\mathrm{M}})_j = \sum_{i=1}^{|\mathcal{A}|}\left(\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)\right)\frac{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)p_t(x_j)}{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)} = \frac{1}{T}\sum_{t=1}^{T} p_t(x_j)\sum_{i=1}^{|\mathcal{A}|} p_t(x_i) = \frac{1}{T}\sum_{t=1}^{T} p_t(x_j),$$

where the first equality holds because the i-th entry of the vector $\widehat{\sigma}$ is exactly $\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)$ and the (i,j)-th entry of $\widehat{\mathrm{M}}$ by definition is $\frac{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)p_t(x_j)}{\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)}$, and the last equality holds because p_t is a distribution over actions so $\sum_{i=1}^{|\mathcal{A}|} p_t(x_i) = 1$. □

Suppose that both players are playing MWU for T rounds. Then Lemma 4.4.4 together with Theorems 4.2.1 and the stability of MWU imply that $\mathbb{E}_{(a,b)\sim\widehat{\sigma}}[u_1(a,b)] \geq \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a}[u_1(a,b)] - O(m/\sqrt{T})$. A similar inequality holds for player 2 and $\widehat{\sigma}_b$. As $T \to \infty$, the inequality above becomes similar to (4.1). This will play a crucial role in the proof of our convergence result, which shows that $\widehat{\sigma}, \widehat{\sigma}_a$ and $\widehat{\sigma}_b$ converge to the set of policy equilibria. We would also like to guarantee that the empirical distributions of observed play $\tilde{\sigma}, \tilde{\sigma}_a$ and $\tilde{\sigma}_b$ also converge to this set. To show this second result, we are going to proof that $\tilde{\sigma}$ approaches $\widehat{\sigma}$ almost surely as T goes to infinity.

Lemma 4.4.5. *Let* $\widehat{\sigma} = \frac{1}{T}\sum_{t=1}^{T} p_t$ *be the empirical distribution after* T *rounds of the game and let* $\tilde{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\delta_t$ *be the empirical distribution of observed play. Then* $\limsup_{T\to\infty}\|\tilde{\sigma} - \widehat{\sigma}\|_1 = 0$ *almost surely. Similarly, for the distributions corresponding to deviation in play we have* $\limsup_{T\to\infty}\|\tilde{\sigma}_a - \widehat{\sigma}_a\|_1 = 0$ *and* $\limsup_{T\to\infty}\|\tilde{\sigma}_b - \widehat{\sigma}_b\|_1 = 0$ *almost surely.*

Our next step is to show that the empirical Markov chains $\widehat{\mathrm{M}}$ converge to a Markov chain M induced by some distribution π over the functional space $\mathcal{F}$. We do so by constructing a sequence of empirical distributions $\widehat{\pi}$ over $\mathcal{F}$, based on the players' strategies, which induce $\widehat{\mathrm{M}}$. We can then consider every convergent subsequence of $(\widehat{\pi})_{T_\ell=1}^{\infty}$ with limit point π and argue that the corresponding sequence $(\widehat{\mathrm{M}})_{T_\ell=1}^{\infty}$ of Markov chains converges to the Markov chain induced by π.

Definition 4.4.8. Let $\widehat{\pi}$ be the distribution over $\mathcal{F}$, such that the probability to sample any fixed $f : \mathcal{A}_2 \to \mathcal{A}_1$ and $g : \mathcal{A}_1 \to \mathcal{A}_2$ is $\widehat{\pi}(f,g) = \prod_{i \in |\mathcal{A}|} \frac{\sum_t p_t(x_i) p_t(y_i)}{\sum_{t=1} p_t(x_i)}$, where $x_i = (a_i, b_i)$ and $y_i = (f(b_i), g(a_i))$. Similarly, let $\widehat{\pi}_a$ and $\widehat{\pi}_b$ be the distributions over $\mathcal{F}$ constructed as above but by using the empirical distribution of deviated play induced by player 1 deviating to action $a \in \mathcal{A}_1$ and player 2 deviating to action $b \in \mathcal{A}_2$.

The next lemma checks that $\widehat{\pi}$ is really a probability distribution.

Lemma 4.4.6. *The functionals $\widehat{\pi}, \widehat{\pi}_a$ and $\widehat{\pi}_b$ are all probability distributions.*

Proof. Consider the space of all transition events for a fixed (a,b) pair i.e. $\mathcal{S}_{(a,b)} = \{((a',b') \times (a,b)) : (a',b') \in \mathcal{A}\}$. There is an inherent probability measure on this set, given by $\mathbb{P}\left[(a',b') \times (a,b)\right] = \frac{\sum_t p_t(a,b) p_t(a',b')}{\sum_t p_t(a,b)}$. It is easy to see that this is a probability measure, since the measure of the whole set is exactly

$$\sum_{(a',b') \in \mathcal{A}} \frac{\sum_t p_t(a,b) p_t(a',b')}{\sum_t p_t(a,b)} = \frac{\sum_t p_t(a,b) \sum_{(a',b') \in \mathcal{A}} p_t(a',b')}{\sum_t p_t(a,b)} = 1.$$

The set of all $\mathcal{F}$ can exactly be thought of as $\times_{(a,b) \in \mathcal{A}} \mathcal{S}_{(a,b)}$ and the function $\widehat{\pi}$ defined in 4.4.8 is precisely the product measure on that set. Similar arguments show that $\widehat{\pi}_a$ and $\widehat{\pi}_b$ are probability distributions. □

The proof of the above lemma reveals something interesting about the construction of $\widehat{\pi}$. Fix the actions $(a,b) \in \mathcal{A}$. Then the probability to sample a function pair (f,g) which map (a,b) to (a',b') i.e. $a' = f(a)$ and $b' = f(b)$ is precisely equal to the entry $\widehat{\mathrm{M}}_{(a,b),(a',b')}$ of the empirical Markov chain. Since every function pair $(f,g) \in \mathcal{F}$ is determined by the way $\mathcal{A}$ is mapped, and we have already have a probability distribution for a fixed mapping (a,b) to (a',b'), we can just extend this to $\widehat{\pi}$ by taking the product distribution over all pairs $(a, b) \in \mathcal{A}$. This construction gives us exactly that $\widehat{\mathrm{M}}$ is induced by $\widehat{\pi}$.

Lemma 4.4.7. *Let $\widehat{\mathrm{M}}$, $\widehat{\mathrm{M}}_a$ and $\widehat{\mathrm{M}}_b$ be the empirical Markov chains defined in 4.4.7, then the induced Markov chain from $\widehat{\pi}$ is exactly $\widehat{\mathrm{M}}$ and the induced Markov chains from $\widehat{\pi}_a$ and $\widehat{\pi}_b$ are exactly $\widehat{\mathrm{M}}_a$ and $\widehat{\mathrm{M}}_b$.*

Proof sketch. The proof is by direct computation. □

The last step of the proof is to show that any limit point π of $(\widehat{\pi})_T$ is necessarily a policy equilibrium. This is done through an argument by contradiction. In particular we assume that a limit point π is not a policy equilibrium. The limit point π induces a Markov chain M, which we can show is the limit point of the corresponding subsequence of $(\widehat{\mathrm{M}})_T$ by using lemma 4.4.7. Since π is not a policy equilibrium, no stationary distribution of M can satisfy the inequalities (4.1). We can now show that the subsequence of $(\widehat{\sigma})_T$ which are stationary distributions of the corresponding $\widehat{\mathrm{M}}$'s, converges to a stationary distribution of M. This, however, is a contradiction because of the next theorem.

Theorem 4.4.8. *Let P be the set of all product distributions $\sigma \times \sigma_a \times \sigma_b$ which satisfy the inequalities in Equation 4.1:*

$$\mathbb{E}_{(a,b)\sim\sigma}\left[u_1(a,b)\right] \geq \mathbb{E}_{(a,b)\sim\sigma_a}\left[u_1(a,b)\right]$$
$$\mathbb{E}_{(a,b)\sim\sigma}\left[u_2(a,b)\right] \geq \mathbb{E}_{(a,b)\sim\sigma_b}\left[u_2(a,b)\right].$$

Let $\widehat{\sigma}^T$ be the empirical distribution of play after T rounds and let $\widehat{\sigma}_a^T$ be the empirical distribution when player 1 switches to playing action a and define $\widehat{\sigma}_b^T$ similarly for player 2. Then the product distribution $\widehat{\sigma}^T \times \widehat{\sigma}_a^T \times \widehat{\sigma}_b^T$ converges to weakly to the set P.

Proof sketch. The proof is by contradiction. We assume that there is a convergent subsequence of $(\widehat{\sigma}^T \times \widehat{\sigma}_a^T \times \widehat{\sigma}_b^T)$. The existence of such a subsequence implies that one of the inequalities in Equation 4.1 is violated which is a contradiction. □

We now sketch how the rest of proof of Theorem 4.4.2. The proof again goes by contradiction. Assume π is not a policy equilibrium, this implies that no stationary distribution of M and corresponding stationary distributions of M_a and M_b can satisfy inequalities (4.1). Since the empirical distributions $\widehat{\sigma}$, $\widehat{\sigma}_a$ and $\widehat{\sigma}_b$ of the play satisfies inequalities (4.1) up to an $o(1)$ additive factor, we can show, in Theorem 4.4.8, that in the limit, the policy equilibrium inequalities are exactly satisfied. Combined with the convergence of $\widehat{\mathrm{M}}$, $\widehat{\mathrm{M}}_a$ and $\widehat{\mathrm{M}}_b$ to M, M_a and M_b, respectively, this implies that stationary distributions of M, M_a and M_b, satisfying Equation 4.1, giving a contradiction.

4.4.3 Relation of policy equlibria to CCEs

So far we have defined a new class of equilibria and shown that they correspond to no policy regret play. Furthermore, we know that if both players in a 2-player game play stable no external regret algorithms, then their play also has sublinear policy regret. It is natural to ask if every CCE is also a policy equilibrium: if σ is a CCE, is there a corresponding policy equilibrium π which induces a Markov chain M for which σ is a stationary distribution satisfying (4.1)? We show that the answer to this question is positive:

Theorem 4.4.9. *For any CCE σ of a 2-player game $\mathcal{G}$, there exists a policy-equilibrium π which induces a Markov chain M with stationary distribution σ.*

Proof sketch. To prove this, we show that for any CCE we can construct stable no-external regret algorithm which converge to it, and so since stable no-external regret algorithms always converge to policy equilibria (Theorem 4.2.1), this implies the CCE is also a policy equilibrium. □

However, we show the converse is not true: policy equilibria can give rise to behavior which is not a CCE. Our proof appeals to a utility sequence which is similar in spirit to the one in Theorem 4.1.1, but is adapted to the game setting.

Theorem 4.4.10. *There exists a 2-player game $\mathcal{G}$ and product distributions $\sigma \times \sigma_a \times \sigma_b \in S$ (where S is defined in Definition 4.4.5 as the possible distributions of play from policy equilibria), such that σ is not a CCE of $\mathcal{G}$.*

In the next section, Section 4.4.4, we give a simple example of a policy equilibrium which is not a CCE.

4.4.4 Simple example of a policy equilibrium

We now present a simple 2-player game with strategies of the players which lead to a policy equilibrium, which in fact is not a CCE. Further these strategies give the asymptotically maximum utility for both row and column players over repeated play of the game. The idea behind the construction is very similar to the one showing incompatibility of policy regret and external regret. The utility matrix for the game is given in Table 4-I. Since the column player has the same payoff for all his actions they will always have no policy and no external regret. The strategy the column player chooses is to always play the function $f : \mathcal{A}_1 \to \mathcal{A}_2$:

Table 4-I: Utility matrix

Player 1\Player 2	**c**	**d**
a	(3/4,1)	(0,1)
b	(1,1)	(0,1)

$$f(x) = \begin{cases} c & x = a \\ d & x = b. \end{cases}$$

In the view of the row player, this strategy corresponds to playing against an adversary which plays the familiar utility functions:

$$u_t(a_{t-1}, a_t) = \begin{cases} 1 & a_{t-1} = a, a_t = b \\ \frac{3}{4} & a_{t-1} = a_t = a \\ 0 & \text{otherwise.} \end{cases}$$

We have already observed that on these utilities, the row player can have either no policy regret or no external regret but not both. What is more the utility of no policy regret play is higher than the utility of any of the no external regret strategies. This already implies that the row player is better off playing according to the no policy regret strategy which consists of always playing the fixed function $g : \mathcal{A}_2 \to \mathcal{A}_1$ given by $g(x) = a$. Below we present the policy equilibrium $\pi \in \Delta\mathcal{F}$, corresponding Markov chain $\mathrm{M} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ and its stationary distribution $\sigma \in \Delta\mathcal{A}$ satisfying the no policy regret requirement.

$$\pi(\tilde{f}, \tilde{g}) = \delta_{(f,g)}, \mathrm{M} = \begin{array}{c} \\ (a,c) \\ (a,d) \\ (b,c) \\ (b,d) \end{array} \begin{array}{c} \begin{array}{cccc} (a,c) & (a,d) & (b,c) & (b,d) \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}, \sigma(x, y) = \delta_{(a,c)}.$$

Suppose the row player was playing any no policy regret strategy, for example one coming from a no policy regret algorithm, as a response to the observed utilities $u_t(\cdot, \cdot)$. Since the only sublinear policy regret play for these utilities is to only deviate from playing a a sublinear number of times we see that the empirical distribution of play for the row player converges to the dirac distribution δ_a. Together with the strategy of the column player, this implies the column player chooses the action d only a sublinear number of times and thus their empirical distribution of play converges to δ_c. It now follows that the empirical distribution of play converges to $\delta_a \times \delta_c = \delta_{(a,c)} \in \Delta\mathcal{A}$. We can similarly verify that the empirical Markov chain will converge to M and the empirical functional distribution $\widehat{\pi}$ converges to π. Theorem 4.4.2 guarantees that because both players incur only sublinear regret π is a policy equilibrium. It should also be intuitively clear why this is the case without the theorem – suppose that the row player switches to playing the fixed action b. The resulting functional distribution, Markov chain and stationary distributions become:

$$\pi_b(\tilde{f}, \tilde{g}) = \delta_{(f, \widehat{g} \equiv b)}, \mathrm{M}_b = \begin{array}{c|cccc} & (a,c) & (a,d) & (b,c) & (b,d) \\ \hline (a,c) & 0 & 0 & 1 & 0 \\ (a,d) & 0 & 0 & 1 & 0 \\ (b,c) & 0 & 0 & 0 & 1 \\ (b,d) & 0 & 0 & 0 & 1 \end{array}, \sigma_b(x,y) = \delta_{(b,d)}.$$

The resulting utility for the row player is now 0, compared to the utility gained from playing according to π, which is 3/4.

4.5 Detailed proofs from Section 4.4

Proof of Lemma 4.4.7. Consider $\widehat{\mathrm{M}}_{(a,b),(a',b')} = \frac{\sum_{t=1}^{T} p_t(a',b') p_t(a,b)}{\sum_{t=1}^{T} p_t(a,b)}$. The transition probability induced by $\widehat{\pi}$ is exactly

$$\begin{aligned} \mathbb{P}\left[(a',b') | (a,b)\right] &= \sum_{(f,g):(f(b),g(a))=(a',b')} \widehat{\pi}(f,g) = \sum_{(f,g):(f(b),g(a))=(a',b')} \prod_{i \in [|\mathcal{A}|]} \frac{\sum_t p_t(x_i) p_t(y_i)}{\sum_{t=1} p_t(x_i)} \\ &= \sum_{(f,g):(f(b),g(a))=(a',b')} \frac{\sum_t p_t(a,b) p_t(a',b')}{\sum_{t=1} p_t(a,b)} \prod_{i \in [|\mathcal{A}|], (x_i,y_i) \neq ((a,b),(a',b')} \frac{\sum_t p_t(x_i) p_t(y_i)}{\sum_{t=1} p_t(x_i)} \\ &= \frac{\sum_t p_t(a,b) p_t(a',b')}{\sum_{t=1} p_t(a,b)} \sum_{(f,g):(f(b),g(a))=(a',b')} \prod_{i \in [|\mathcal{A}|], (x_i,y_i) \neq ((a,b),(a',b')} \frac{\sum_t p_t(x_i) p_t(y_i)}{\sum_{t=1} p_t(x_i)} \\ &= \frac{\sum_t p_t(a,b) p_t(a',b')}{\sum_{t=1} p_t(a,b)}, \end{aligned}$$

where the last equality holds, because for fixed (f,g) with $x_i = (a_i, b_i)$ and $y_i = (f(b_i), g(a_i))$, the product $\prod_{i \in [|\mathcal{A}|], (x_i,y_i) \neq ((a,b),(a',b')} \frac{\sum_t p_t(x_i) p_t(y_i)}{\sum_{t=1} p_t(x_i)}$ is exactly the conditional probability $\widehat{\pi}((f,g) | (f(b), g(a)) = (a',b'))$. The result for $\widehat{\pi}_a$ and $\widehat{\pi}_b$ is shown similarly.

□

Proof of Theorem 4.4.8. Theorem 4.4.8 follows from the fact that convergence in the Prokhorov metric implies weak convergence. First notice that by Prokhorov's Theorem $\mathcal{P}(\mathcal{A})$ is a compact metric space with the Prokhorov metric. Thus by Tychonoff's Theorem the product space $\mathcal{P}(\mathcal{A})^3$ is compact in the maximum metric. Suppose for a contradiction that the sequence $(\widehat{\sigma}^T \times \widehat{\sigma}_a^T \times \widehat{\sigma}_b^T)_T$ does not converge to the set S.

This implies that there exists some subsequence $(\widehat{\sigma}^k \times \widehat{\sigma}_a^k \times \widehat{\sigma}_b^k)_k$, converging to some $\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b \notin S$. If $\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b \notin S$, then either $\mathbb{E}_{(a,b)\sim\sigma}\left[u_1(a,b)\right] < \mathbb{E}_{(a,b)\sim\sigma_a}\left[u_1(a,b)\right]$ or $\mathbb{E}_{(a,b)\sim\sigma}\left[u_2(a,b)\right] < \mathbb{E}_{(a,b)\sim\sigma_b}\left[u_2(a,b)\right]$. WLOG suppose the first inequality holds. From our assumption, the continuity of u_1 and the definition of the maximum metric we have $\lim_{k\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}^k}\left[u_1(a,b)\right] = \mathbb{E}_{(a,b)\sim\widehat{\sigma}}\left[u_1(a,b)\right]$ and $\lim_{k\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^k}\left[u_1(a,b)\right] = \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a}\left[u_1(a,b)\right]$. Notice that by the fact $\widehat{\sigma}_a^k$ is the average empirical distribution if player 1 changed its play to the fixed action $a \in \mathcal{A}_1$ and $\widehat{\sigma}^k$ being the average empirical distribution it holds that $\mathbb{E}_{(a,b)\sim\widehat{\sigma}^k}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^k}\left[u_1(a,b)\right] \geq -o(1)$ and thus $\lim_{k\to\infty}\left[\mathbb{E}_{(a,b)\sim\widehat{\sigma}^k}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^k}\left[u_1(a,b)\right]\right] \geq 0$. The above implies:

$$\begin{aligned}
0 &\leq \lim_{k\to\infty}\left[\mathbb{E}_{(a,b)\sim\widehat{\sigma}^k}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^k}\left[u_1(a,b)\right]\right] \\
&= \lim_{k\to\infty}\mathbb{E}_{(a,b)\sim\widehat{\sigma}^k}\left[u_1(a,b)\right] - \lim_{k\to\infty}\mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^k}\left[u_1(a,b)\right] \\
&= \mathbb{E}_{(a,b)\sim\widehat{\sigma}}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a}\left[u_1(a,b)\right] < 0,
\end{aligned}$$

which is a contradiction. Since $\mathcal{A} \times \mathcal{A} \times \mathcal{A}$ is separable then convergence in the Prokhorov metric in $\mathcal{P}(\mathcal{A} \times \mathcal{A} \times \mathcal{A})$ is equivalent to weak convergence. Again we can argue by contradiction – if we assume that $(\widehat{\sigma}^T \times \widehat{\sigma}_a^T \times \widehat{\sigma}_b^T)_T$ doesn't converge to the set S in the Prokhorov metric, then there exists some subsequence $(\widehat{\sigma}^k \times \widehat{\sigma}_a^k \times \widehat{\sigma}_b^k)_k$ which converges to some $\mu \in \mathcal{P}(\mathcal{A}^3)$ such that $\mu \notin S$. First we argue that μ must be a product measure i.e. $\mu = (\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b)$. Let $(\widehat{\sigma}^j \times \widehat{\sigma}_a^j \times \widehat{\sigma}_b^j)_j$ be a convergence subsequence of $(\widehat{\sigma}^k \times \widehat{\sigma}_a^k \times \widehat{\sigma}_b^k)_k$ in $\mathcal{P}(\mathcal{A})^3$, with limit $(\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b)$, then each of $\widehat{\sigma}^j$, $\widehat{\sigma}_a^j$ and $\widehat{\sigma}_b^j$ converge weakly to $\widehat{\sigma}$, $\widehat{\sigma}_a$ and $\widehat{\sigma}_b$ respectively and thus $(\widehat{\sigma}^j \times \widehat{\sigma}_a^j \times \widehat{\sigma}_b^j)_j$ converges weakly to $\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b$ and thus it converges in the Prokhorov metric of $\mathcal{P}(\mathcal{A}^3)$. This implies that $(\widehat{\sigma}^k \times \widehat{\sigma}_a^k \times \widehat{\sigma}_b^k)_k$ also converges weakly to $\widehat{\sigma} \times \widehat{\sigma}_a \times \widehat{\sigma}_b$ and so μ is a product measure. Again since $\mu \notin S$, assume WLOG $\mathbb{E}_{(a,b)\sim\widehat{\sigma}}\left[u_1(a,b)\right] < \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a}\left[u_1(a,b)\right]$. Define $f : \mathcal{A}^3 \to \mathbb{R}$, $f(a,b,c,d,e,f) = u_1(a,b) - u_1(c,d)$. f is continuous and from the no-policy regret of the pair $\widehat{\sigma}^k, \widehat{\sigma}_a^k$ we have:

$$0 \leq \lim_{k\to\infty}\mathbb{E}_{(a,b,c,d,e,f)\sim(\widehat{\sigma}^k\times\widehat{\sigma}_a^k\times\widehat{\sigma}_b^k)_k}\left[f(a,b,c,d,e,f)\right] = \mathbb{E}_{(a,b,c,d,e,f)\sim\mu}\left[f(a,b,c,d,e,f)\right] < 0,$$

which is again a contradiction. □

Proof of Theorem 4.4.2. We consider the sequence of empirical distributions $\widehat{\pi}^T$ defined in 4.4.8, over the functional space $\mathcal{F}_1 \times \mathcal{F}_2$ and show that this sequence must converge to the set of all policy equilibria Π in the Prokhorov metric on $\mathcal{P}(\mathcal{F}_1 \times \mathcal{F}_2)$. First, notice that since the functions $f : \mathcal{A}_2 \to \mathcal{A}_1$ are from finite sets of actions to finite sets of actions, we can consider the set $\mathcal{F}_1$ as a subset of a finite dimensional vector space, with the underlying field of real numbers and the metric induced by the l_1 norm. Similarly, we can also equip $\mathcal{F}_2$ with the l_1 norm. Since both $\mathcal{F}_1$ and $\mathcal{F}_2$ are closed sets with respect to this metric and they are clearly bounded, they are compact. Thus the set $\mathcal{F}_1 \times \mathcal{F}_2$ is a compact set with the underlying metric d being the maximum metric. By Prokhorov's Theorem we know that $\mathcal{P}(\mathcal{F}_1 \times \mathcal{F}_2)$ is a compact metric space with the Prokhorov metric. Suppose that the sequence $(\widehat{\pi}^T)_T$ does not converge to Π. This implies that there is some convergent subsequence $(\widehat{\pi}^t)_t$ with a limit π outside of Π. Let M be the Markov chain induced by π and let $\widehat{\mathrm{M}}^T$ be the Markov chain induced by $\widehat{\pi}^T$.

First we show that $\lim_{t\to\infty} \|\widehat{\mathrm{M}}^t - \mathrm{M}\|_1 = 0$. Recall that $\mathrm{M}_{(a,b),(a',b')} = \sum_{(f,g):f(b)=a',g(a)=b'} \pi(f,g)$ and that by lemma 4.4.7 $\widehat{\mathrm{M}}^t_{(a,b),(a',b')} = \sum_{(f,g):f(b)=a',g(a)=b'} \widehat{\pi}^t(f,g)$. Notice that f, g are continuous functions on $\mathcal{F}_1$ and $\mathcal{F}_2$, since the topology induced by the l_1 metric on both sets is exactly the the discrete topology and every function from a topological space equipped with the discrete topology is continuous. Since convergence in the Prokhorov metric implies weak convergence, we have that for any fixed f, g, $\lim_{t\to\infty} \widehat{\pi}^t(f,g) = \pi(f,g)$. Since the sum $\sum_{(f,g):f(b)=a',g(a)=b'} \widehat{\pi}^t(f,g)$ is finite this implies that $\lim_{t\to\infty} \sum_{(f,g):f(b)=a',g(a)=b'} \widehat{\pi}^t(f,g) = \sum_{(f,g):f(b)=a',g(a)=b'} \pi(f,g)$ and so $\lim_{t\to\infty} \|\widehat{\mathrm{M}}^t - \mathrm{M}\|_1 = 0$.

Next we show that any convergent subsequence $(\widehat{\sigma}^k)_k$ of $(\widehat{\sigma}^t)_t$ in the Prokhorov metric, converges to a stationary distribution σ of M. First notice that $(\widehat{\sigma}^k)_k$ exists, since $\mathcal{P}(\mathcal{A})$ is compact. Next, suppose σ is the limit of $(\widehat{\sigma}^k)_k$ in the Prokhorov metric. This implies that $\lim_{k\to\infty} \widehat{\sigma}^k(a,b) = \sigma(a,b)$, in particular if we consider $\mathcal{A} \subset \mathbb{R}^{|\mathcal{A}|}$ and $\widehat{\sigma}^k, \sigma \in \mathbb{R}^{|\mathcal{A}|}$ as vectors, then the above implies that $\lim_{k\to\infty} \|\sigma - \widehat{\sigma}^k\|_1 = 0$. Next we construct the following sequence $(\sigma_{k_n})_{k_n}$ of stationary distributions of M – choose k_n large enough, so that $\|\mathrm{M} - \widehat{\mathrm{M}}^{k_n}\| \leq \frac{1}{n}$. Such a k_n exists, because $(\widehat{\mathrm{M}}^k)_k$ is a subsequence of $(\widehat{\mathrm{M}}^t)_t$ which converges to M. By lemma 4.4.4, there exists a stationary distribution σ_{k_n} of M such that $\|\widehat{\sigma}_{k_n} - \sigma_{k_n}\|_1 \leq \frac{c}{n}$, for some constant c. We show that σ_{k_n} converges to σ. Fix some $\epsilon > 0$, we find an N, such that for any $n \geq N$ we have $\|\sigma_{k_n} - \sigma\|_1 < \epsilon$. Notice that $\|\sigma_{k_n} - \sigma\|_1 \leq \|\sigma_{k_n} - \widehat{\sigma}_{k_n}\|_1 + \|\widehat{\sigma}_{k_n} - \sigma\|_1$. Since $\|\sigma_{k_n} - \widehat{\sigma}_{k_n}\|_1 \leq \frac{c}{n}$ and by convergence, we know that for $\frac{\epsilon}{2}$, there exists N' such that for any $n \geq N'$, $\|\widehat{\sigma}_{k_n} - \sigma\|_1 < \frac{\epsilon}{2}$, we can set $N = \max\left(\frac{2}{c\epsilon}, N_1\right)$. Suppose, for a contradiction, that σ is not a stationary distribution of M. Then there exists some ϵ such that $\|\sigma^\top \mathrm{M} - \sigma\|_2 > \epsilon$. This implies:

$$\epsilon < \|\sigma^\top \mathrm{M} - \sigma\|_2 \leq \|\sigma^\top \mathrm{M} - \sigma_{k_n}^\top \mathrm{M}\|_2 + \|\sigma_{k_n}^\top \mathrm{M} - \sigma\|_2 < 2\|\sigma_{k_n} - \sigma\|_2,$$

where the last inequality holds by the fact $\sigma - \sigma_{k_n}$ is not a stationary distribution of M and thus M can only shrink the difference as a stochastic matrix. The inequality $2\|\sigma_{k_n} - \sigma\|_2 > \epsilon$ is a contradiction since we know that σ_{k_n} converges to σ and thus σ is a stationary distribution of M. Since strong convergence, implies weak convergence, which in hand implies convergence in the Prokhorov metric for separable metric spaces, we have shown that every convergent subsequence of $(\widehat{\sigma}_t)_t$ converges to a stationary distribution of M in the Prokhorov metric.

Next, we show that $(\widehat{\pi}_a^t)_t$ converges to π_a. By assumption $(\widehat{\pi}^t)_t$ converges weakly to π. Since we are are in a finite dimensional space, we also have strong convergence. In particular, for any $g \in \mathcal{F}_2$, we have

$$\lim_{t\to\infty} \sum_{f\in\mathcal{F}_1} \widehat{\pi}^t(f,g) = \sum_{f\in\mathcal{F}_1} \lim_{t\to\infty} \widehat{\pi}^t(f,g) = \sum_{f\in\mathcal{F}_1} \pi(f,g)$$

and so the sequence of marginal distribution also converges to the respective marginal of π. Since $\widehat{\pi}^t_a$ is exactly the product distribution of the dirac distribution over $\mathcal{F}_1$ putting all weight on the constant function mapping everything to the fixed action a and the marginal of $\widehat{\pi}^t$ over $\mathcal{F}_2$, by the convergence of marginals we conclude that $(\widehat{\pi}^t_a)_t$ converges to π_a in the strong sense and thus in the Prokhorov metric. In the same way we can show that $(\widehat{\pi}^t_b)_t$ converges to π_b.

With a similar argument as for $(\widehat{\sigma}^t)$ we show that every convergent subsequence of $(\widehat{\sigma}^t_a)_t$ converges to a stationary distribution of M_a and any convergent subsequence of $(\widehat{\sigma}^t_b)_t$ converges to a stationary distribution of M_b. Because of the construction in Theorem 4.4.1 and the convergence of $\widehat{\pi}_a$ to π_a, we can guarantee that $(\widehat{\sigma}^t_a)_t$ converges precisely to σ_a:

$$\begin{aligned}\sigma_a(a,b) &= \sum_{f,g:g(a)=b} \pi_a(f,g) = \sum_{f,g:g(a)=b} \lim_{t\to\infty} \widehat{\pi}^t_a(f,g)\\ &= \lim_{t\to\infty} \sum_{f,g:g(a)=b} \widehat{\pi}^t_a(f,g) = \lim_{t\to\infty} \widehat{\sigma}^t_a(a,b).\end{aligned}$$

Similarly $\widehat{\sigma}^t_b$ converges to $\widehat{\sigma}_b$. However, we assumed that π is not a policy equilibrium and thus no stationary distributions of M, M_a and M_b can satisfy the policy equilibrium inequalities. We now arrive at a contradiction since by Theorem 4.4.8 and the above, we have that any limit point of $(\widehat{\sigma}^t)_t$ and the corresponding distributions for fixed actions a and b are stationary distributions of M, M_a and M_b, respectively, which satisfy the policy equilibrium inequalities. □

4.5.1 Concentration of the estimated Markov chain

Lemma 4.5.1. *With probability at least* $1-|\mathcal{A}|6\exp\left(-\frac{T\epsilon^2}{4}\right)$ *it holds that* $|\frac{1}{T}\sum_{t=1}^T p_{t-1}(x_i)p_t(x_j) - \frac{1}{T}\sum_{t=1}^T \delta_{t-1}(x_i)\delta_t(x_j)| < \epsilon$ *and* $|\frac{1}{T}\sum_{t=1}^T p_t(x_i) - \frac{1}{T}\sum_{t=1}^T \delta_t(x_i)| < \epsilon$*, simultaneously for all* i.

Proof. We consider the random variable $Z_t = \delta_t(x_i) - p_t(x_i)$, notice that $\mathbb{E}[Z_t|p_1,\cdots,p_{t-1}] = 0$ so that $\{Z_t\}_t$ is a bounded martingale sequence with $|Z_t| < 1$ and thus by Azuma's inequality we have $\mathbb{P}\left[\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| \geq \epsilon\right] < 2\exp\left(-\frac{T\epsilon^2}{2}\right)$ which shows that $\frac{1}{T}\sum_{t=1}^{T}\delta_t(a_i)$ concentrates around $\frac{1}{T}\sum_{t=1}^{T} p_t(x_i)$. Let $R_t = \delta_{t-1}(x_i)\delta_t(x_j) - p_{t-1}(x_i)p_t(x_j)$ and consider the filtration $\{\mathcal{F}_t\}_t$, where $\mathcal{F}_1 = \emptyset$, $\mathcal{F}_t = \Sigma(\delta_1, \cdots, \delta_t)$ is the sigma algebra generated by the random variables δ_1 to δ_t. Then $|R_{2t}| \leq 1$ and $\mathbb{E}[R_{2t}|\mathcal{F}_1, \cdots, \mathcal{F}_{2t-2}] = 0$, so $\{R_{2t}\}_t$ is also a bounded martingale difference and thus $\mathbb{P}\left[\left|\frac{1}{T}\sum_{t=1}^{\frac{T}{2}} R_{2t}\right| \geq \frac{\epsilon}{2}\right] < 2\exp\left(-\frac{T\epsilon^2}{4}\right)$. A similar argument allows us to bound the sum of the R_{2t+1}'s and a union bound gives us $\mathbb{P}\left[\left|\frac{1}{T}\sum_{t=1}^{T} R_t\right| \geq \epsilon\right] < 4\exp\left(-\frac{T\epsilon^2}{4}\right)$. A union bound over all i finishes the proof. □

Definition 4.5.1. Define the perturbed distribution of player i at time t to be $\tilde{p}_t^i = (1 - \sqrt{|\mathcal{A}|\tilde{\epsilon}})p_t^i + \mathbf{1}\frac{\sqrt{|\mathcal{A}|\tilde{\epsilon}}}{|\mathcal{A}_i|}$.

Lemma 4.5.2. *The difference of expected utilities from playing according to $(\tilde{p}_t^i)_{t=1}^T$ instead of $(p_t^i)_{t=1}^T$ is at most $2T\sqrt{|\mathcal{A}|\tilde{\epsilon}}$*

Proof. From lemma 4.5.4 at each time step the difference of expected utility is bounded by $\sqrt{|\mathcal{A}|\tilde{\epsilon}}$ in absolute value. □

Theorem 4.5.3. *If at time t player i plays according to $\tilde{p}_t^i$, where $\tilde{\epsilon} = \frac{T^{-1/4}}{|\mathcal{A}|}$ and $p_t = \tilde{p}_t^1(\tilde{p}_t^2)^\top$, then the regret for playing according to $\tilde{p}_t^i$ is at most $O(T^{7/8})$. Further $\limsup_{T\to\infty}\|\tilde{\mathrm{M}} - \widehat{\mathrm{M}}\|_2 = 0$, almost surely. Additionally if $\tilde{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\delta_t$ is the stationary distribution of $\tilde{\mathrm{M}}$ corresponding to the observed play and $\widehat{\sigma} = \frac{1}{T}\sum_{t=1}^{T} p_t$ is the stationary distribution of $\widehat{\mathrm{M}}$ corresponding to the averaged empirical distribution, then $\limsup_{T\to\infty}\|\tilde{\sigma} - \widehat{\sigma}\|_1 = 0$ almost surely.*

Proof. Set $\tilde{\epsilon} = \frac{T^{-1/4}}{|\mathcal{A}|}$. The regret bound of the no-external regret algorithms now becomes $O(T^{7/8})$. We can, however, now guarantee that $\sum_{t=1}^{T} p_t(x_i) \geq \frac{T^{-1/4}}{|\mathcal{A}|}$ and thus, combining this with the high probability bound we obtain that with probability at least $1 - C\exp\left(-\frac{T\epsilon^2}{4}\right)$ it holds that $|\tilde{\mathrm{M}}_{i,j} - \widehat{\mathrm{M}}_{i,j}| < 2\epsilon\frac{T^{1/4}}{|\mathcal{A}|}$. To see this, let $x = \frac{1}{T}\sum_{t=1}^{T} p_{t-1}(x_i)p_t(x_j), \widehat{x} = \frac{1}{T}\sum_{t=1}^{T}\delta_{t-1}(x_i)\delta_t(x_j), y = \frac{1}{T}\sum_{t=1}^{T} p_t(x_i), \widehat{y} = \frac{1}{T}\sum_{t=1}^{T}\delta_t(x_i)$.

Then

$$|\tilde{\mathrm{M}}_{i,j} - \widehat{\mathrm{M}}_{i,j}| = \left|\frac{x}{y} - \frac{\widehat{x}}{\widehat{y}}\right| \leq \frac{|x - \widehat{x}|}{|y|} + \frac{|\widehat{x}|}{|1/y - 1/\widehat{y}|} \leq \frac{\epsilon}{\tilde{\epsilon}} + \frac{|\widehat{x}||y - \widehat{y}|}{|y\widehat{y}|} \leq 2\frac{\epsilon}{\tilde{\epsilon}},$$

where the last inequality holds because $\widehat{x} \leq \widehat{y}$. Setting $\epsilon = T^{-1/3}$ and a union bound we arrive at $\mathbb{P}\left[\|\tilde{\mathrm{M}} - \widehat{\mathrm{M}}\|_2 > T^{-1/12}\right] < C\exp\left(-\frac{T^{1/3}}{4}\right)$. By Borel-Cantelli lemma we have $\limsup_{T\to\infty}\|\tilde{\mathrm{M}} - \widehat{\mathrm{M}}\|_2 = 0$ almost surely. From lemma 4.5.1 and a union bound we know that with $\mathbb{P}\left[\|\tilde{\sigma} - \widehat{\sigma}\|_1 > \epsilon\right] < 2|\mathcal{A}|\exp\left(-\frac{T\epsilon^2}{4}\right)$. Setting $\epsilon = T^{-1/3}$ and again using Borel-Cantelli's lemma we see that $\limsup_{T\to\infty}\|\tilde{\sigma} - \widehat{\sigma}\|_1 = 0$. □

4.5.2 Auxiliary results

Lemma 4.5.4. *Let σ and σ' be two distributions supported on a finite set and let f be a utility/loss function uniformly bounded by 1. If $\|\sigma - \sigma'\|_1 \leq \epsilon$ then $|\mathbb{E}_{a\sim\sigma}[f(a)] - \mathbb{E}_{a\sim\sigma'}[f(a)]| \leq \epsilon$.*

Proof.

$$\begin{aligned}|\mathbb{E}_{s\sim\sigma}[f(s)] - \mathbb{E}_{s'\sim\sigma'}[f(s)]| &= |\sum_{s\in S}\sigma(s)f(s) - \sum_{s\in S}\sigma'(s)f(s)| \\ = |\sum_{s\in S}f(s)(\sigma(s) - \sigma'(s))| &\leq \sum_{s\in S}|\sigma(s) - \sigma'(s)| = \|\sigma - \sigma'\|_1 \leq \epsilon.\end{aligned}$$

□

Lemma 4.5.5. *Let $\mathrm{M} \in \mathbb{R}^{d\times d}$ and $\widehat{\mathrm{M}} \in \mathbb{R}^{d\times d}$ be two row-stochastic matrices, such that $\|\mathrm{M} - \widehat{\mathrm{M}}\| \leq \epsilon$, then for any stationary distribution $\widehat{\sigma}$ of $\widehat{\mathrm{M}}$, there exists a stationary distribution σ of M, such that $\|\widehat{\sigma} - \sigma\|_1 \leq \frac{4d^2\epsilon}{\delta}$.*

Proof. Let $\mathrm{U} \in \mathbb{R}^{d\times k}$ be the left singular vectors corresponding to the singular value 1 of M and let $\widehat{\mathrm{U}} \in \mathbb{R}^{d\times l}$ be the left singular vectors corresponding to the singular value 1 of $\widehat{\mathrm{M}}$. First notice that

$$\|\mathrm{M}\mathrm{M}^\top - \widehat{\mathrm{M}}\widehat{\mathrm{M}}^\top\| \leq \|\mathrm{M}\mathrm{M}^\top - \mathrm{M}\widehat{\mathrm{M}}^\top\| + \|\mathrm{M}\widehat{\mathrm{M}}^\top - \widehat{\mathrm{M}}\widehat{\mathrm{M}}^\top\| \leq (\|\mathrm{M}\| + \|\widehat{\mathrm{M}}\|)\|\mathrm{M} - \widehat{\mathrm{M}}\| \leq 2\epsilon$$

Denote the eigen-gap of M by δ, then by Wedin's theorem (see for example lemma B.3 in Allen-Zhu and Li (2016)) we have

$$\|\widehat{\mathrm{U}}^\top \mathrm{U}^\perp\| \leq \frac{\|\mathrm{M}\mathrm{M}^\top - \widehat{\mathrm{M}}\widehat{\mathrm{M}}^\top\|}{\delta} \leq \frac{2\epsilon}{\delta}.$$

WLOG assume $\widehat{\sigma} = \frac{(\widehat{\mathrm{U}})_i}{\|(\widehat{\mathrm{U}})_i\|_1}$. This implies that $\|\widehat{\sigma}^\top \mathrm{U}^\perp\|_2 \leq \frac{2d\epsilon}{\delta}$ and thus:

$$\|\mathrm{U}\mathrm{U}^\top \widehat{\sigma} - \widehat{\sigma}\|_2 = \|(I - \mathrm{U}\mathrm{U}^\top)\widehat{\sigma}\|_2 = \|\mathrm{U}^\perp(\mathrm{U}^\perp)^\top \widehat{\sigma}\|_2 \leq \|\widehat{\sigma}^\top \mathrm{U}^\perp\|_2 \leq \frac{2d\epsilon}{\delta}.$$

Let $\sigma_i = \frac{\mathrm{U}_i}{\|\mathrm{U}_i\|_1}$ be the stationary distribution of M, corresponding to the i-th left singular vector and let $\alpha_i = (\mathrm{U}^\top \widehat{\sigma})_i \|\mathrm{U}_i\|_1 \geq 0$. Then we have $\|\sum_i \alpha_i \sigma_i - \widehat{\sigma}\|_1 \leq \frac{2d^2\epsilon}{\delta}$, where the inequality follows from the derivation above and the inequality between l_1 and l_2 norms. Let $\sigma = \frac{\sum_i \alpha_i \sigma_i}{\|\sum_i \alpha_i \sigma_i\|_1}$. This is a stationary distribution of M, since

$$\sigma^\top \mathrm{M} = \frac{1}{\|\sum_i \alpha_i \sigma_i\|_1} \sum_i \alpha_i \sigma_i^\top \mathrm{M} = \frac{\sum_i \alpha_i \sigma_i}{\|\sum_i \alpha_i \sigma_i\|_1} = \sigma.$$

Notice that by reverse triangle inequality we have

$$|\|\sum_i \alpha_i \sigma_i\|_1 - \|\widehat{\sigma}\|_1| \leq \frac{2d^2\epsilon}{\delta},$$

or equivalently

$$|\|\sum_i \alpha_i \sigma_i\|_1 - 1| \leq \frac{2d^2\epsilon}{\delta}.$$

Thus we have:

$$\|\sigma - \widehat{\sigma}\|_1 \leq \|\sum_i \alpha_i \sigma_i - \widehat{\sigma}\|_1 + \|\sigma - \sum_i \alpha_i \sigma_i - \widehat{\sigma}\|_1 \leq \frac{2d^2\epsilon}{\delta} + \|\sigma\|_1 |1 - \|\sum_i \alpha_i \sigma_i\|_1| \leq \frac{4d^2\epsilon}{\delta}.$$

□

Corollary 4.5.6. *Let the empirical distribution of observed play be $\tilde{\sigma}^T = \frac{1}{T}\sum_{t=1}^T \delta_t$, the empirical distribution of play if player 1 deviated to playing fixed action $a \in \mathcal{A}_1$ be $\tilde{\sigma}_a^T$ and the empirical distribution of play if player 2 to action $b \in \mathcal{A}_2$ be $\tilde{\sigma}_b^T$. The sequence $(\tilde{\sigma}^T, \tilde{\sigma}_a^T, \tilde{\sigma}_b^T)_T$ converges to the set P almost surely.*

Proof. Lemma 4.5.4, together with Theorem 4.5.3 imply that

$$\limsup_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0$$

almost surely i.e.

$$\mathbb{P}\left[\limsup_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0\right] = 1.$$

Since

$$\begin{aligned}&\limsup_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| \geq \\ &\liminf_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| \geq 0,\end{aligned}$$

this implies that

$$\mathbb{P}\left[\liminf_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0\right] = 1.$$

On the other hand this implies

$$\begin{aligned}\mathbb{P}\Bigg[&\liminf_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0 \bigcap \\ &\limsup_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0\Bigg] \geq 1\end{aligned}$$

and so

$$\mathbb{P}\left[\lim_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\right| = 0\right] = 1.$$

In a similar way we can get $\lim_{T\to\infty} \left|\mathbb{E}_{(a,b)\sim\tilde{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right]\right| = 0$ a.s. The above imply that $\lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right] = 0$ a.s. and $\lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\tilde{\sigma}_a^T}\left[u_1(a,b)\right] = 0$ a.s. and thus:

$$\begin{aligned}0 &= \lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right] + \lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\tilde{\sigma}_a^T}\left[u_1(a,b)\right] \\ &= \lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\tilde{\sigma}_a^T}\left[u_1(a,b)\right] + \lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right]\end{aligned}$$

a.s.. Since $\mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right] < o(1)$, this implies that

$$\begin{aligned}0 &\leq -\lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\widehat{\sigma}_a^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\widehat{\sigma}^T}\left[u_1(a,b)\right] \\ &= \lim_{T\to\infty} \mathbb{E}_{(a,b)\sim\tilde{\sigma}^T}\left[u_1(a,b)\right] - \mathbb{E}_{(a,b)\sim\tilde{\sigma}_a^T}\left[u_1(a,b)\right]\end{aligned}$$

a.s.. Now we can proceed as in the proof of Theorem 4.4.8. □

Chapter 5

Limits of learning in Tabular Reinforcement Learning

We provide improved gap-dependent regret bounds for reinforcement learning in finite episodic Markov decision processes. Compared to prior work, our bounds depend on alternative definitions of gaps. These definitions are based on the insight that, in order to achieve a favorable regret, an algorithm does not need to learn how to behave optimally in states that are not reached by an optimal policy. We prove tighter upper regret bounds for optimistic algorithms and accompany them with new information-theoretic lower bounds for a large class of MDPs. Our results show that optimistic algorithms can not achieve the information-theoretic lower bounds even in deterministic MDPs unless there is a unique optimal policy. The main contributions of this chapter are based on work carried out during my internship at Google Research, New York. This work was done in collaboration with Dr. Christoph Dann, Dr. Mehryar Mohri, and Dr. Julian Zimmert.

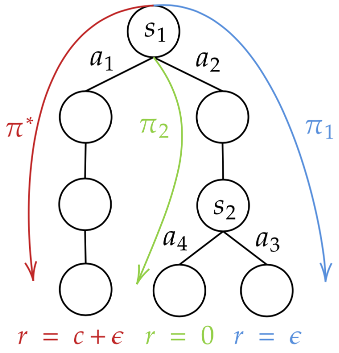
5.1 Instance dependent bounds in prior work and limitations

While the performance of RL algorithms in tabular Markov decision processes has been the subject of many studies in the past (e.g. Fiechter, 1994; Kakade, 2003;

Osband et al., 2013; Dann et al., 2017; Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019; Dann, 2019), the vast majority of existing analyses focuses on worst-case problem-independent regret bounds, which only take into account the size of the MDP, the horizon H and the number of episodes K.

Recently, however, some significant progress has been achieved towards deriving more optimistic problem-dependent guarantees. This includes more refined regret bounds for the tabular episodic setting that depend on structural properties of the specific MDP considered (Simchowitz and Jamieson, 2019; Lykouris et al., 2019; Jin and Luo, 2020; Foster et al., 2020b; He et al., 2020). Motivated by instance-dependent analyses in multi-armed bandits (Lai and Robbins, 1985), these analyses derive gap-dependent regret-bounds of the form $O\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{H\log(K)}{\text{gap}(s,a)}\right)$, where the sum is over state-actions pairs (s, a) and where the gap notion is defined as the difference of the optimal value function V^* of the Bellman optimal policy π^* and the Q-function of π^* at a sub-optimal action: $\text{gap}(s, a) = V^*(s) - Q^*(s, a)$. We will refer to this gap definition as ***value-function gap*** in the following. We note that a similar notion of gap has been used in the infinite horizon setting to achieve instance-dependent bounds (Auer and Ortner, 2007; Tewari and Bartlett, 2008; Auer et al., 2009; Filippi et al., 2010; Ok et al., 2018), however, a strong assumption about irreducability of the MDP is required. We discuss such bounds in Section 5.3.

While regret bounds based on these value function gaps generalize the bounds available in the multi-armed bandit setting, we argue that they have a major limitation. The bound at each state-action pair depends only on the gap at the pair and treats all state-action pairs equally, ignoring their topological ordering in the MDP. This can have a major impact on the derived bound. In this paper, we address this issue and formalize the following key observation about the difficulty of RL in an episodic MDP through improved instance-dependent regret bounds:



	Value-function gap (prior)	Return gap (ours)
General Regret bounds	$O\Big(\sum_{s,a}\frac{H\log(K)}{\mathrm{gap}(s,a)}\Big)$ $\Omega\Big(\sum_{s,a:\ s\in\pi^*}\frac{\log(K)}{\mathrm{gap}(s,a)}\Big)$	$O\Big(\sum_{s,a}\frac{\log(K)}{\overline{\mathrm{gap}}(s,a)}\Big)$ $\Omega\Big(\sum_{s,a}\frac{\log(K)}{H\overline{\mathrm{gap}}(s,a)}\Big)$
Example on the left	$\mathrm{gap}(s_1,a_2) = c$ $\mathrm{gap}(s_2,a_4) = \epsilon$ $O\Big(\frac{SH\log(K)}{\epsilon}\Big)$	$\overline{\mathrm{gap}}(s_1,a_2) = c$ $\overline{\mathrm{gap}}(s_2,a_4) = \frac{c+\epsilon}{H} \approx c$ $O\Big(\frac{SH\log(K)}{c}\Big)$

Figure 5-1: Comparison of our contributions in MDPs with deterministic transitions. Bounds only include the main terms and all sums over (s,a) are understood to only include terms where the respective gap is nonzero. $\overline{\text{gap}}$ is a our alternative ***return gap*** definition introduced later (Definition 5.4.1).

> ***Learning a policy with optimal return does not require an RL agent to distinguish between actions with similar outcomes (small value-function gap) in states that can only be reached by taking highly suboptimal actions (large value-function gap).***

To illustrate this insight, consider autonomous driving, where each episode corresponds to driving from a start to a destination. If the RL agent decides to run a red light on a crowded intersection, then a car crash is inevitable. Even though the agent could slightly affect the severity of the car crash by steering, this effect is small and, hence, a good RL agent does not need to learn how to best steer after running a red light. Instead, it would only need a few samples to learn to obey the traffic light in the first place as the action of disregarding a red light has a very large value-function gap.

To understand how this observation translates into regret bounds, consider the toy example in Figure 5-1. This MDP has deterministic transitions and only terminal rewards with $c \gg \epsilon > 0$. There are two decision points, s_1 and s_2, with two actions each, and all other states have a single action. There are three policies which govern the regret bounds: π^* (red path) which takes action a_1 in state s_1; π_1 which takes action a_2 at s_1 and a_3 at s_2 (blue path); and π_2 which takes action a_2 at s_1 and a_4 at s_2 (green path). Since π^* follows the red path, it never reaches s_2 and achieves optimal return $c + \epsilon$, while π_1 and π_2 are both suboptimal with return ϵ and 0 respectively. Existing value-function gaps evaluate to $\text{gap}(s_1, a_2) = c$ and $\text{gap}(s_2, a_4) = \epsilon$ which yields a regret bound of order $H \log(K)(1/c + 1/\epsilon)$. The idea behind these bounds is to capture the necessary number of episodes to distinguish the value of the optimal policy π^* from the value of any other sub-optimal policy ***on all states***. However, since π^* will never reach s_2 it is not necessary to distinguish it from any other policy at s_2. A good algorithm only needs to determine that a_2 is sub-optimal in s_1, which eliminates both π_1 and π_2 as optimal policies after only $\log(K)/c^2$ episodes. This suggests a regret of order $O(\log(K)/c)$. The bounds presented in this paper achieve this rate up to factors of H by replacing the gaps at every state-action pair with the average of all gaps along certain paths containing the state action pair. We call these averaged gaps ***return gaps***. The return gap at (s, a) is denoted as $\overline{\text{gap}}(s, a)$. Our new bounds replace $\text{gap}(s_2, a_4) = \epsilon$ by $\overline{\text{gap}}(s_2, a_4) \approx \frac{1}{2}\text{gap}(s_1, a_2) + \frac{1}{2}\text{gap}(s_2, a_4) = \Omega(c)$. Notice that ϵ and c can be selected arbitrarily in this example. In particular, if we take $c = 0.5$ and $\epsilon = 1/\sqrt{K}$ our bounds remain logarithmic $O(\log(K))$, while prior regret bounds scale as $\sqrt{K}$.

This work is motivated by the insight just discussed. First, we show that improved regret bounds are indeed possible by proving a tighter regret bound for StrongEuler, an existing algorithm based on the optimism-in-the-face-of-uncertainty (OFU) principle (Simchowitz and Jamieson, 2019). Our regret bound is stated in terms of our new return gaps that capture the problem difficulty more accurately and avoid explicit dependencies on the smallest value function gap $\text{gap}_{\min}$. Our technique applies to optimistic algorithms in general and as a by-product improves the dependency on episode length H of prior results. Second, we investigate the difficulty of RL in episodic MDPs from an information-theoretic perspective by deriving regret lower-bounds. We show that existing value-function gaps are indeed sufficient to capture difficulty of problems but only when each state is visited by an optimal policy with some probability. Finally, we prove a new lower bound when the transitions of the MDP are deterministic that depends only on the difference in return of the optimal policy and suboptimal policies, which is closely related to our notion of return gap.

5.2 Problem setting and notation

We now recall some of the notation and the RL setting introduced in Section 1.5.

We consider reinforcement learning in episodic tabular MDPs with a fixed horizon. An MDP can be described as a tuple $(\mathcal{S}, \mathcal{A}, P, R, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are state- and action-space of size S and A respectively, P is the state transition distribution with $P(\cdot|s,a) \in \Delta^{S-1}$ the next state probability distribution, given that action a was taken in the current state s. R is the reward distribution defined over $\mathcal{S} \times \mathcal{A}$ and $r(s,a) = \mathbb{E}[R(s,a)] \in [0,1]$. Episodes admit a fixed length or ***horizon*** H.

We consider ***layered*** MDPs: each state $s \in \mathcal{S}$ belongs to a layer $\kappa(s) \in [H]$ and the only non-zero transitions are between states s, s' in consecutive layers, with $\kappa(s') = \kappa(s) + 1$. This common assumption (see e.g. Krishnamurthy et al., 2016) corresponds to MDPs with time-dependent transitions, as in (Jin et al., 2018; Dann et al., 2017), but allows us to omit an explicit time-index in value-functions and policies. For ease of presentation, we assume there is a unique start state s_1 with $\kappa(s_1) = 1$ but our results can be generalized to multiple (possibly adversarial) start states. Similarly, for convenience, we assume that all states are reachable by some policy with non-zero probability, but not necessarily all policies or the same policy.

We denote by K the number of episodes during which the MDP is visited. Before each episode $k \in [K]$, the agent selects a deterministic policy $\pi_k \colon \mathcal{S} \to \mathcal{A}$ out of a set of all policies Π and π_k is then executed for all H time steps in episode k. For each policy π, we denote by $w^\pi(s,a) = \mathbb{P}(S_{\kappa(s)} = s, A_{\kappa(s)} = a \mid A_h = \pi(S_h)\ \forall h \in [H])$ and $w^\pi(s) = \sum_a w^\pi(s,a)$ probability of reaching state-action pair (s,a) and state s

respectively when executing π. For convenience, $supp(\pi) = \{s \in \mathcal{S} \colon w^\pi(s) > 0\}$ is the set of states visited by π with non-zero probability. The Q- and value function of a policy π are

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{h=\kappa(s)}^{H} r(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s, A_{\kappa(s)} = a\right], \quad \text{and} \quad V^\pi(s) = Q^\pi(s, \pi(s))$$

and the regret incurred by the agent is the sum of its regret over K episodes

$$R(K) = \sum_{k=1}^{K} v^* - v^{\pi_k} = \sum_{k=1}^{K} V^*(s_1) - V^{\pi_k}(s_1), \tag{5.1}$$

where $v^\pi = V^\pi(s_1)$ is the expected total sum of rewards or ***return*** of π and V^* is the optimal value function $V^*(s) = \max_{\pi\in\Pi} V^\pi(s)$. Finally, the set of optimal policies is denoted as $\Pi^* = \{\pi \in \Pi : V^\pi = V^*\}$. Note that we only call a policy optimal if it satisfies the Bellman equation in every state, as is common in literature, but there may be policies outside of Π^* that also achieve maximum return because they only take suboptimal actions outside of their support. The variance of the Q function at a state-action pair (s,a) of the optimal policy is $\mathcal{V}^*(s,a) = \mathbb{V}[R(s,a)] + \mathbb{V}_{s'\sim P(\cdot|s,a)}[V^*(s')]$, where $\mathbb{V}[X]$ denotes the variance of the r.v. X. The maximum variance over all state-action pairs is $\mathcal{V}^* = \max_{(s,a)} \mathcal{V}^*(s,a)$. Finally, our proofs will make use of the following clipping operator $\mathrm{clip}[a|b] = \chi(a \geq b)a$ that sets a to zero if it is smaller than b.

5.3 Related work

Instance dependent regret lower bounds for the MAB were first introduced in Lai and Robbins (1985). Later Graves and Lai (1997) extend such instance dependent lower bounds to the setting of controlled Markov chains, while assuming infinite horizon and certain properties of the stationary distribution of each policy. Building on their work, more recently Combes et al. (2017) establish instance dependent lower bounds for the Structured Stochastic Bandit problem. Very recently, in the stochastic MAB, Garivier et al. (2019) generalize and simplify the techniques of Lai and Robbins (1985) to completely characterize the behavior of uniformly good algorithms. The work of Ok et al. (2018) builds on these ideas to provide an instance dependent lower bound for infinite horizon MDPs, again under assumptions of how the stationary distributions of each policy will behave and irreducibility of the Markov chain. The idea behind deriving the above bounds is to use the uniform goodness of the studied algorithm to argue that the algorithm must select a certain policy or action at least a fixed number of times. This number is governed by a change of environment under which said policy/action is now the best overall. The reasoning now is that unless the algorithm is able to distinguish between these two environments it will have to incur linear regret asymptotically. Since the algorithm is uniformly good this can not happen.

For infinite horizon MDPs with additional assumptions the works of Auer and Ortner (2007); Tewari and Bartlett (2008); Auer et al. (2009); Filippi et al. (2010); Ok et al. (2018) establish logarithmic in horizon regret bounds of the form $O(D^2S^2A\log{()}T)/\delta)$, where δ is a gap-like quantity and D is a diameter measure. We now discuss the works of (Tewari and Bartlett, 2008; Ok et al., 2018), which should give more intuition about how the infinite horizon setting differs from our setting. Both works consider the non-episodic problem and therefore make some assumptions about the MDP $\mathcal{M}$. The main assumption, which allows for computationally tractable algorithms is that of irreducibility. Formally both works require that under any policy the induced Markov chain is irreducible. Intuitively, the notion of irreducibility allows for coming up with exploration strategies, which are close to min-max optimal and are easy to compute. In (Ok et al., 2018) this is done by considering the same semi-infinite LP 5.27 as in our work. Unlike our work, however, assuming that the Markov chain induced by the optimal policy π^* is irreducible allows for a nice characterization of the set $\Lambda(\theta)$ of "confusing" environments. In particular the authors manage to show that at every state s it is enough to consider the change of environment which makes the reward of any action $a : (s,a) \notin \pi^*$ equal to the reward of $a' : (s,a') \in \pi^*$. Because of the irreducability assumption we know that the support of $P(\cdot|s,a)$ is the same as the support of $P(\cdot|s,a')$ and this implies that the above change of environment makes the policy π which plays (s,a) and then coincides with π^* optimal. Some more work shows that considering only such changes of environment is sufficient for an equivalent formulation to the LP5.27. Since this is an LP with at most $S \times A$ constraints it is solvable in polynomial time and hence a version of the algorithm in (Combes et al., 2017) results in asymptotic min-max rates for the problem. The exploration in (Tewari and Bartlett, 2008) is also based on a similar LP, however, slightly more sophisticated.

Very recently there has been a renewed interest in proposing instance dependent regret bounds for finite horizon tabular MDPs (Simchowitz and Jamieson, 2019; Lykouris et al., 2019; Jin and Luo, 2020). The works of (Simchowitz and Jamieson, 2019; Lykouris et al., 2019) are based on the OFU principle and the proposed regret bounds scale as $O(\sum_{(s,a)\notin\pi^*} H\log{()T}/\operatorname{gap}(s,a) + SH\log{()T}/\operatorname{gap}_{\min})$, disregarding variance terms and terms depending only poli-logarithmically on the gaps. The setting in (Lykouris et al., 2019) also considers adversarial corruptions to the MDP, unknown to the algorithm, and their bound scales with the amount of corruption. Jin and Luo (2020) derive similar upper bounds, however, the authors assume a known transition kernel and take the approach of modelling the problem as an instance of Online Linear Optimization, through using occupancy measures (Zimin and Neu, 2013). For the problem of Q-learning, Yang et al. (2020); Du et al. (2020), also propose algorithms with regret scaling as $O(SAH^6\log{()T}/\operatorname{gap}_{\min})$. All of these bounds scale at least as $\Omega(SH\log{()T}/\operatorname{gap}_{\min})$. Simchowitz and Jamieson (2019) show an MDP instance on which no optimistic algorithm can hope to do better.

5.4 Novel upper bounds for optimistic algorithms

5.4.1 Optimistic algorithms and StrongEuler

We begin this section by describing one approach to solving the RL problem in the above setting which is through optimistic algorithms. Optimistic algorithms maintain estimators of the Q-functions at every state-action pair such that there exists at least one policy π for which the estimator, $\bar{Q}^{\pi}$, overestimates the Q-function of the optimal policy, that is $\bar{Q}^{\pi}(s,a) \geq Q^*(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. During episode $k \in [K]$, the optimistic algorithm selects the policy π_k with highest optimistic value function $\bar{V}_k$. By definition, it holds that $\bar{V}_k(s) \geq V^*(s)$. The optimistic value and Q-functions are constructed through finite-sample estimators of the true rewards $r(s,a)$ and the transition kernel $\mathbb{P}(\cdot|s,a)$ plus bias terms, similar to estimators for the UCB MAB algorithm. Careful construction of these bias terms is crucial for deriving min-max optimal regret bounds in S, A and H (Azar et al., 2017). Bias terms which yield the tightest known bounds come from concentration of martingales results such as Freedman's inequality (Freedman, 1975) and empirical Bernstein's inequality for martingales (Maurer and Pontil, 2009).

StrongEuler is the optimistic algorithm proposed by Simchowitz and Jamieson (2019). The algorithm satisfies a stronger notion of optimism called ***strong optimism***. To define strong optimism we need the notion of ***surplus*** which roughly measures the optimism at a fixed state-action pair. Formally the surplus at (s,a) during episode k is defined as

$$E_k(s,a) = \bar{Q}_k(s,a) - r(s,a) - \langle P(\cdot|s,a), \bar{V}_k \rangle \ . \tag{5.2}$$

We say that an algorithm is strongly optimistic if $E_k(s,a) \geq 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, k \in [K]$. Surpluses are also central to our new regret bounds and we will carefully discuss their use in Section 5.4.3.

5.4.2 Prior optimistic regret bounds and opportunities for improvement

As hinted to in the introduction, the way prior regret bounds treat value-function gaps independently at each state-action pair can lead to excessively loose guarantees. Bounds that use value-function gaps (Simchowitz and Jamieson, 2019; Lykouris et al., 2019; Jin and Luo, 2020) scale at least as

$$\sum_{s,a\colon\ \operatorname{gap}(s,a)>0} \frac{H\log(K)}{\operatorname{gap}(s,a)} + \sum_{s,a\colon\ \operatorname{gap}(s,a)=0} \frac{H\log(K)}{\operatorname{gap}_{\min}},$$

where state-action pairs with zero gap appear, with $\operatorname{gap}_{\min} = \min_{s,a\colon\ \operatorname{gap}(s,a)>0} \operatorname{gap}(s,a)$, the smallest positive gap. To illustrate where these bounds are loose, let us revisit the example in Figure 5-1. Here, these bounds evaluate to $\frac{H\log(K)}{c} + \frac{H\log(K)}{\epsilon} + \frac{SH\log(K)}{\epsilon}$, where the first two terms come from state-action pairs with positive value-function gaps and the last term comes from all the state-action pairs with zero gaps. There are several opportunities for improvement:

O.1 State-action pairs that can only be visited by taking optimal actions: We should not pay the $1/\operatorname{gap}_{\min}$ factor for such (s,a) as there are no other suboptimal policies π to distinguish from π^* in such states.

O.2 State-action pairs that can only be visited by taking at least one suboptimal action: We should not pay the $1/\operatorname{gap}(s_2,a_3)$ factor for state-action pair (s_2,a_3) and the $1/\operatorname{gap}_{\min}$ factor for (s_2,a_4) because no optimal policy visits s_2. Such state-action pairs should only be accounted for with the price to learn that a_2 is not optimal in state s_1. After all, learning to distinguish between π_1 and π_2 is unnecessary for optimal return.

Both opportunities suggest that the price $\frac{1}{\operatorname{gap}(s,a)}$ or $\frac{1}{\operatorname{gap}_{\min}}$ that each state-action pair (s,a) contributes to the regret bound can be reduced by taking into account the regret incurred by the time (s,a) is reached. Opportunity **O.1** postulates that if no regret can be incurred up to (and including) the time step (s, a) is reached, then this state-action pair should not appear in the regret bound. Similarly, if this regret is necessarily large, then the agent can learn this with few observations and stop reaching (s, a) earlier than $\text{gap}(s, a)$ may suggest. To illustrate Opportunity **O.2** better we consider an additional example. Our example can be found in Figure 5-2. The MDP is an extension of the

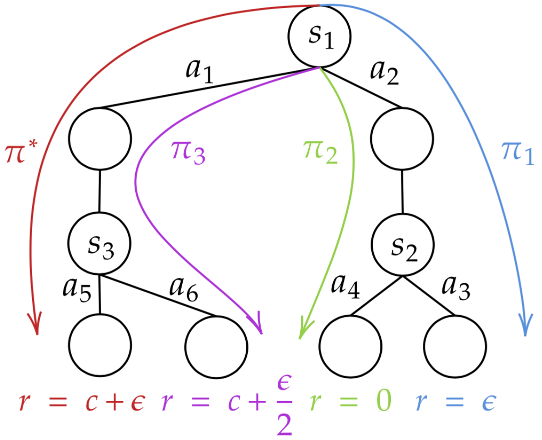


Figure 5-2: Example for Opportunity **O.2**

one presented in Figure 5-1 with the new addition of actions a_5 and a_6 in state s_3 and the new state following action a_6. Again there is only a single action available at all other states than s_1, s_2, s_3. The reward of the state following action a_6 is set as $r = c + \epsilon/2$. This defines a new sub-optimal policy π_3 and the gap $\text{gap}(s_3, a_6) = \frac{\epsilon}{2}$. Information theoretically it is impossible to distinguish π_3 as sub-optimal in less than $\Omega(\log(K)/\epsilon^2)$ rounds and so any uniformly good algorithm would have to pay at least $O(\log(K)/\epsilon)$ regret. However, what we observed previously still holds true, i.e., we should not have to play more than $\log(K)/c^2$ rounds to eliminate both π_1 and π_2 as sub-optimal policies. Prior work now suffers Opportunity **O.2** as it would pay $\log(K)/\epsilon$ regret for all zero gap state-action pairs belonging to either π_1 or π_2, essentially evaluating to $SA\log(K)/\epsilon$. On the other hand our bounds will only pay $\log(K)/\epsilon$ regret for zero gap state-action pairs belonging to π_3.

Since the total regret incurred during one episode by a policy π is simply the expected sum of value-function gaps visited (Lemma 5.5.1 in Section 5.5),

$$v^* - v^\pi = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h) \right], \tag{5.3}$$

we can measure the regret incurred up to reaching (S_t, A_t) by the sum of value function gaps $\sum_{h=1}^{t} \operatorname{gap}(S_h, A_h)$ up to this point t. We are interested in the regret incurred up to visiting a certain state-action pair (s, a) which π may visit only with some probability. We therefore need to take the expectation of such gaps conditioned on the event that (s, a) is actually visited. We further condition on the event that this regret is nonzero, which is exactly the case when the agent encounters a positive value-function gap within the first $\kappa(s)$ time steps. We arrive at

$$\mathbb{E}_\pi \left[\sum_{h=1}^{\kappa(s)} \operatorname{gap}(S_h, A_h) \;\middle|\; S_{\kappa(s)} = s, A_{\kappa(s)} = a, B \leq \kappa(s) \right],$$

where $B = \min\{h \in [H+1] \colon \operatorname{gap}(S_h, A_h) > 0\}$ is the first time a non-zero gap is visited. This quantity measures the regret incurred up to visiting (s, a) through suboptimal actions. If this quantity is large for all policies π, then a learner will stop visiting this state-action pair after few observations because it can rule out all actions that lead to (s, a) quickly. Conversely, if the event that we condition on has zero probability under any policy, then (s, a) can only be reached through optimal action choices (including a in s) and incurs no regret. This motivates our new definition of gaps that combines value function gaps with the regret incurred up to visiting the state-action pair:

Definition 5.4.1 (Return gap)**.** For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ define $\mathcal{B}(s, a) \equiv \{B \leq \kappa(s), S_{\kappa(s)} = s, A_{\kappa(s)} = a\}$, where B is the first time a non-zero gap is encountered. $\mathcal{B}(s, a)$ denotes the event that state-action pair (s, a) is visited and that a suboptimal action was played at any time up to visiting (s, a). We define the return gap as

$$\overline{\operatorname{gap}}(s, a) \equiv \operatorname{gap}(s, a) \vee \min_{\substack{\pi \in \Pi: \\ \mathbb{P}_\pi(\mathcal{B}(s,a)) > 0}} \frac{1}{H} \mathbb{E}_\pi \left[\sum_{h=1}^{\kappa(s)} \operatorname{gap}(S_h, A_h) \;\middle|\; \mathcal{B}(s, a) \right]$$

if there is a policy $\pi \in \Pi$ with $\mathbb{P}_\pi(\mathcal{B}(s,a)) > 0$ and $\overline{\operatorname{gap}}(s,a) \equiv 0$ otherwise.

The additional $1/H$ factor in the second term is a required normalization suggesting that it is the average gap rather than their sum that matters. Equipped with this definition, we are ready to state our main upper bound which pertains to the StrongEuler algorithm proposed by Simchowitz and Jamieson (2019).

Theorem 5.4.1 (Main Result (Informal))**.** *The regret $R(K)$ of* StrongEuler *is bounded with high probability for all number of episodes K as*

$$R(K) \lessapprox \sum_{\substack{(s,a)\in\mathcal{S}\times\mathcal{A}:\\ \overline{\operatorname{gap}}(s,a)>0}} \frac{\mathcal{V}^*(s,a)}{\overline{\operatorname{gap}}(s,a)} \log(K)\,.$$

In the above, we have restricted the bound to only those terms that have inverse polynomial dependence on the gaps.

Comparison with existing gap-dependent bounds. We now compare our bound to the existing gap-dependent bound for StrongEuler by Simchowitz and Jamieson (2019, Corollary B.1)

$$R(K) \lessapprox \sum_{\substack{(s,a)\in\mathcal{S}\times\mathcal{A}:\\ \operatorname{gap}(s,a)>0}} \frac{H\mathcal{V}^*(s,a)}{\operatorname{gap}(s,a)} \log(K) + \sum_{\substack{(s,a)\in\mathcal{S}\times\mathcal{A}:\\ \operatorname{gap}(s,a)=0}} \frac{H\mathcal{V}^*}{\operatorname{gap}_{\min}} \log(K)\,. \tag{5.4}$$

We here focus only on terms that admit a dependency on K and an inverse-polynomial dependency on gaps as all other terms are comparable. Most notable is the absence of the second term of Equation 5.4 in our bound in Theorem 5.4.1. Thus, while state-action pairs with $\overline{\operatorname{gap}}(s,a) = 0$ do not contribute to our regret bound, they appear with a $1/\operatorname{gap}_{\min}$ factor in existing bounds. Therefore, our bound addresses Opportunity **O.1** because it does not pay for state-action pairs that can only be visited through optimal actions. Further, state-action pairs that do contribute to our bound satisfy $\frac{1}{\overline{\operatorname{gap}}(s,a)} \le \frac{1}{\operatorname{gap}(s,a)} \wedge \frac{H}{\operatorname{gap}_{\min}}$ and thus never contribute more than in the existing bound in Equation 5.4. Therefore, our regret bound is never worse. In fact, it is significantly tighter when there are states that are only reachable by taking severely suboptimal actions, i.e., when the average value-function gaps are much larger than $\mathrm{gap}(s,a)$ or $\mathrm{gap}_{\min}$. By our definition of return gaps, we only pay the inverse of these larger gaps instead of $\mathrm{gap}_{\min}$. Thus, our bound also addresses **O.2** and achieves the desired $\log(K)/c$ regret bound in the motivating example of Figure 5-1 as opposed to the $\log(K)/\epsilon$ bound of prior work.

Regret bound when transitions are deterministic. We now interpret Definition 5.4.1 for MDPs with deterministic transitions and derive an alternative form of our bound in this case. Let $\Pi_{s,a}$ be the set of all policies that visit (s,a) and have taken a suboptimal action up to that visit, that is,

$$\Pi_{s,a} \equiv \Big\{ \pi \in \Pi \, : \, s^{\pi}_{\kappa(s)} = s, a^{\pi}_{\kappa(s)} = a, \exists \; h \leq \kappa(s), \mathrm{gap}(s^{\pi}_{h}, a^{\pi}_{h}) > 0 \Big\}.$$

where $(s^{\pi}_1, a^{\pi}_1, s^{\pi}_2, \dots, s^{\pi}_H, a^{\pi}_H)$ are the state-action pairs visited (deterministically) by π. Further, let $v^*_{s,a} = \max_{\pi \in \Pi_{s,a}} v^{\pi}$ be the best return of such policies. Definition 5.4.1 now evaluates to $\overline{\mathrm{gap}}(s,a) = \mathrm{gap}(s,a) \vee \frac{1}{H}(v^* - v^*_{s,a})$ and the bound in Theorem 5.4.1 can be written as

$$R(K) \lessapprox \sum_{s,a \colon \Pi_{s,a} \neq \varnothing} \frac{H \log(K)}{v^* - v^*_{s,a}} \,. \tag{5.5}$$

We show in Section 5.4.6, that it is possible to further improve this bound when the optimal policy is unique by only summing over state-action pairs which are not visited by the optimal policy.

5.4.3 Regret analysis with improved clipping: from minimum gap to average gap

In this section, we present the main technical innovations of our tighter regret analysis. Our framework applies to ***optimistic*** algorithms that maintain a Q-function estimate, $\bar{Q}_k(s,a)$, which overestimates the optimal Q-function $Q^*(s,a)$ with high probability in all states s, actions a and episodes k. We first give an overview of gap-dependent analyses and then describe our approach.

Overview of gap-dependent analyses. As already alluded to in Section 5.4.1 the central quantity in regret analyses of optimistic algorithms are the surpluses $E_k(s,a)$. Worst-case regret analyses bound the regret in episode k as $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_{\pi_k}(s,a)E_k(s,a)$, the expected surpluses under the optimistic policy π_k executed in that episode. Instead, gap-dependent analyses rely on a tighter version and bound the instantaneous regret by the ***clipped surpluses*** (e.g. Proposition 3.1 Simchowitz and Jamieson, 2019)

$$V^*(s_1) - V^{\pi_k}(s_1) \leq 2e\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[E_k(s,a) \;\middle|\; \frac{1}{4H}\mathrm{gap}(s,a) \vee \frac{\mathrm{gap}_{\min}}{2H}\right]. \tag{5.6}$$

Using concentration arguments one can show that $E_k(s,a)$ shrinks at a rate of $\sqrt{\frac{1}{n_k(s,a)}}$, with $n_k(s,a)$ the current number of times (s,a) has been visited. Thus, roughly speaking, after $\left(\frac{\mathrm{gap}(s,a)\vee\mathrm{gap}_{\min}}{H}\right)^{-2}$ visits to (s,a), the surplus falls below the threshold and does not contribute any more to the regret. Since each visit incurs at most $\frac{\mathrm{gap}(s,a)\vee\mathrm{gap}_{\min}}{H}$ regret, this leads to a $\frac{H}{\mathrm{gap}(s,a)\vee\mathrm{gap}_{\min}}$ dependency on gaps for each state-action pair.

Sharper clipping with general thresholds. Our main technical contribution for achieving a regret bound in terms of return gaps $\overline{\mathrm{gap}}(s,a)$ is the following improved surplus clipping bound:

Proposition 5.4.2 (Improved surplus clipping bound)**.** *Let the surpluses $E_k(s,a)$ be generated by an optimistic algorithm. Then the instantaneous regret of π_k is bounded as follows:*

$$V^*(s_1) - V^{\pi_k}(s_1) \leq 4\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[E_k(s,a) \;\middle|\; \frac{1}{4}\mathrm{gap}(s,a) \vee \epsilon_k(s,a)\right],$$

where $\epsilon_k \colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}_0^+$ is any clipping threshold function that satisfies

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h,A_h)\right] \leq \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \mathrm{gap}(S_h,A_h)\right].$$

Compared to previous surplus clipping bounds in Equation 5.6, there are several notable differences. First, instead of $\mathrm{gap}_{\min}/2H$, we can now pair $\mathrm{gap}(s,a)$ with more general clipping thresholds $\epsilon_k(s,a)$, as long as their expected sum over time steps after the first non-zero gap was encountered is at most half the expected sum of gaps. We will provide some intuition for this condition below. Note that $\epsilon_k(s,a) \equiv \frac{\mathrm{gap}_{\min}}{2H}$ satisfies the condition because the LHS evaluates to $\frac{\mathrm{gap}_{\min}}{2H}\mathbb{P}_{\pi_k}(B \leq H)$ and there must be at least one positive gap in the sum $\sum_{h=1}^H \mathrm{gap}(S_h, A_h)$ on the RHS in event $\{B \leq H\}$. Thus our bound recovers existing results. In addition, the first term in our clipping thresholds is $\frac{1}{4}\mathrm{gap}(s,a)$ instead of $\frac{1}{4H}\mathrm{gap}(s,a)$. Simchowitz and Jamieson (2019) are able to remove this spurious H factor only if the problem instance happens to be a bandit instance and the algorithm satisfies strong optimism where surpluses have to be non-negative. Our analysis does not require such conditions and therefore generalizes these existing results.[1]

Intuition for threshold condition in Proposition 5.4.2. The key to proving Proposition 5.4.2 is the following self-bounding trick for the instantaneous regret $V^*(s_1) - V^{\pi_k}(s_1)$. The self-bounding trick works in the following way. If we have a $\rho > 0$ which is upper bounded by $\gamma \geq \rho$ and lower bounded by $0 < \beta \leq \rho$, we can further bound $(1-c)\rho \leq \gamma - c\beta$ for any constant $c \in [0,1]$. We can now use this trick twice with $\rho = V^*(s_1) - V^{\pi_k}(s_1)$, $\gamma = \mathbb{E}_{\pi_k}[\sum_{h=1}^H E_k(S_h, A_h)]$, $c = 1/4$ and β equal to either the expected sum of gaps or the assumed lower bound with clipping functions, that is $\beta = \mathbb{E}_{\pi_k}[\sum_{h=1}^H \mathrm{gap}(S_h, A_h)]$ or $\beta = \mathbb{E}_{\pi_k}[\sum_{h=1}^H \epsilon_k(S_h, A_h)]$. This implies that one half of the instantaneous regret is bounded as

$$\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)) \leq \sum_{h=1}^H \mathbb{E}_{\pi_k}\left[E_k(S_h, A_h) - \frac{\mathrm{gap}(S_h, A_h) + \epsilon_k(S_h, A_h)}{4}\right].$$

Using the fact that the clip operator satisfies $a - b \leq \mathrm{clip}[a|b]$ gives the desired statement. This implies that, in order to achieve the tightest regret bound, we should

[1] Our layered state space assumption affects the H dependencies in lower-order terms in our final regret compared to Simchowitz and Jamieson (2019). However, Proposition 5.4.2 directly applies to their setting without any penalty in H.

clip $E_k(S_h, A_h)$ to the largest possible value. Thus, the goal is to lower bound the expected sum of gaps as tightly as possible by the clipping function. Besides this insight, introducing the stopping time B in the condition is key to addressing **O.1** and requires a careful treatment laid out in the full proof in Section 5.5.

Choice of clipping thresholds for return gaps. The condition in Proposition 5.4.2 suggests that one can set $\epsilon_k(S_h, A_h)$ to be proportional to the average expected gap under policy π_k:

$$\epsilon_k(s,a) = \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h) \,\middle|\, \mathcal{B}(s,a)\right]. \tag{5.7}$$

if $\mathbb{P}_{\pi_k}(\mathcal{B}(s,a)) > 0$ and $\epsilon_k(s,a) = \infty$ otherwise. Lemma 5.5.5 in Section 5.5 shows that this choice indeed satisfies the condition in Proposition 5.4.2. If we now take the minimum over all policies for π_k, then we can proceed with the standard analysis and derive our main result in Theorem 5.4.1. However, by avoiding the minimum over policies, we can derive a stronger regret bound that depends on the actual policies executed by the algorithm. We present this bound in the next section.

5.4.4 Policy-dependent regret bound

We will now show how our results from the previous sections translate into a stronger regret bound on the concrete example of StrongEuler. Ignoring lower-order terms, Simchowitz and Jamieson (2019) showed that the surpluses of this algorithm are bounded as $E_k(s,a) \lessapprox \sqrt{\frac{\mathcal{V}^*(s,a)\log(\bar{n}_k(s,a))}{\bar{n}_k(s,a)}}$ where $\bar{n}_k(s,a) = \sum_{j=1}^{k-1} w^{\pi_j}(s,a)$ are the expected number of samples for (s,a) up to episode k and $\mathcal{V}^*(s,a) = \mathbb{V}[R(s,a)] + \mathbb{V}_{s'\sim P(\cdot|s,a)}[V^*(s')]$ is the one-step variance term w.r.t. the optimal value function. In this case, Proposition 5.4.2 with the clipping to the average gap gives:

$$R(K) \lessapprox \sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[\sqrt{\frac{\mathcal{V}^*(s,a)\log\left(\bar{n}_k(s,a)\right)}{\bar{n}_k(s,a)}} \,\middle|\, \operatorname{gap}(s,a) \vee \epsilon_k(s,a)\right]. \tag{5.8}$$

An existing analysis now translates such a clipping bound into a $\log(K)$ regret bound using an integration argument (e.g. Simchowitz and Jamieson, 2019). However, we cannot rely on these arguments since our clipping thresholds may not be constant across episodes. To address this technical challenge, we derive the following lemma based on an optimization view on such terms

Lemma 5.4.3. *For any sequence of thresholds $\gamma_1, \dots \gamma_K > 0$, consider the problem with $x_0 = 1$*

$$\underset{x_1,\dots x_K \in [0,1]}{\text{maximize}} \sum_{k=1}^{K} x_k \sqrt{\frac{\log\left(\sum_{j=0}^{k} x_j\right)}{\sum_{j=0}^{k} x_j}} \quad \text{s.t.} \quad \text{for} \quad \text{all} \quad k \in [K] : \qquad \sqrt{\frac{\log\left(\sum_{j=0}^{k} x_j\right)}{\sum_{j=0}^{k} x_j}} \geq \gamma_k.$$

The optimal value is bounded for any $t \in [K]$ from above as $\frac{\log(t)}{\epsilon_t} + \sqrt{(K-t)\log(K)}$.

Applying this lemma for each (s,a) with $x_k = w^{\pi_k}(s,a)$ and appropriate clipping thresholds $\gamma_k \approx (\text{gap}(s,a) \vee \epsilon_k(s,a))/\sqrt{\mathcal{V}^*(s,a)}$ we can derive our main result:

Theorem 5.4.4. *When `StrongEuler` is run with confidence parameter δ, then with probability at least $1-\delta$, its regret is bounded for all number of episodes K as*

$$R(K) \lessapprox \sum_{s,a} \min_{t \in [K_{(s,a)}]} \left\{ \frac{\mathcal{V}^*(s,a) \log\left(\frac{M}{\delta}\right) \log\left(\frac{\mathcal{V}^*(s,a)\log(M/\delta)}{\text{gap}(s,a) \vee \epsilon_t(s,a)}\right)}{\text{gap}(s,a) \vee \epsilon_t(s,a)} + \sqrt{\mathcal{V}^*(s,a)(K_{(s,a)} - t)} \right\} \log(K)$$
$$+ S^2 A H^4 \log\left(\frac{MK}{\delta}\right) \log\left(\frac{MH}{\overline{\text{gap}}_{\min}}\right),$$

where $M \leq (SAH)^3$, $\overline{\text{gap}}_{\min} = \min_{(s,a)} \overline{\text{gap}}(s,a)$ and $K_{(s,a)}$ is the last episode during which a policy that may visit (s,a) was played.

A slightly more refined version of this bound is stated Theorem 5.5.11 in Section 5.5. Note that our regret bound depends through $\epsilon_t(s,a)$, the average gap encountered by π_t, on the policies that the algorithm played. We can smoothly interpolate between the worst-case rate of $\sqrt{K}$ achieved for $t \ll K$ when all gaps and average gaps are $O(\sqrt{\mathcal{V}^*(s,a)K})$ and the gap-dependent regret rate achieved when $t \approx K$ that scales inversely with gaps. Our bound depends on the gaps in all episodes and can benefit from choices for t that yields a large gap or average gap in late episodes.

Comparing with the bound in Simchowitz and Jamieson (2019). We now proceed to compare our bound directly to the one stated in Corollary B.1 (Simchowitz and Jamieson, 2019). We will ignore the factors with only poly-logarithmic dependence on gaps as they are are common between both bounds. We now recall the regret bound presented in Corollary B.1, modulo said factors:

$$R(K) \leq O\left(\sum_{(s,a)\in\mathcal{Z}_{sub}} \frac{\alpha H\mathcal{V}^*(s,a)}{\text{gap}(s,a)}\mathcal{LOG}(M/\delta, K, \text{gap}(s,a)) + |\mathcal{Z}_{opt}|\frac{H\mathcal{V}^*}{\text{gap}_{\min}}\mathcal{LOG}(M/\delta, K, \text{gap}_{\min})\right),$$

where $\mathcal{V}^* = \max_{(s,a)} \mathcal{V}(s,a)$, $\mathcal{Z}_{opt}$ is the set on which $\text{gap}(s,\pi^*(s)) = 0$, i.e., the set of state-action pairs assigned to π^* according to the Bellman optimality condition, and $\mathcal{Z}_{sub}$ is the complement of $\mathcal{Z}_{opt}$, and

$$\mathcal{LOG}(M/\delta, t, \breve{\text{gap}}_t(s,a)) = \log\left(\frac{M}{\delta}\right)\log\left(t \wedge 1 + \frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\text{gap}}_t(s,a)^2}\right).$$

If we take $t = K$ in Theorem 5.4.4, we have the following upper bound:

$$R(K) \leq O\left(\sum_{(s,a)\in\mathcal{Z}_{sub}} \frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta, K, \text{gap}(s,a))}{\text{gap}(s,a)} + \frac{H\mathcal{V}^*|\mathcal{S}_{opt}|\mathcal{LOG}(M/\delta, K, \text{gap}_{\min})}{\min_{k,s,a}\epsilon_k(s,a)}\right),$$

where $\mathcal{S}_{opt}$ is the set of all states for $s \in \mathcal{S}$ for which $\text{gap}(s,\pi^*(s)) = 0$ and there exists at least one state s' with $\kappa(s') < s$ for which $\text{gap}(s',\pi^*(s)) > 0$. We note that this set is no larger than the set $\mathcal{Z}_{opt}$ and further that even the smallest $\epsilon_k(s,a)$ can still be much larger than $\text{gap}_{\min}$, as it is the conditional average of the gaps. In particular, this leads to an arbitrary improvement in our example in Figure 5-1 and an improvement of SA in the example in Figure 5-2.

5.4.5 Nearly tight bounds for deterministic transition MDPs

We recall that for deterministic MDPs, $\epsilon_k(s,a) = \frac{V^*(s_1)-V^{\pi_k}(s_1)}{2H}, \forall a$ and the definition of the set $\Pi_{s,a}$:

$$\Pi_{s,a} \equiv \{\pi \in \Pi : s^\pi_{\kappa(s)} = s, a^\pi_{\kappa(s)} = a, \exists\ h \leq \kappa(s), \text{gap}(s^\pi_h, a^\pi_h) > 0\}.$$

We note that $\mathcal{V}(s,a) \leq 1$ as this is just the variance of the reward at (s,a). Theorem 5.5.11 immediately yields the following regret bound by taking $t = K$.

Corollary 5.4.5 (Explicit bound from Equation 5.5)**.** *Suppose the transition kernel of the MDP consists only of point-masses. Then with probability* $1-\delta$*,* `StrongEuler`*'s regret is bounded as*

$$R(K) \leq O\Bigg(\sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H\mathcal{LOG}\left(M/\delta, K, \overline{\mathrm{gap}}(s,a)\right)}{v^* - v^*_{s,a}}$$

$$+ \sum_{s,a} SH^3 \log\left(\frac{MK}{\delta}\right) \min\left\{\log\left(\frac{MK}{\delta}\right), \log\left(\frac{MH}{\overline{\mathrm{gap}}(s,a)}\right)\right\}$$

$$+ SAH^3(S \vee H)\log\left(\frac{M}{\delta}\right)\Bigg),$$

where $v^*_{s,a} = \max_{\pi\in\Pi_{s,a}} v^\pi$.

We now compare the above bound with the one in (Simchowitz and Jamieson, 2019) again. For simplicity we are going to take K to be the smaller of the two quantities in the logarithm. To compare the bounds, we compare $\sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H(\log(KM/\delta)))}{v^*-v^*_{(s,a)}}$ to $\sum_{(s,a)\in\mathcal{Z}_{sub}} \frac{\alpha H \log(KM/\delta)}{\mathrm{gap}(s,a)} + \frac{|\mathcal{Z}_{opt}|H}{\mathrm{gap}_{\min}}$. Recall that $\alpha \in [0,1]$ is defined as the smallest value such that for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ it holds that

$$P(s'|s,a) - P(s'|s,\pi^*(s)) \leq \alpha P(s'|s,a).$$

For any deterministic transition MDP with more than one layer and one sub-optimal action it holds that $\alpha = 1$. We will compare $V^*(s_1) - V^{\pi^*_{(s,a)}}(s_1)$ to $\mathrm{gap}(s,a) = Q^*(s,\pi^*(s)) - Q^*(s,a)$. This comparison is easy as by Lemma 5.5.1 we can write

$$V^*(s_1) - V^{\pi^*_{(s,a)}}(s_1) = \sum_{(s',a')\in\pi^*_{(s,a)}} w_{\pi^*_{(s,a)}(s',a')}\,\mathrm{gap}(s',a') = \sum_{(s',a')\in\pi^*_{(s,a)}} \mathrm{gap}(s',a') \geq \mathrm{gap}(s,a).$$

Hence, our bound in the worst case matches the one in Simchowitz and Jamieson (2019) and can actually be significantly better. We would further like to remark that we have essentially solved all of the issues presented in the example MDP in Figure 5-1.

In particular we do not pay any gap-dependent factors for states which are only visited by π^*, we do not pay a $\text{gap}_{\min}$ factor for any state and we never pay any factors for distinguishing between two suboptimal policies. Finally, we compare this bound to the lower bound derived Theorem 5.6.4 only with respect to number of episodes and gaps. Let $\mathcal{S}^*$ be the set of all states in the support of an optimal policy

$$\sum_{(s,a)\in\mathcal{S}\setminus\mathcal{S}^*\times\mathcal{A}} \frac{\log(K)}{H(v^* - v^{\pi^*_{(s,a)}}(s_1))} \le R(K) \le \sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H\log(K)}{v^* - v^*_{s,a}}.$$

The difference between the two bounds, outside of an extra H^2 factor, is in the sets $\mathcal{S}^*$ and the set $\{s, a : \Pi_{s,a} = \emptyset\}$. We note that $\{s, a : \Pi_{s,a} = \emptyset\} \subseteq \mathcal{S}^*$. Unfortunately there are examples in which $\{s, a : \Pi_{s,a} = \emptyset\}$ is $O(1)$ and $\mathcal{S}^* = \Omega(S)$ leading to a discrepancy between the upper and lower bounds of the order $\Omega(S)$. As we show in Theorem 5.6.7 this discrepancy can not really be avoided by optimistic algorithms.

5.4.6 Tighter bounds for unique optimal policy.

If we further assume that the optimal policy is unique on its support, then we can show StrongEuler will only incur regret on sub-optimal state-action pairs. This matches the information theoretic lower bound up to horizon factors. The formal regret bound is found below:

Corollary 5.4.6. *Suppose the transition kernel of the MDP consists only of point-masses and there exists a unique optimal* π^*. *Then with probability* $1-\delta$, ***StrongEuler****'s regret is bounded as*

$$R(K) \le O\Bigg(\sum_{(s,a)\notin\pi^*} \frac{\mathcal{LOG}\left(M/\delta, K, \overline{\text{gap}}(s,a)\right)}{\overline{\text{gap}}(s,a)} \\ + \sum_{(s,a)\notin\pi^*} SH^3 \log\left(\frac{MK}{\delta}\right) \min\left\{\log\left(\frac{MK}{\delta}\right), \log\left(\frac{MH}{\overline{\text{gap}}(s,a)}\right)\right\} \\ + SAH^3(S \vee H)\log\left(\frac{M}{\delta}\right)\Bigg).$$

Comparing terms which depend polynomially on $1/\overline{\text{gap}}$ to the information theoretic lower bound in Theorem 5.6.4 we observe only a multiplicative difference of H^2.

5.5 Detailed proofs for Section 5.4

5.5.1 Useful decomposition lemmas

We start by providing the following lemma that establishes that the instantaneous regret can be decomposed into gaps defined w.r.t. any optimal (and not necessarily Bellman optimal) policy.

Lemma 5.5.1 (General policy gap decomposition)**.** *Let* $\text{gap}^{\widehat{\pi}}(s,a) = V^{\widehat{\pi}}(s) - Q^{\widehat{\pi}}(s,a)$ *for any optimal policy* $\widehat{\pi} \in \Pi^*$*. Then the difference in values of* $\widehat{\pi}$ *and any policy* $\pi \in \Pi$ *is*

$$V^{\widehat{\pi}}(s) - V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} \text{gap}^{\widehat{\pi}}(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s\right] \tag{5.9}$$

and, further, the instantaneous regret of π *is*

$$v^* - v^{\pi} = \sum_{s,a} w^{\pi}(s,a)\,\text{gap}^{\widehat{\pi}}(s,a). \tag{5.10}$$

Proof. We start by establishing a recursive bound for the value difference of π and $\widehat{\pi}$ for any s

$$\begin{aligned} V^{\widehat{\pi}}(s) - V^{\pi}(s) &= V^{\widehat{\pi}}(s) - Q^{\widehat{\pi}}(s,\pi(s)) + Q^{\widehat{\pi}}(s,\pi(s)) - V^{\pi}(s) \\ &= \text{gap}^{\widehat{\pi}}(s,\pi(s)) + Q^{\widehat{\pi}}(s,\pi(s)) - Q^{\pi}(s,\pi(s)) \\ &= \text{gap}^{\widehat{\pi}}(s,\pi(s)) + \sum_{s'} P_\theta(s'|s,\pi(s))[V^{\widehat{\pi}}(s') - V^{\pi}(s')]. \end{aligned}$$

Unrolling this recursion for all layers gives

$$V^{\widehat{\pi}}(s) - V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} \text{gap}^{\widehat{\pi}}(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s\right].$$

To show the second identity, consider $s = s_1$ and note that $v^{\pi} = V^{\pi}(s_1)$ and $v^* = v^{\widehat{\pi}} = V^{\widehat{\pi}}(s_1)$ because $\widehat{\pi}$ is an optimal policy. □

For the rest of the paper we are going to focus only on the Bellman optimal policy from each state and hence only consider $\mathrm{gap}^{\widehat{\pi}}(s,a) = \mathrm{gap}(s,a)$. All of our analysis will also go through for arbitrary $\mathrm{gap}^{\widehat{\pi}}, \widehat{\pi} \in \Pi^*$, however, this did not provide us with improved regret bounds.

We now show the following technical lemma which generalizes the decomposition of value function differences and will be useful in the surplus clipping analysis.

Lemma 5.5.2. *Let $\Psi : \mathcal{S} \to \mathbb{R}$, $\Delta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be functions satisfying $\Psi(s) = 0$ for any s with $\kappa(s) = H+1$ and $\pi\colon \mathcal{S} \to \mathcal{A}$ a deterministic policy. Further, assume that the following relation holds*

$$\Psi(s) = \Delta(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \Psi\rangle,$$

and let $\mathcal{A}$ be any event that is $\mathcal{H}_h$-measurable where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \ldots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Then, for any $h \in [H]$ and $h' \in \mathbb{N}$ with $h \leq h' \leq H+1$, it holds that

$$\mathbb{E}_\pi[\chi\left(\mathcal{A}\right)\Psi(S_h))] = \mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\left(\sum_{t=h}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'+1})\right)\right] = \mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\sum_{t=h}^{H}\Delta(S_t, A_t)\right].$$

Proof. First apply the assumption of Ψ recursively to get

$$\Psi(s) = \mathbb{E}_\pi\left[\sum_{t=\kappa(s)}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'}) \;\middle|\; S_{\kappa(s)} = s\right].$$

Plugging this identity into $\mathbb{E}_\pi[\chi\left(\mathcal{A}\right)\Psi(S_h))]$ yields

$$\begin{aligned}
\mathbb{E}_\pi[\chi\left(\mathcal{A}\right)\Psi(S_h))] &= \mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\mathbb{E}_\pi\left[\sum_{t=h}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'}) \;\middle|\; S_h\right]\right]\\
&\overset{(i)}{=} \mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\mathbb{E}_\pi\left[\sum_{t=h}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'}) \;\middle|\; \mathcal{H}_h\right]\right]\\
&\overset{(ii)}{=} \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\left(\sum_{t=h}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'})\right) \;\middle|\; \mathcal{H}_h\right]\right]\\
&\overset{(iii)}{=} \mathbb{E}_\pi\left[\chi\left(\mathcal{A}\right)\left(\sum_{t=h}^{h'-1}\Delta(S_t, A_t) + \Psi(S_{h'})\right)\right]
\end{aligned}$$

where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \dots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Identity (i) holds because of the Markov-property and (ii) holds because $\mathcal{A}$ is $\mathcal{H}_h$-measurable. The final identity (iii) uses the tower-property of conditional expectations. □

5.5.2 General surplus clipping for strongly optimistic algorithms

Clipped operators. One of the main arguments to derive instance dependent bounds is to write the instantaneous regret in terms of the surpluses which are clipped to the minimum positive gap. We now define the clipping threshold $\epsilon_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ and associated clipped surpluses

$$\ddot{E}_k(s,a) = \operatorname{clip}\left[E_k(s,a) \mid \epsilon_k(s,a)\right] = \chi\left(E_k(s,a) \geq \epsilon_k(s,a)\right) E_k(s,a). \tag{5.11}$$

Next, define the clipped Q- and value-function as

$$\ddot{Q}_k(s,a) = \ddot{E}_k(s,a) + r(s,a) + \langle P(\cdot|s,a), \ddot{V}_k, \rangle \quad \text{and} \quad \ddot{V}_k(s) = \ddot{Q}_k(s, \pi_k(s)). \tag{5.12}$$

The random variable which is the state visited by π_k at time h throughout episode k is denoted by S_h and A_h is the action at time h.

Events about encountered gaps Define the event $\mathcal{E}_h = \{\operatorname{gap}(S_h, A_h) > 0\}$ that at time h an action with a positive gap played, the $\mathcal{P}_{1:h} = \bigcap_{h'=1}^{h-1} \mathcal{E}_{h'}^c$ that only actions with zero gap have been played until h and the event $\mathcal{A}_h = \mathcal{E}_h \cap \mathcal{P}_{1:h}$ that the first positive gap was encountered at time h. Let $\mathcal{A}_{H+1} = \mathcal{P}_{1:H}$ be the event that only zero gaps were encountered. Further, let

$$B = \min\{h \in [H+1] \colon \operatorname{gap}(S_h, A_h) > 0\}$$

be the first time a non-zero gap is encountered. Note that B is a stopping time w.r.t. the filtration $\mathcal{F}_h = \sigma(S_1, A_1, \dots, S_h, A_h)$.

The proof of Simchowitz and Jamieson (2019) consists of two main steps. First show that for their definition of clipped value functions one can bound $\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(\bar{V}_k(s_1) - V^{\pi_k}(s_1))$. Next, using optimism together with the fact that π_k has highest value function at episode k it follows that $\bar{V}_k(s_1) - V^{\pi_k}(s_1) \geq V^*(s_1) - V^{\pi_k}(s_1)$. The second main step is to use a high-probability bound on the clipped surpluses to relate them to the probability to visit the respective state-action pair and the proof is finished via an integration lemma. We now show that the first step can be carried out in greater generality by defining a less restrictive clipping operator. This operator is independent of the details in the definition of gap at each state-action pair but rather only uses a certain property which allows us to decompose the episodic regret as a sum over gaps. We will also further show that one does not need to use an integration lemma for the second step but can rather reformulate the regret bound as an optimization problem. This will allow us to clip surpluses at state-action pairs with zero gaps beyond the gap_{min} rate.

5.5.2.1 Clipping with an arbitrary threshold and proof of Proposition 5.4.2

Recall the definition of the clipped surpluses and clipped value function in Equation 5.11 and Equation 5.12. We begin by showing a general relation between the clipped value function difference and the non-clipped surpluses for any clipping threshold $\epsilon_k : \mathcal{S} \to \mathbb{R}$. This will help in establishing $\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(\bar{V}_k(s_1) - V^{\pi_k}(s_1))$.

Lemma 5.5.3. *Let $\epsilon_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ be arbitrary. Then for any strongly optimistic algorithm it holds that*

$$\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} (\text{gap}(S_h, A_h) - \epsilon_k(S_h, A_h))\right]. \tag{5.13}$$

Proof. We use $W_k(s) = \ddot{V}_k(s) - V^{\pi_k}(s)$ in the following and first show that $W(s_1) \geq \mathbb{E}_{\pi_k}[W_k(S_B)]$. As a precursor, we prove

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_h)\right] \geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_{h+1}\right) W_k(S_{h+1})\right] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h+1}\right) W_k(S_{h+1})\right]. \tag{5.14}$$

To see this, plug the definitions into $W_k(s)$ which gives $W_k(s) = \ddot{V}_k(s) - V^{\pi_k}(s) = \ddot{E}_k(s, \pi_k(s)) + \langle P(\cdot|s, \pi_k(s)), W_k\rangle$ and use this in the LHS of Equation 5.14 as

$$\begin{aligned}
\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_h)\right] &= \mathbb{E}_{\pi_k}\Big[\chi\left(\mathcal{P}_{1:h}\right) \underbrace{\ddot{E}_k(S_h, A_h)}_{\geq 0}\Big] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) \mathbb{E}[W_k(S_{h+1}) \mid S_h]\right] \\
&\overset{(i)}{\geq} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) \mathbb{E}_{\pi_k}[W_k(S_{h+1}) \mid \mathcal{H}_h]\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1}) \mid \mathcal{H}_h]\right] = \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1})\right]
\end{aligned}$$

where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \ldots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Step (i) follows from strong optimism and the Markov property and (ii) holds because $\mathcal{P}_{1:h}$ is $\mathcal{H}_h$-measurable. We now rewrite the RHS by splitting the expectation based on whether event $\mathcal{E}_{h+1}$ occurred as

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1})\right] = \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h+1}\right) W_k(S_{h+1})\right] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_{h+1}\right) W_k(S_{h+1})\right].$$

We have now shown Equation 5.14, which we will now use to lower-bound $W_k(s_1)$ as

$$\begin{aligned}
W_k(s_1) &= \mathbb{E}_{\pi_k}[\chi\left(\mathcal{E}_1\right) W_1(S_1)] + \mathbb{E}_{\pi_k}[\chi\left(\mathcal{E}_1^c\right) W_1(S_1)] \\
&= \mathbb{E}_{\pi_k}[\chi\left(\mathcal{A}_1\right) W_1(S_1)] + \mathbb{E}_{\pi_k}[\chi\left(\mathcal{P}_{1:1}\right) W_1(S_1)] \\
&\geq \mathbb{E}_{\pi_k}[\chi\left(\mathcal{A}_1\right) W_1(S_1)] + \sum_{h=2}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) W_k(S_h)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) W_k(S_h)\right] = \mathbb{E}_{\pi_k}[W_k(S_B)].
\end{aligned}$$

Applying Lemma 5.5.2 with $\mathcal{A} = \mathcal{A}_h$, $\Psi = W_k$ and $\Delta = \ddot{E}_k$ yields

$$\begin{aligned}
W_k(s_1) &\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \ddot{E}_k(S_{h'}, A_{h'})\right] \\
&\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \epsilon_k(S_{h'}, A_{h'})\right],
\end{aligned}$$

where we applied the definition clipped surpluses which gives $\ddot{E}_k(s, a) = \operatorname{clip}[E_k(s, a) \mid \epsilon_k(s, a)] \geq E_k(s, a) - \epsilon_k(s, a)$. It only remains to show that

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] \geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right].$$

To do so, we apply Lemma 5.5.2 twice, first with $\mathcal{A} = \mathcal{A}_h$, $\Psi = \bar{V}_k - V^{\pi_k}$ and $\Delta = E_k$ and then again with $\mathcal{A} = \mathcal{A}_h$, $\Psi = V^* - V^{\pi_k}$ and $\Delta = \text{gap}$ which gives

$$\begin{aligned}
\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] &= \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right)(\bar{V}_k(S_h) - V^{\pi_k}(S_h))\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right)(V^*(S_h) - V^{\pi_k}(S_h))\right] \\
&= \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \text{gap}(S_{h'}, A_{h'})\right].
\end{aligned}$$

Thus, we have shown that

$$\begin{aligned}
&\ddot{V}_k(s_1) - V^{\pi_k}(s_1) = W_k(s_1) \\
&\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \text{gap}(S_{h'}, A_{h'})\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \epsilon_k(S_{h'}, A_{h'})\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} (\text{gap}(S_{h'}, A_{h'}) - \epsilon_k(S_{h'}, A_{h'}))\right] \\
&= \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} (\text{gap}(S_h, A_h) - \epsilon_k(S_h, A_h))\right]
\end{aligned}$$

where the last equality uses the definition of B, the first time step at which a non-zero gap was encountered. □

Lemma 5.5.4 (Optimism of clipped value function)**.** *Let the clipping thresholds* $\epsilon_k \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ *used in the definition of* $\ddot{V}_k$ *satisfy*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h, A_h)\right] \leq \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \text{gap}(S_h, A_h)\right]$$

for some optimal policy $\widehat{\pi}$*. Then scaled optimism holds for the clipped value function, i.e.,*

$$\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)).$$

Proof. The proof works by establishing the following chain of inequalities:

$$\begin{aligned}
\frac{V^*(s_1) - V^{\pi_k}(s_1)}{2} &\overset{(a)}{=} \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right] \overset{(b)}{=} \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h))\right] \\
&\overset{(c)}{=} \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\left(\operatorname{gap}(S_h, A_h)) - \frac{1}{2}\operatorname{gap}(S_h, A_h))\right)\right] \\
&\overset{(d)}{\leq} \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}(\operatorname{gap}(S_h, A_h)) - \epsilon_k(S_h, A_h)))\right] \\
&\overset{(e)}{\leq} \ddot{V}_k(s_1) - V^{\pi_k}(s_1).
\end{aligned}$$

Here, (a) uses Lemma 5.5.1 and (b) uses the definition of B. Step (c) is just algebra and step (d) uses the assumption on the threshold function. The last step (e) follows from Lemma 5.5.3. □

We are now ready to present Proposition 5.4.2.

Proof of Proposition 5.4.2. Applying Lemma 5.5.4 which ensures scaled optimism of the clipped value function gives

$$V^*(s_1) - V^{\pi_k}(s_1) \leq 2(\ddot{V}_k(s_1) - V^{\pi_k}(s_1)) = 2\sum_{s,a} w^{\pi_k}(s,a)\ddot{E}_k(s,a),$$

where the equality follows from the definition of $\ddot{V}_k(s_1)$ and Lemma 5.5.2. Subtracting $\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1))$ from both sides gives

$$\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)) \leq 2\sum_{s,a} w^{\pi_k}(s,a)\left(\ddot{E}_k(s,a) - \frac{\operatorname{gap}(s,a)}{4}\right)$$

because Lemma 5.5.1 ensures that $\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)) = \frac{1}{2}\sum_{s,a} w^{\pi_k}(s,a)\operatorname{gap}(s,a)$. Reordering terms yields

$$\begin{aligned}
V^*(s_1) - V^{\pi_k}(s_1) &\leq 4\sum_{s,a} w^{\pi_k}(s,a)\left(\ddot{E}_k(s,a) - \frac{\operatorname{gap}(s,a)}{4}\right) \\
&= 4\sum_{s,a} w^{\pi_k}(s,a)\left(\operatorname{clip}\left[E_k(s,a)\,\middle|\,\epsilon_k(s,a)\right] - \frac{\operatorname{gap}(s,a)}{4}\right) \\
&\leq 4\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[E_k(s,a)\,\middle|\,\epsilon_k(s,a) \vee \frac{\operatorname{gap}(s,a)}{4}\right],
\end{aligned}$$

where the final inequality follows from the general properties of the clipping operator, which satisfies

$$\operatorname{clip}[a|b] - c = \begin{cases} a - c \le a & \text{for } a \ge b \vee c \\ 0 - c \le 0 & \text{for } a \le b \\ a - c \le 0 & \text{for } a \le c \end{cases} \le \operatorname{clip}[a|b \vee c].$$

□

5.5.3 Definition of valid clipping thresholds ϵ_k

Proposition 5.4.2 establishes a sufficient condition on the clipping thresholds ϵ_k that ensures that the penalized surplus clipping bounds holds. We now discuss several choices for this threshold that satisfy this condition.

Minimum positive gap $\text{gap}_{\min}$**:** We now make the quick observation that taking $\epsilon_k \equiv \frac{\text{gap}_{\min}}{2H}$ will satisfy the condition of Proposition 5.4.2, because on the event $\mathcal{B} \equiv \mathcal{A}_{H+1}^c$ there exists at least one positive gap in the sum $\sum_{h=1}^H \text{gap}(S_h, A_h)$, which, by definition, is at least $\text{gap}_{\min}$. This shows that our results already can recover the bounds in prior work, with significantly less effort.

Average gaps: Instead of the minimum gap which was used in existing analyses, we now show that we can also use the marginalized average gap which we will define now. Recall that $B = \min\{h \in [H+1] \colon \text{gap}(S_h, A_h) > 0\}$ is the first time a non-zero gap is encountered. Note that B is a stopping time w.r.t. the filtration $\mathcal{F}_h = \sigma(S_1, A_1, \dots, S_h, A_h)$. Further let

$$\mathcal{B}(s,a) \equiv \{B \le \kappa(s), S_{\kappa(s)} = s, A_{\kappa(s)} = a\} \tag{5.15}$$

be the event that (s, a) was visited after a non-zero gap in the episode. We now define this clipping threshold

$$\epsilon_k(s,a) \equiv \begin{cases} \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \text{gap}(S_h, A_h) \,\middle|\, \mathcal{B}(s,a)\right] & \text{if } \mathbb{P}_{\pi_k}(\mathcal{B}(s,a)) > 0 \\ \infty & \text{otherwise} \end{cases} \tag{5.16}$$

As the following lemma shows, this is a valid choice which satisfies the condition of Proposition 5.4.2.

Lemma 5.5.5. *The expected sum of clipping thresholds in Equation* (5.16) *over all state-action pairs encountered after a positive gap is at most half the expected total gaps per episode. That is,*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h, A_h)\right] \leq \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right].$$

Proof. We rewrite the LHS of the inequality to show as $\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \chi\left(B \leq h\right) \epsilon_k(S_h, A_h)\right]$ and from now on consider the random variable $f_h(B, S_h, A_h) = \chi\left(B \leq h\right) \epsilon_k(S_h, A_h)$ where $f_h(b, s, a) = \chi\left(b \leq h\right) \epsilon_k(s, a)$ is a deterministic function[2]. We will show below that $\mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \leq \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h)\right]$. This is sufficient to prove the statement, because

$$\begin{aligned}
\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h, A_h)\right] &= \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \\
&\leq \frac{1}{2H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right] \\
&= \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h)\right] = \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right].
\end{aligned}$$

To bound the expected value of $f_h(B, S_h, A_h)$, we first write f_h for all triples b, s, a such that $\mathbb{P}_{\pi_k}(B = b, A_h = a, S_h = s) > 0$ as

$$\begin{aligned}
f_h(b, s, a) &\overset{(i)}{=} \chi\left(b \leq h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \,\middle|\, B \leq h,\ S_h = s, A_h = a\right] \\
&\overset{(ii)}{=} \chi\left(b \leq h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \,\middle|\, B \leq h,\ S_h = s, A_h = a\right] \\
&\quad + \chi\left(b \leq h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \,\middle|\, S_h = s, A_h = a\right],
\end{aligned}$$

where (i) expands the definition of ϵ_k and (ii) decomposes the sum inside the conditional expectation and uses the Markov-property to simplify the conditioning for terms after

[2] It may still depend on the current policy π_k which is determined by observations in episodes 1 to $k-1$. But, crucially, f_h does not depend on any realization in the k-th episode

h. Before taking the expectation of $f_h(B, S_h, A_h)$, we first rewrite the conditional expectation in the first term above, which will be useful later.

$$
\begin{aligned}
&\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; B \leq h,\ S_h = s, A_h = a\right] \\
&\overset{(i)}{=} \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \chi\left(A_h = a, S_h = s\right) \chi\left(B \leq h\right)\right]}{\mathbb{P}_{\pi_k}\left[B \leq h,\ S_h = s, A_h = a\right]} \\
&\overset{(ii)}{=} \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \chi\left(A_h = a, S_h = s\right)\right]}{\mathbb{P}_{\pi_k}\left[B \leq h,\ S_h = s, A_h = a\right]} \\
&= \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h = s, A_h = a\right]}{\mathbb{P}_{\pi_k}\left[B \leq h \;\mid\; S_h = s, A_h = a\right]}.
\end{aligned}
$$

Here, step (i) uses the property of conditional expectations with respect to an event with nonzero probability and (ii) follows from the definition of B: When $B > h$, the sum of gaps until h is zero. Consider now the expectation of $f_h(B, S_h, A_h)$

$$
\begin{aligned}
&\mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \leq h\right) \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \leq h \;\mid\; S_h, A_h\right]}\right] \qquad (5.17) \\
&\quad + \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \leq h\right) \mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\, S_h, A_h\right]\right] \qquad (5.18)
\end{aligned}
$$

The term in (5.18) can be bounded using the tower-property of expectations as

$$
\begin{aligned}
&\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \leq h\right) \mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\, S_h, A_h\right]\right] \\
&\leq \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\, S_h, A_h\right]\right] = \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right].
\end{aligned}
$$

For the term in (5.17), we also use the tower-property to rewrite it as

$$\begin{aligned}
&\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right)\frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h}\operatorname{gap}(S_{h'},A_{h'})\;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\mid\; S_h, A_h\right]}\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right)\frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h}\operatorname{gap}(S_{h'},A_{h'})\;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\mid\; S_h, A_h\right]}\;\middle|\; S_h, A_h\right]\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right)\;\middle|\; S_h, A_h\right]\frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h}\operatorname{gap}(S_{h'},A_{h'})\;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\mid\; S_h, A_h\right]}\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h}\operatorname{gap}(S_{h'},A_{h'})\;\middle|\; S_h, A_h\right]\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h}\operatorname{gap}(S_{h'},A_{h'})\right].
\end{aligned}$$

Summing both terms yields the required upper-bound $\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\operatorname{gap}(S_h,A_h)\right]$ on the expectation $\mathbb{E}_{\pi_k}\left[f_h(B,S_h,A_h)\right]$. □

5.5.4 Policy-dependent regret bound for StrongEuler

We now show how to derive a regret bound for StrongEuler algorithm in Simchowitz and Jamieson (2019) that depends on the gaps of the played policies throughout the K episodes.

To build on parts of the analysis in Simchowitz and Jamieson (2019), we first define some useful notation analogous to Simchowitz and Jamieson (2019) but adapted to our setting:

$$\begin{aligned}
\bar{n}_k(s,a) &= \sum_{j=1}^{k} w^{\pi_k}(s,a), \\
M &= (SAH)^3, \\
\mathcal{V}^{\pi}(s,a) &= \mathbb{V}[R(s,a)] + \mathbb{V}_{s'\sim P(\cdot|s,a)}[V^{\pi}(s')], \\
\mathcal{V}_k(s,a) &= \mathcal{V}^{\pi_k}(s,a) \wedge \mathcal{V}^{*}(s,a)
\end{aligned}$$

We will use their following results:

Proposition 5.5.6 (Proposition F.1, F.9 and B.4 in Simchowitz and Jamieson (2019))**.** *There is a good event $\mathcal{A}^{\text{conc}}$ that holds with probability $1-\delta/2$. In this event,* StrongEuler *is strongly optimistic (as well as optimistic). Further, there is a universal constant $c \geq 1$ so that for all $k \geq 1$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, the surpluses are bounded as*

$$0 \leq \frac{1}{c} E_k(s,a) \leq B_k^{\text{lead}}(s,a) + \sum_{h=\kappa(s)}^{H} \mathbb{E}_{\pi_k}\left[B_k^{\text{fut}}(S_h, A_h) \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right],$$

where B^{lead}, B^{fut} are defined as

$$B_k^{\text{lead}}(s,a) = H \wedge \sqrt{\frac{\mathcal{V}_k(s,a)\log\left(Mn_k(s,a)/\delta\right)}{n_k(s,a)}},$$
$$B_k^{\text{fut}}(s,a) = H^3 \wedge H^3\left(\sqrt{\frac{S\log\left(Mn_k(s,a)/\delta\right)}{n_k(s,a)}} + \frac{S\log\left(Mn_k(s,a)/\delta\right)}{n_k(s,a)}\right)^2.$$

Lemma 5.5.7 (Lemma B.3 in Simchowitz and Jamieson (2019))**.** *Let $m \geq 2$, $a_1, \dots, a_m \geq 0$ and $\epsilon \geq 0$. Then* $\operatorname{clip}\left[\sum_{i=1}^{m} a_i \middle| \epsilon\right] \leq 2\sum_{i=1}^{m} \operatorname{clip}\left[a_i \middle| \frac{\epsilon}{2m}\right]$.

Equipped with these results and our improved surplus clipping proposition in Proposition 5.5.6, we can now derive the following bound on the regret of StrongEuler

Lemma 5.5.8. *In event $\mathcal{A}^{\text{conc}}$, the regret of* StrongEuler *is bounded for all $k \geq 1$ as*

$$\begin{aligned} R(K) \leq & 8\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) \operatorname{clip}\left[cB_k^{\text{lead}}(s,a) \;\middle|\; \frac{\breve{\text{gap}}_k(s,a)}{4}\right] \\ & + 16\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) \operatorname{clip}\left[cB_k^{\text{fut}}(s,a) \;\middle|\; \frac{\breve{\text{gap}}_k(s,a)}{8SA}\right], \end{aligned}$$

with a universal constant $c \geq 1$ and $\breve{\text{gap}}_k(s,a) = \frac{\text{gap}(s,a)}{4} \vee \epsilon_k(s,a)$.

Proof. We now use our improved surplus clipping result from Proposition 5.4.2 as a starting point to bound the instantaneous regret of StrongEuler in the kth episode as

$$V^*(s_1) - V^{\pi_k}(s_1) \le 4 \sum_{s,a} w^{\pi_k}(s,a) \operatorname{clip}\left[E_k(s,a) \,\middle|\, \breve{\operatorname{gap}}_k(s,a) \right]. \tag{5.19}$$

Next, we write the bound on the surpluses from Proposition 5.5.6 as

$$\begin{aligned} &E_k(s,a) \le cB_k^{\text{lead}}(s,a) \\ &\qquad + c\sum_{s',a'} \chi\left(\kappa(s') \ge \kappa(s)\right) \mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right] B_k^{\text{fut}}(s',a') \end{aligned}$$

and plugging it in Equation 5.19 and applying Lemma 5.5.7 gives

$$\begin{aligned} V^*(s_1) - V^{\pi_k}(s_1) \le\ & 8\sum_{s,a} w^{\pi_k}(s,a) \operatorname{clip}\left[cB_k^{\text{lead}}(s,a) \,\middle|\, \frac{\breve{\operatorname{gap}}_k(s,a)}{4} \right] \\ & + 16 \sum_{s,a} w^{\pi_k}(s,a) \operatorname{clip}\left[cB_k^{\text{fut}}(s,a) \,\middle|\, \frac{\breve{\operatorname{gap}}_k(s,a)}{8SA} \right]. \end{aligned}$$

The statement to show follows now by summing over $k \in [K]$. The form of the second term in the previous display follows from the inequality

$$\begin{aligned} &\sum_{s,a} w^{\pi_k}(s,a) \chi\left(\kappa(s') \ge \kappa(s)\right) \mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right] \\ &\le \sum_{s,a} w^{\pi_k}(s,a) \mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right] = w^{\pi_k}(s',a'). \end{aligned}$$

□

We note that if $\pi_k \equiv \widehat{\pi}$ for any $\widehat{\pi} \in \Pi^*$ then $V^*(s_1) - V^{\pi_k}(s_1) = 0$, and WLOG we can disregard such terms in the total regret.

The next step is to relate $\bar{n}_k(s,a)$ to $n_k(s,a)$ via the following lemma.

Lemma 5.5.9 (Lemma B.7 in Simchowitz and Jamieson (2019))**.** *Define the event* $\mathcal{A}^{\text{samp}}$

$$\mathcal{A}^{\text{samp}} = \left\{ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall k \ge \tau(s,a) \colon n_k(s,a) \ge \frac{\bar{n}_k(s,a)}{4} \right\},$$

where $\tau(s,a) = \inf\{k : \bar{n}_k(s,a) \ge H_{\text{samp}}\}$ *and* $H_{\text{samp}} = c' \log(M/\delta)$ *for a universal constant* c'*. Then event* $\mathcal{A}^{\text{samp}}$ *holds with probability* $1 - \delta/2$*.*

Proof. This can be proved analogously to Lemma B.7 in Simchowitz and Jamieson (2019) and Lemma 6 in Dann et al. (2019) with the difference that in our case, there can only be at most one observation of (s,a) per episode for each (s,a) due to our layered assumption. Thus, there is no need to sum over observations accumulated for each $h \in [H]$ and our $H_{\text{samp}} = O(\log(H))$ as opposed to $O(H\log(H))$. □

Lemma 5.5.10. *Let $f_{s,a} \colon \mathbb{N} \to \mathbb{R}$ be non-increasing with $\sup_u f_{s,a}(u) \le \widehat{f} < \infty$ for all $s,a \in \mathcal{S} \times \mathcal{A}$. Then on event $\mathcal{A}^{\text{samp}}$ in Lemma 5.5.9, we have*

$$\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \le SA\widehat{f}H_{\text{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4).$$

Proof.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \\
&= \sum_{s,a}\sum_{k=1}^{\tau(s,a)-1} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \\
&\le \sum_{s,a}\left(\sum_{k=1}^{\tau(s,a)-1} w^{\pi_k}(s,a)\right)\widehat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4) \\
&= \sum_{s,a} n_{\tau(s,a)}(s,a)\widehat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4) \\
&\le SAH_{\text{samp}}\widehat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4).
\end{aligned}$$

□

Theorem 5.5.11 (Regret Bound for StrongEuler). *With probability at least $1-\delta$, the regret of StrongEuler is bounded for all number of episodes $K \in \mathbb{N}$ as*

$$\begin{aligned}
R(K) \lesssim \sum_{s,a} \min_{t \in [K_{(s,a)}]} &\left\{ \frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta, t, \breve{\text{gap}}_t(s,a))}{\breve{\text{gap}}_t(s,a)} \right. \\
&\left. + \sqrt{(K_{(s,a)} - t)\mathcal{LOG}(M/\delta, K_{(s,a)}, \breve{\text{gap}}_{K_{(s,a)}}(s,a))} \right\} \\
&+ \sum_{s,a} SH^3 \log\left(\frac{MK}{\delta}\right) \min\left\{\log\left(\frac{MK}{\delta}\right), \log\left(\frac{MH}{\breve{\text{gap}}_{\min}(s,a)}\right)\right\} \\
&+ SAH^3(S \vee H)\log\left(\frac{M}{\delta}\right).
\end{aligned}$$

Here, $K_{(s,a)}$ is the last round during which a policy π was played such that $w^{\pi}(s,a) > 0$, $\breve{\mathrm{gap}}_t(s,a) = \mathrm{gap}(s,a) \vee \epsilon_t(s,a)$, $\breve{\mathrm{gap}}_{\min}(s,a) = \min_{k\in[K]:\ \breve{\mathrm{gap}}_k(s,a)>0} \breve{\mathrm{gap}}_k(s,a)$ is the smallest gap encountered for each (s,a), and

$$\mathcal{LOG}(M/\delta, t, \breve{\mathrm{gap}}_t(s,a)) = \log\left(\frac{M}{\delta}\right)\log\left(t \wedge 1 + \frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\mathrm{gap}}_t(s,a)^2}\right).$$

Proof. We here consider the event $\mathcal{A}^{\mathrm{conc}} \cap \mathcal{A}^{\mathrm{samp}}$ which has probability at least $1-\delta$ by Proposition 5.5.6 and Lemma 5.5.9. We now start with the regret bound in Lemma 5.5.8 and bound the two terms individually in the following:

Bounding the B^{lead} term. We have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{lead}}(s,a) \;\middle|\; \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right] \\
&\overset{(i)}{\leq} SAHH_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)\,\mathrm{clip}\left[c\sqrt{\frac{4\mathcal{V}_k(s,a)\log(M\bar{n}_k(s,a)/4\delta)}{\bar{n}_k(s,a)}} \;\middle|\; \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right] \\
&\overset{(ii)}{\leq} SAHH_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K_{(s,a)}} w^{\pi_k}(s,a)\mathrm{clip}\left[2c\sqrt{\mathcal{V}^*(s,a)\log\left(\frac{M}{\delta}\right)}\sqrt{\frac{\log(\bar{n}_k(s,a))}{\bar{n}_k(s,a)}} \;\middle|\; \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right],
\end{aligned} \tag{5.20}$$

where step (i) applies Lemma 5.5.10 and (ii) follows from the definition of $\mathcal{V}_k(s,a)$, the definition of $K_{(s,a)}$ and

$$\begin{aligned}
&\log\left(\frac{M\bar{n}_k(s,a)}{4\delta}\right) = \log\left(\frac{M}{4\delta}\right) + \log(\bar{n}_k(s,a)) \\
&\leq \left(\log\left(\frac{M}{4\delta}\right)+1\right)\log(\bar{n}_k(s,a)) = \log\left(\frac{Me}{4\delta}\right)\log(\bar{n}_k(s,a)) \leq \log(M/\delta)\log(\bar{n}_k(s,a)).
\end{aligned}$$

We now apply our optimization lemma (Lemma 5.5.14) with $x_k = w^{\pi_k}(s,a)$, $v_k = 2c\sqrt{\mathcal{V}^*(s,a)\log(M/\delta)}$, and $\epsilon_k = \frac{\breve{\mathrm{gap}}_k(s,a)}{4v_k}$ to bound each (s,a)-term in Equation 5.20 for any $t \in [K]$ as

$$\begin{aligned}
&4\frac{v_t}{\epsilon_t}\log\left(t \wedge 1 + \frac{1}{\epsilon_t^2}\right) + 4v_t\sqrt{\log\left(K \wedge 1 + \frac{1}{\epsilon_K^2}(K-t)\right)} \\
&= \frac{32c^2\mathcal{V}^*(s,a)\log\left(\frac{M}{\delta}\right)\log\left(t \wedge 1 + \frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\mathrm{gap}}_t(s,a)^2}\right)}{\breve{\mathrm{gap}}_t(s,a)} \\
&+8c\sqrt{(K-t)\mathcal{V}^*(s,a)\log\left(\frac{M}{\delta}\right)\log\left(K \wedge 1 + \frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\mathrm{gap}}_K(s,a)^2}\right)}.
\end{aligned}$$

We have

$$\begin{aligned}&\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)\,\mathrm{clip}\left[2c\sqrt{\mathcal{V}^*(s,a)\log(M/4\delta)}\sqrt{\frac{\log(\bar{n}_k(s,a))}{\bar{n}_k(s,a)}} \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right]\\ \le&\frac{32c^2\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta,t,\breve{\mathrm{gap}}_t(s,a))}{\breve{\mathrm{gap}}_t(s,a)} + 8c\sqrt{(K-t)\mathcal{LOG}(M/\delta,K,\breve{\mathrm{gap}}_K(s,a))}.\end{aligned}$$

Plugging this bound back in Equation 5.20 gives

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{lead}}(s,a) \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right]\\ &\lesssim SAH\log\left(\frac{M}{\delta}\right)\\ &+\sum_{s,a}\min_{t\in[K_{(s,a)}]}\left\{\frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta,t,\breve{\mathrm{gap}}_t(s,a))}{\breve{\mathrm{gap}}_t(s,a)} + \sqrt{(K_{(s,a)}-t)\mathcal{LOG}(M/\delta,K,\breve{\mathrm{gap}}_{K_{(s,a)}}(s,a))}\right\}\end{aligned}$$

where $\lesssim$ only ignores absolute constant factors.

Bounding the B^{fut} term. Consider the second term in Lemma 5.5.8 and event $\mathcal{A}^{\mathrm{conc}} \cap \mathcal{A}^{\mathrm{samp}}$. Then by Lemma 5.5.10

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a) \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{8SA}\right]\\ &\le SAH^3H_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)f_{s,a}(\bar{n}_k(s,a))\end{aligned}$$

where $f_{s,a}$ is

$$\begin{aligned}&f_{s,a}(\bar{n}_k(s,a)) =\\ &\mathrm{clip}\left[2cH^3 \wedge 2cH^3\left(\sqrt{\frac{S\log(M\bar{n}_k(s,a)/\delta)}{\bar{n}_k(s,a)}} + \frac{S\log(M\bar{n}_k(s,a)/\delta)}{\bar{n}_k(s,a)}\right)^2 \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right].\end{aligned}$$

We now apply Lemma C.1 by Simchowitz and Jamieson (2019) which gives

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a) \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{8SA}\right]\\ &\le SAH^3H_{\mathrm{samp}} + \sum_{s,a}Hf_{s,a}(H) + \sum_{s,a}\int_{H}^{\bar{n}_K(s,a)} f_{s,a}(u)du\\ &\le SAH^4c'\log(M/\delta) + \sum_{s,a}\int_{H}^{\bar{n}_K(s,a)} f_{s,a}(u)du.\end{aligned}$$

The remaining integral term is bounded with Lemma B.9 (b) by Simchowitz and Jamieson (2019) with $C' = S, C = H^3$ and $\epsilon = \check{\mathrm{gap}}_{\,\mathrm{min}}(s,a) = \min_{k\in[K_{(s,a)}]:\ \check{\mathrm{gap}}_{\,k}(s,a)>0} \check{\mathrm{gap}}_{\,k}(s,a)$ as follows.

$$\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a)\ \middle|\ \frac{\check{\mathrm{gap}}_{\,k}(s,a)}{8SA}\right]$$
$$\lesssim SAH^4 \log\left(\frac{M}{\delta}\right) + \sum_{s,a}\left(SH^3\log\left(\frac{M}{\delta}\right)\right.$$
$$\left. + SH^3\log\left(\frac{MK}{\delta}\right)\min\left\{\log\left(\frac{MK}{\delta}\right), \log\left(\frac{MH}{\check{\mathrm{gap}}_{\,\mathrm{min}}(s,a)}\right)\right\}\right)$$
$$\lesssim SAH^3(S\vee H)\log\left(\frac{M}{\delta}\right) + \sum_{s,a} SH^3\log\left(\frac{MK}{\delta}\right)\min\left\{\log\left(\frac{MK}{\delta}\right), \log\left(\frac{MH}{\check{\mathrm{gap}}_{\,\mathrm{min}}(s,a)}\right)\right\}.$$

□

5.5.5 Proof of Corollary 5.4.6

We begin by showing a different type of upper bound on the expected gaps by the surpluses. Define the set $\beta_k = range(B)$ where B is the r.v. which is the stopping time with respect to π_k. For any π^*, define the set

$$\begin{aligned}\mathcal{O}_k(\pi^*) = \bigcup_{s_b\in\beta_k}\{(s,a)\in\mathcal{S}\times\mathcal{A} :&\mathbb{P}_{\pi^*}((S_h,A_h)=(s,a)|S_{\kappa(s_b)}=s_b)\\ &\geq \mathbb{P}_{\pi_k}((S_h,A_h)=(s,a)|S_{\kappa(s_b)}=s_b)\}.\end{aligned}$$

This set has the following intuitive definition – whenever $\mathcal{A}_B$ occurs we restrict our attention to the MDP with initial state S_B. On this restricted MDP, $\mathcal{O}_k$ is the set of state-action pairs which have greater probability to be visited by the optimal π^* than by π_k.

Lemma 5.5.12. *Assume strong optimism and greedy $\bar{V}_k$, i.e., $\bar{V}_k(s) \geq \max_a \bar{Q}_k(s,a)$ for all $s\in\mathcal{S}$. Then there exists an optimal π^* for which*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\mathrm{gap}(S_h,A_h)\right] \leq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\chi(S_h,A_h\notin\mathcal{O}_k(\pi^*))E_k(S_h,A_h)\right].$$

Proof. One can write the optimistic value function for any s and π as follows

$$\begin{aligned}\bar{V}^{\pi}(s) &= \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} E_k(S_h, A_h) + r(S_h, A_h)\Big|S_{\kappa(s)} = s\right]\\ &= E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}^{\pi}\rangle.\end{aligned}$$

By backwards induction on H we show that for any s, $\kappa(s) \leq H$ $\bar{V}^{\pi} \leq \bar{V}_k$. The base case holds from the fact that on all $s : \kappa(s) = H$, $\bar{V}_k(s)$ is just the largest optimistic reward over all actions at s. For the induction step it holds that

$$\begin{aligned}\bar{V}^{\pi}(s) &= E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}^{\pi}\rangle\\ &\leq E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}_k\rangle\\ &= \bar{Q}_k(s, \pi(s)) \leq \bar{V}_k(s),\end{aligned}$$

where the first inequality holds from the induction hypothesis and the second inequality holds by definition of the value function. We now have

$$\begin{aligned}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \mathrm{gap}(S_h, A_h)\right] &= \mathbb{E}_{\pi_k}\left[V^*(S_B) - V_k(S_B)\right]\\ &\leq \mathbb{E}_{\pi_k}\left[\bar{V}_k(S_B) - V_k(S_B)\right] - \mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right].\end{aligned}$$

Let us focus on the term $\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right]$

$$\begin{aligned}\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right] &= \mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)|S_B\right]\right]\\ &= \mathbb{E}_{\pi_k}\left[\sum_{s} \frac{\bar{V}^*(s) - V^*(s)}{\mathbb{P}_{\pi_k}(S_B = s)}\chi(S_B = s)\right]\\ &= \mathbb{E}_{\pi_k}\left[\sum_{s} \frac{\mathbb{E}_{\pi^*}\left[\sum_{h=\kappa(s)}^{H} E_k(S_h, A_h)|S_{\kappa(s)} = s\right]}{\mathbb{P}_{\pi_k}(S_B = s)}\chi(S_B = s)\right].\end{aligned}$$

We can similarly expand the term $\mathbb{E}_{\pi_k}\left[\bar{V}_k(S_B) - V_k(S_B)\right]$. By the definition of $\mathcal{O}_k(\pi^*)$ it holds that for any $h \geq \kappa(s)$

$$\begin{aligned}&\mathbb{E}_{\pi_k}\left[E_k(S_h, A_h)|S_{\kappa(s)} = s\right] - \mathbb{E}_{\pi^*}\left[E_k(S_h, A_h)|S_{\kappa(s)} = s\right]\\ &\qquad\leq \mathbb{E}_{\pi_k}\left[\chi(S_h, A_h \notin \mathcal{O}_k(\pi^*))E_k(S_h, A_h)|S_{\kappa(s)} = s\right].\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right] \\
&\leq \mathbb{E}_{\pi_k}\left[\sum_s \frac{\mathbb{E}_{\pi_k}\left[\sum_{h=\kappa(s)}^H \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h) | S_{\kappa(s)} = s\right]}{\mathbb{P}_{\pi_k}(S_B = s)} \chi(S_B = s)\right] \\
&= \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h)\right].
\end{aligned}$$

□

We next show a version of Lemma 5.5.3 which takes into account the set $\mathcal{O}_k(\pi^*)$.

Lemma 5.5.13. *With the same assumptions as in Lemma 5.5.12, there exists an optimal π^* for which*

$$\ddot{V}_k(s_1) - V_k(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \operatorname{gap}(S_h, A_h) - \sum_{h=B}^H \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) \epsilon_k(S_h, A_h)\right],$$

where ϵ_k is arbitrary.

Proof. Since $\ddot{E}_k$ is non-negative on all state-action pairs we have

$$\begin{aligned}
\ddot{V}_k(s_1) - V^{\pi_k}(s_1) &= \mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \ddot{E}_k(S_h, A_h)\right] \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \ddot{E}_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \chi\left((S_h, A_h) \notin \mathcal{O}_k\right) \ddot{E}_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \chi((S_h, A_h) \notin \mathcal{O}_k) E_k(S_h, A_h)\right] \\
&\quad - \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \chi((S_h, A_h) \notin \mathcal{O}_k) \epsilon_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \operatorname{gap}(S_h, A_H)\right] - \mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \chi((S_h, A_h) \notin \mathcal{O}_k) \epsilon_k(S_h, A_h)\right],
\end{aligned}$$

where the second to last inequality follows from the definition of $\ddot{E}_k$ and the last inequality follows from Lemma 5.5.12. □

Next, we define $\bar{\epsilon}_k$ in the following way. Let

$$\bar{\epsilon}_k(s, a) \equiv \begin{cases} \epsilon_k(s, a) & \text{if } (s, a) \notin \mathcal{O}_k(\pi^*) \\ \infty & \text{otherwise,} \end{cases} \tag{5.21}$$

where ϵ_k is the clipping function defined in Equation 5.16. Lemma 5.5.13 now implies that

$$\ddot{V}_k(s_1) - V_k(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h) - \sum_{h=B}^{H} \bar{\epsilon}_k(S_h, A_h)\right].$$

This is sufficient to argue Lemma 5.5.8 with $\breve{\operatorname{gap}}_k(s,a) = \frac{\operatorname{gap}(s,a)}{4} \vee \bar{\epsilon}_k(s,a)$ and hence arrive at a version of Corollary 5.4.5 which uses $\bar{\epsilon}_k$ as the clipping thresholds. Let us now argue that $\bar{\epsilon}_k(s,a) = \infty$ for all $(s,a) \in \pi^*$ whenever π^* is the unique optimal policy for the deterministic MDP. To do so consider $(s,a) \in \pi^*$ and $\pi_k \neq \pi^*$. Since the MDP is deterministic, β_k is a singleton and is the the first state s_b at which π_k differs from π^*. We now observe that if $\kappa(s) < \kappa(s_b)$, this implies $\epsilon_k(s,a) = \infty$ as $\mathcal{B}(s,a)$ does not occur. Further, the conditional probabilities $\mathbb{P}_{\pi^*}((S_h, A_h) = (s,a)|S_{\kappa(s_b)} = s_b)$ and $\mathbb{P}_{\pi_k}((S_h, A_h) = (s,a)|S_{\kappa(s_b)} = s_b)$ are both equal to 1 if $\kappa(s) > \kappa(s_b)$ and so $(s,a) \in \mathcal{O}_k(\pi^*)$ which implies $\bar{\epsilon}_k(s,a) = \infty$. Thus we can clip all gaps at $(s,a) \in \pi^*$ to infinity and they will never appear in the regret bound.

5.5.6 Alternative to integration lemmas

Lemma 5.4.3 is a simplified version of the following stronger result:

Lemma 5.5.14. *Consider the following optimization problem*

$$\begin{aligned} &\underset{x_1,\dots,x_K}{\mathit{maximize}} && \sum_{k=1}^{K} \frac{v_k x_k \sqrt{\log\left(\sum_{j=1}^{k} x_j\right)}}{\sqrt{\sum_{j=1}^{k} x_j}} \\ &\mathit{s.t.} && 1 \leq x_1, \quad 0 \leq x_k \leq 1, \quad \frac{\sqrt{\log\left(\sum_{j=1}^{k} x_j\right)}}{\sqrt{\sum_{j=1}^{k} x_j}} \geq \epsilon_k \qquad \forall\ k \in [K], \end{aligned} \tag{5.22}$$

with $(v_i)_{i\in[K]} \in \mathbb{R}_+^K$ and $(\epsilon_i)_{i\in[K]} \in \mathbb{R}_+^K$. Then the optimal value of Problem 5.22 is bounded for any $t \in [K]$ as

$$4\frac{\bar{v}_t}{\epsilon_t}\log\left(t \wedge 1 + \frac{1}{\epsilon_t^2}\right) + 4v_t^*\sqrt{\log\left(K \wedge 1 + \frac{1}{\epsilon_K^2}\right)(K-t)}, \tag{5.23}$$

where $\bar{v}_t = \max_{k\in[t]} v_k$ and $v_t^* = \max_{K\geq k\geq t} v_k$.

Proof. Denote by $X_k = \sum_{t=1}^{k} x_t$ the cumulative sum of x_t. The proof consists of splitting the objective of Equation 5.22 into two terms:

$$\sum_{k=1}^{t} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} + \sum_{k=t+1}^{K} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} \tag{5.24}$$

and bounding each by the corresponding one in Equation 5.23 respectively.

Before doing so, we derive the following bound on the sum of $\frac{x_k}{\sqrt{X_k}}$ terms:

$$\sum_{k=m+1}^{M} \frac{x_k}{\sqrt{X_k}} = \sum_{k=m+1}^{M} \frac{X_k - X_{k-1}}{\sqrt{X_k}} \leq \int_{X_m}^{X_M} \frac{1}{\sqrt{x}}\, dx = 2(\sqrt{X_M} - \sqrt{X_m})\,, \tag{5.25}$$

where the inequality is due to X_k being non-decreasing.

Consider now each term in the objective in Equation 5.24 separately.

Summands up to t: Since X_k is non-decreasing, we can bound

$$\begin{aligned}\sum_{k=1}^{t} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} \leq \bar{v}_t \sqrt{\log(X_t)} \sum_{k=1}^{t} \frac{x_k}{\sqrt{X_k}} &\overset{(i)}{\leq} 2\bar{v}_t \sqrt{\log(X_t)}\sqrt{X_t} \\ &\overset{(ii)}{\leq} 2\frac{\bar{v}_t}{\epsilon_t} \log(X_t)\,,\end{aligned}$$

where (i) follows from Equation 5.25 using the convention $X_0 = 0$ and (ii) from the optimization constraint $\sqrt{\log(X_t)} \geq \epsilon_t \sqrt{X_t}$. It remains to bound $\log(X_t)$ by $2\log\left(t \wedge 1 + \frac{1}{\epsilon_t^2}\right)$. Since all increments x_j are at most 1, the bound $\log(X_t) \leq \log(t)$ holds.

We claim the following:

Claim 5.5.15. *For any x s.t. $\log(x) \leq \log(\log(x)/a)$ it holds that $\log(x) \leq 2\log(1 + 1/a)$.*

Proof. First, we note that if $0 < x \leq e$, then $\log(\log(x)) < 0$ and thus the assumption of the claim implies $\log(x) \leq \log(1/a)$. Next, assume that $x > e$. Then we have $\frac{\log(\log(x))}{\log(x)} \leq 1/e$, which together with the assumption of the claim implies $\log(x) \leq 1/e \log(x) + \log(1/a)$ or equivalently $\log(x) \leq \frac{e}{e-1}\log(1/a)$. Noting that $e/(e-1) \leq 2$ completes the proof. □

The constraints of the problem enforce $\sqrt{X_k} \leq \frac{\sqrt{\log(X_k)}}{\epsilon_k}$, which implies after squaring and taking the logarithm: $\log(X_k) \leq \log(\log(X_k)/\epsilon_k^2)$. Thus, using Claim 5.5.15 yields:

$$\log(X_k) \leq 2\log\left(k \wedge 1 + 1/\epsilon_k^2\right). \tag{5.26}$$

Summands larger than t: Let $v_t^* = \max_{k:\, t<k\leq K} v_k$. For this term, we have

$$\begin{aligned}
\sum_{k=t+1}^{K} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} &\overset{5.26}{\leq} 2v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)} \sum_{k=t+1}^{K} \frac{x_k}{\sqrt{X_k}} \\
&\overset{5.25}{\leq} 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{X_K} - \sqrt{X_t}) \\
&\leq 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{X_K - X_t}) \\
&\leq 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{K - t}),
\end{aligned}$$

where we first bounded $\log(X_k) \leq \log(X_K)$, because X_k is non-decreasing, and used the upper bound on $\log(X_K)$. Then we applied Equation 5.25 and finally used $0 \leq x_k \leq 1$. □

5.6 Instance-dependent lower bounds

We here shed light on what properties on an episodic MDP determine the statistical difficulty of RL by deriving information-theoretic lower bounds on the asymptotic expected regret of any (good) algorithm. To that end, we first derive a general result that expresses a lower bound as the optimal value of a certain optimization problem and then derive closed-form lower-bounds from this optimization problem that depend on certain notions of gaps for two special cases of episodic MDPs.

Specifically, in those special cases, we assume that the rewards follow a Gaussian distribution with variance $1/2$. We further assume that the optimal value function is bounded in the same range as individual rewards, e.g. as $0 \leq V^*(s) < 1$ for all $s \in \mathcal{S}$. This assumption is common in the literature (e.g. Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018) and can be considered harder than a normalization of $V^*(s) \in [0, H]$ (see Jiang and Agarwal (2018)).

5.6.1 General instance-dependent lower bound as an optimization problem

The idea behind deriving instance-dependent lower bounds for the stochastic MAB problem (Lai and Robbins, 1985; Combes et al., 2017; Garivier et al., 2019) and infinite horizon MDPs (Graves and Lai, 1997; Ok et al., 2018) are based on first assuming that the algorithm studied is ***uniformly good***, that is, on any instance of the problem and for any $\alpha > 0$, the algorithm incurs regret at most $o(T^\alpha)$, and then argue that, to achieve that guarantee, the algorithm must select a certain policy or action at least some number of times as it would otherwise not be able to distinguish the current MDP from another MDP that requires a different optimal strategy.

Since comparison between different MDPs is central to lower-bound constructions, it is convenient to make the problem-instance explicit in the notation. To that end, let Θ be the problem class of possible MDPs and we use subscripts θ and λ for value functions, return, MDP parameters etc., to denote specific problem instances $\theta, \lambda \in \Theta$ of those quantities. Further, for a policy π and MDP θ, $\mathbb{P}_\theta^\pi$ denotes the law of one episode, i.e., the distribution of $(S_1, A_1, R_1, S_2, A_2, R_2, \dots, S_{H+1})$. To state the general regret lower-bound we need to introduce the set of ***confusing*** MDPs. This set consists of all MDPs λ in which there is at least one optimal policy π such that $\pi \notin \Pi_\theta^*$, i.e., π is not optimal for the original MDP and no policy in Π_θ^* has been changed.

Definition 5.6.1. For any problem instance $\theta \in \Theta$ we define the set of confusing MDPs $\Lambda(\theta)$ as

$$\Lambda(\theta) := \{\lambda \in \Theta \colon \Pi_\lambda^* \setminus \Pi_\theta^* \neq \varnothing \text{ and } KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = 0 \ \forall \pi \in \Pi_\theta^*\}.$$

We are now ready to state our general regret lower-bound for episodic MDPs:

Theorem 5.6.1 (General instance-dependent lower bound for episodic MDPs)**.** *Let ψ be a uniformly good RL algorithm for Θ, that is, for all problem instances $\theta \in \Theta$ and exponents $\alpha > 0$, the regret of ψ is bounded as $\mathbb{E}[R_\theta(K)] \leq o(K^\alpha)$, and assume that $v^*_\theta < H$. Then, for any $\theta \in \Theta$, the regret of ψ satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[R_\theta(K)]}{\log(K)} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem

$$\begin{aligned} &\underset{\eta(\pi)\geq 0}{\textit{minimize:}} && \sum_{\pi\in\Pi} \eta(\pi)\,(v^*_\theta - v^\pi_\theta) \\ &\textit{subject to:} && \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta) \end{aligned}. \tag{5.27}$$

The optimization problem in Theorem 5.6.1 can be interpreted as follows. The variables $\eta(\pi)$ are the (expected) number of times the algorithm chooses to play policy π which makes the objective the total expected regret incurred by the algorithm. The constraints encode that any uniformly good algorithm needs to be able to distinguish the true instance θ from all confusing instances $\lambda \in \Lambda(\theta)$, because otherwise it would incur linear regret. To do so, a uniformly good algorithm needs to play policies π that induce different behavior in λ and θ which is precisely captured by the constraints $\sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \geq 1$.

Although Theorem 5.6.1 has the flavor of results in the bandit and RL literature, there are a few notable differences. Compared to lower-bounds in the infinite-horizon MDP setting (Graves and Lai, 1997; Tewari and Bartlett, 2008; Ok et al., 2018), we for example do not assume that the Markov chain induced by an optimal policy π^* is irreducible. That irreducability plays a key role in converting the semi-infinite linear program Equation 5.27, which typically has uncountably many constraints, into a linear program with only $O(SA)$ constraints. While for infinite horizon MDPs, irreducibility is somewhat necessary to facilitate exploration, this is not the case for the finite horizon setting and in general we cannot obtain a convenient reduction of the set of constraints $\Lambda(\theta)$. See Section 5.7.2 for a more in-depth discussion.

5.6.2 Gap-dependent lower bound when optimal policies visit all states

To derive closed-form gap-dependent bounds from the general optimization problem Equation 5.27, we need to identify a finite subset of confusing MDPs $\Lambda(\theta)$ that each require the RL agent to play a distinct set of policies that do not help to distinguish the other confusing MDPs. To do so, we restrict our attention to the special case of MDPs where every state is visited with non-zero probability by some optimal policy, similar to the irreducibility assumptions in the infinite-horizon setting (Tewari and Bartlett, 2008; Ok et al., 2018). In this case, it is sufficient to raise the expected immediate reward of a suboptimal (s, a) by $\operatorname{gap}_\theta(s, a)$ in order to create a confusing MDP, as proven in the following lemma.

Lemma 5.6.2. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1 and let $\theta \in \Theta$ be an MDP in this class. Then for any suboptimal state-action pair (s, a) with $\operatorname{gap}_\theta(s, a) > 0$ that is visited by some optimal policy with non-zero probability, there exists a confusing MDP $\lambda \in \Lambda(\theta)$ with*

- *λ and θ only differ in the immediate reward at (s, a)*
- *$KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \leq \operatorname{gap}_\theta(s, a)^2$ for all $\pi \in \Pi$.*

By relaxing the problem in Equation 5.27 to only consider constraints from the confusing MDPs in Lemma 5.6.2 with $KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \leq \operatorname{gap}_\theta(s, a)^2$, for every (s, a), we can derive the following closed-form bound:

Theorem 5.6.3 (Gap-dependent lower bound when optimal policies visit all states)**.** *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance where every state is visited by some optimal policy with non-zero probability. Then any uniformly good* *algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[R_\theta(K)]}{\log(K)} \geq \sum_{s,a\colon\ \mathrm{gap}_\theta(s,a)>0} \frac{1}{\mathrm{gap}_\theta(s,a)}.$$

Theorem 5.6.3 can be viewed as a generalization of Proposition 2.2 in Simchowitz and Jamieson (2019), which gives a lower bound of order $\sum_{s,a\colon\ \mathrm{gap}_\theta(s,a)>0} \frac{H}{\mathrm{gap}_\theta(s,a)}$ for a certain set of MDPs.[3] While our lower bound is a factor of H worse, it is significantly more general and holds in any MDP where optimal policies visit all states and with appropriate normalization of the value function. Theorem 5.6.3 indicates that value-function gaps characterize the instance-optimal regret when optimal policies cover the entire state space.

5.6.3 Gap-dependent lower bound for deterministic-transition MDPs

We expect that optimal policies do not visit all states in most MDPs of practical interest (e.g. because certain parts of the state space can only be reached by making an egregious error). We therefore now consider the general case where $\bigcup_{\pi\in\Pi^*_\theta} supp(\pi) \subsetneq \mathcal{S}$ but restrict our attention to MDPs with deterministic transitions where we are able to give an intuitive closed-form lower bound. Note that deterministic transitions imply $\forall \pi, s, a \colon\ w^\pi(s,a) \in \{0,1\}$. Here, a confusing MDP can be created by simply raising the reward of any (s,a) by

$$v^*_\theta - \max_{\pi\colon\ w^\pi_\theta(s,a)>0} v^\pi_\theta\ , \tag{5.28}$$

the regret of the best policy that visits (s,a), as long as it is positive and (s,a) is not visited by any optimal policy. Equation 5.28 is positive when no optimal policy visits (s,a) in which case suboptimal actions have to be taken to reach (s,a) and $\overline{\mathrm{gap}}_\theta(s,a) > 0$. Let $\pi^*_{(s,a)}$ be any maximizer in Equation 5.28, which has to act

[3] We translated their bound and construction to our setting where $V^* \leq 1$ which reduces the bound by a factor of H.

optimally after visiting (s, a). From the regret decomposition in Equation 5.3 and the fact that $\pi^*_{(s,a)}$ visits (s, a) with probability 1, it follows that $v^*_\theta - v^{\pi^*_{(s,a)}}_\theta \geq \mathrm{gap}_\theta(s, a)$. We further have $v^*_\theta - v^{\pi^*_{(s,a)}}_\theta \leq H\overline{\mathrm{gap}}_\theta(s, a)$. Equipped with the subset of confusing MDPs λ that each raise the reward of a single (s, a) as $r_\lambda(s, a) = r_\theta(s, a) + v^*_\theta - v^{\pi^*_{(s,a)}}_\theta$, we can derive the following gap-dependent lower bound:

Theorem 5.6.4. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance with deterministic transitions. Then any uniformly good algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[R_\theta(K)]}{\log{(K)}} \geq \sum_{s,a \in \mathcal{Z}_\theta \colon \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{1}{H \cdot (v^*_\theta - v^{\pi^*_{(s,a)}}_\theta)} \geq \sum_{s,a \in \mathcal{Z}_\theta \colon \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{1}{H^2 \cdot \overline{\mathrm{gap}}_\theta(s,a)},$$

where $\mathcal{Z}_\theta = \{(s, a) \in \mathcal{S} \times \mathcal{A} \colon \forall \pi^* \in \Pi^*_\theta \quad w^{\pi^*}_\theta(s, a) = 0\}$ is the set of state-action pairs that no optimal policy in θ visits.

We now compare the above lower bound to the upper bound guaranteed by StrongEuler in Equation 5.5. The comparison is only with respect to number of episodes and gaps[4]

$$\sum_{s,a \in \mathcal{X} \colon \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{\log{(K)}}{H^2 \overline{\mathrm{gap}}_\theta(s,a)} \leq \mathbb{E}_\theta[R(K)] \leq \sum_{s,a \colon \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{\log{(K)}}{\overline{\mathrm{gap}}_\theta(s,a)}.$$

The difference between the two bounds, besides the extra H^2 factor, is the fact that (s, a) pairs that are visited by any optimal policy $(s, a \neq \mathcal{Z}_\theta)$ do not appear in the lower-bound while the upper-bound pays for such pairs if they can also be visited after playing a suboptimal action. This could result in cases where the number of terms in the lower bound is $O(1)$ but the number of terms in the upper bound is $\Omega(SA)$ leading to a large discrepancy.

[4] We carry out the comparison in expectation, since our lower bounds do not apply with high probability.

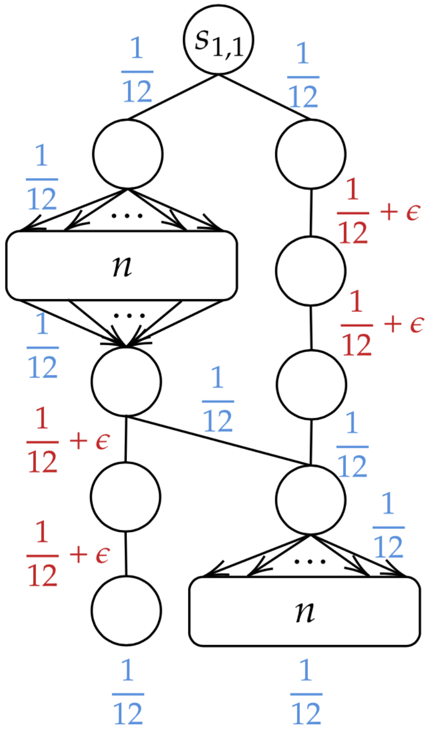


Figure 5-3: Deterministic MDP instance for optimistic lower bound

5.6.4 Lower bounds for optimistic algorithms in MDPs with deterministic transitions

In this section we show a lower bound on the regret of optimistic algorithms, demonstrating that optimistic algorithms can not hope to achieve the information-theoretic lower bounds even if the MDPs have deterministic transitions. While the result might seem similar to the one proposed by Simchowitz and Jamieson (2019) (Theorem 2.3) we would like to emphasize that the construction of Simchowitz and Jamieson (2019) does not apply to MDPs with deterministic transitions, and that the idea behind our construction is significantly different.

Consider the MDP in Figure 5-3. This MDP has $2n+9$ states and $4n+8$ actions. The rewards for each action are either $1/12$ or $1/12+\epsilon/2$ and can be found next to the transitions from the respective states. We are going to label the states according to their layer and their position in the layer so that the first state is $s_{1,1}$ the state which is to the left of $s_{1,1}$ in layer 2 is $s_{2,1}$ and to the right $s_{2,2}$. In general the i-th state in layer h is denoted as $s_{h,i}$. The rewards in all states are deterministic, with a single exception of a Bernoulli reward from state $s_{4,1}$ to $s_{5,2}$ with mean $1/12$. From the construction it is clear that $V^*(s_{1,1}) = 1/2 + \epsilon$. Further there are two sets of optimal policies with the above value function – the n optimal policies which visit state $s_{2,2}$ and the n optimal policies which visit $s_{5,1}$. Notice that the information-theoretic lower bound for this MDP is in $O(\log(K)/\epsilon)$ as only the transition from state $s_{4,1}$ to $s_{5,2}$ does not belong to an optimal policy. In particular, there is no dependence on n. Next we try to show that the class of optimistic algorithms will incur regret at least $\Omega(n \log(\delta^{-1})/\epsilon)$.

Class of algorithms. We adopt the class of algorithms from Section G.2 in (Simchowitz and Jamieson, 2019) with an additional assumption which we clarify momentarily. Recall that the class of algorithms assumes access to an optimistic value function $\bar{V}_k(s) \geq V^*(s)$ and optimistic Q-functions. In particular the algorithms construct optimistic Q and value functions as

$$\bar{V}_k(s) = \max_{a \in \mathcal{A}} \bar{Q}_k(s,a)$$
$$Q_k(s,a) = \widehat{r}_k(s,a) + b_k^{rw}(s,a) + \widehat{p}_k(s,a)^\top \bar{V}_k + b_k(s,a).$$

We assume that there exists a $c \geq 1$ such that

$$\frac{c}{2}\sqrt{\frac{\log(M(1 \vee n_k(s,a)))/\delta}{(1 \vee n_k(s,a))}} \leq b_k^{rw}(s,a) \leq c\sqrt{\frac{\log(M(1 \vee n_k(s,a)))/\delta}{(1 \vee n_k(s,a))}},$$

where $M = \theta(n)$ and $b_k(s,a) \sim \sqrt{S} f_k(s,a) b_k^{rw}(s,a)$, where f_k is a decreasing function in the number of visits to (s,a) given by $n_k(s,a)$. For $n_k(s,a) = \Omega(n \log(n))$, we assume $b_k(s,a) \leq b_k^{rw}(s,a)$. One can verify that this is true for the the Q and value functions of StrongEuler.

Lower bound. Let $\epsilon > 0$ be sufficiently small to be specified later and let N be such that

$$N = \lfloor \frac{c^2 n \log\left(MN/(n\delta)\right)}{16\epsilon^2} \rfloor .$$

Lemma 5.6.5. *There exists n_0, ϵ_0 such that for any pair of $n \geq n_0$ and $\epsilon \leq \epsilon_0$ and any $k \leq N$, with probability at least $1-\delta$, it holds that either $n_k(s_{5,1}) < N/4$, or $\bar{Q}_k(s_{4,1},1) < \bar{Q}_k(s_{4,1},2)$.*

We can show the same for the upper part of the MDP.

Lemma 5.6.6. *There exists n_0, ϵ_0 such that for any pair of $n \geq n_0$ and $\epsilon \leq \epsilon_0$ and any $k \leq N$, with probability at least $1-\delta$, it holds that either $n_k(s_{1,2}) < N/4$, or $\bar{Q}_k(s_{1,1},2) < \bar{Q}_k(s_{1,1},1)$.*

Theorem 5.6.7. *There exists an MDP instance with deterministic transitions on which any optimistic algorithm with confidence parameter δ will incur expected regret of at least $\Omega(S\log\left(\delta^{-1}\right)/\epsilon)$ while it is asymptotically possible to achieve $\Omega(\log\left(K\right)/\epsilon)$ regret.*

Proof of Theorem 5.6.7. Taking the MDP from Figure 5-3. Applying Lemma 5.6.5 and 5.6.6 shows that after N episodes with probability at least $1-2\delta$, the visitation count of $s_{2,2}$ and $s_{5,1}$ each do not exceed $N/4$. Hence there are at least $N/2$ episodes in which neither of them is visited, which means an ϵ-suboptimal policy is taken. Hence the expected regret after N episodes is at least

$$(1-2\delta)\epsilon N/2 = \Omega\left(\frac{S\log\left(\delta^{-1}\right)}{\epsilon}\right) .$$

□

Theorem 5.6.7 has two implications for optimistic algorithms in MDPs with deterministic transitions.

- It is impossible to be asymptotically optimal if the confidence parameter δ is tuned to the time horizon K.

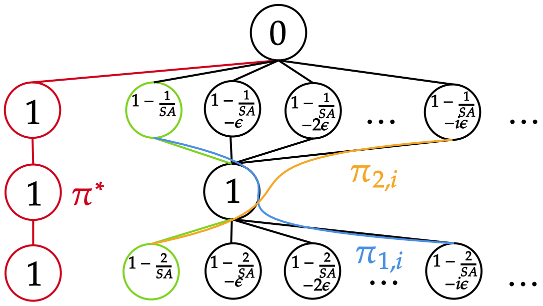

0
1
$1-\frac{1}{SA}$
$1-\frac{1}{SA}-\epsilon$
$1-\frac{1}{SA}-2\epsilon$
…
$1-\frac{1}{SA}-i\epsilon$
…
1
π^*
1
$\pi_{2,i}$
1
$1-\frac{2}{SA}$
$1-\frac{2}{SA}-\epsilon$
$1-\frac{2}{SA}-2\epsilon$
$\pi_{1,i}$
…
$1-\frac{2}{SA}-i\epsilon$
…

Figure 5-4: Issue with restricting LP to Π^*

- It is impossible to have an anytime bound matching the information-theoretic lower bound.

5.6.4.1 Issue with deriving a general bound

We now try to give some intuition regarding why we could not derive a generic lower bound for deterministic transition MDPs. We have already outlined our general approach of restricting the set Π and $\Lambda(\theta)$ to finite subsets of manageable size and then showing that the value of the LP on these restricted sets is not much smaller than the value of the original LP. One natural restriction of Π is the set Π^* from Theorem 5.6.4. Suppose we restrict ourselves to the same set and consider only environments making policies in Π^* optimal as the restriction for $\Lambda(\theta)$. We now give an example of an MDP for which such a restriction will lead to an $\Omega(SA)$ multiplicative discrepancy between the value of the original semi-infinite LP and the restricted LP. The MDP can be found in Figure 5-4. The rewards for each action for a fixed state s are equal and are shown in the vertices corresponding to the states. The number of states in the second and last layer of the MDP are equal to $(SA-3)/2$. The optimal policy takes the red path and has value $V^{\pi^*}=3$. The set Π^* consists of all policies $\pi_{j,i}$ which visit one of the states in green. The policies $\pi_{1,i}$, in blue, visit the green state in the second layer of the MDP and one of the states in the final layer, following the paths in blue. Similarly the policies $\pi_{2,i}$, in orange, visit one of the state in the second layer and the green state in the last layer, following the orange paths. The value function of $\pi_{j,i}$ is $V^{\pi_{j,i}} = 3 - \frac{3}{SA} - i\epsilon$, where $0 \leq i \leq (SA-4)/2$. We claim that playing each $\pi_{j,i}$ $\eta(\pi_{j,i}) = \Omega(SA)$ times is a feasible solution to the LP restricted to Π^*. Fix i, the $\lambda_{\pi_{1,i}}$ must put weight at least $1/SA$ on the green state in layer 2. Coupling with the fact that for all i' the rewards $\pi_{1,i'}$ are also changed under this environment we know that the constraint of the restricted LP with respect to $\lambda_{\pi_{1,i}}$ is lower bounded by $\sum_{i'} \eta(\pi_{1,i'})/(SA)^2$. Since there are $\Omega(SA)$ policies $\{\pi_{1,i'}\}_{i'}$, this implies that $\eta(\pi_{1,i}) = \Omega(SA)$ is feasible. A similar argument holds for any $\pi_{2,i}$. Thus the value of the restricted LP is at most $O(SA)$, for any $\epsilon \ll SA$.

However, we claim that the value of the semi-infinite LP which actually characterizes the regret is at least $\Omega(S^2A^2)$. First, to see that the above assignment of η is not feasible for the semi-infinite LP, consider any policy $\pi \not\in \Pi^*$, e.g. take the policy which visits the state in layer 2 with reward $1 - 1/SA - \epsilon$ and the state in layer 4 with reward $1 - 2/SA - \epsilon$. Each of these states have been visited $O(SA)$ times and $\eta(\pi) = 0$ hence the constraint for the environment λ_π is upper bounded by $SA\left(\left(\frac{1}{SA} + \epsilon\right)^2 + \left(\left(\frac{2}{SA} + \epsilon\right)^2\right)\right) \approx 1/SA$. In general each of the states in black in the second layer and the fourth layer have been visited $1/SA$ times less than what is necessary to distinguish any $\pi \not\in \Pi^*$ as sub-optimal. If we define the i-th column of the MDP as the pair consisting of the states with rewards $1 - 1/SA - i\epsilon$ and $1 - 2/SA - i\epsilon$ then to distinguish the policy visiting both of these states as sub-optimal we need to visit at least one of these $\Omega(S^2A^2)$ times. This implies we need to visit each column of the MDP $\Omega(S^2A^2)$ times and thus any strategy must incur regret at least $\Omega\left(\sum_i S^2A^2 \frac{1}{SA}\right) = \Omega(S^2A^2)$, leading to the promised multiplicative gap of $\Omega(SA)$ between the values of the two LPs.

Why does such a gap arise and how can we hope to fix it this issue? Any feasible solution to the LP restricted to Π^* essentially needs to visit the states in green $\Theta(S^2A^2)$ times. This is sufficient to distinguish the green states as sub-optimal to visit and hence any strategy visiting these states would be also deemed sub-optimal. This is achievable by playing each strategy in Π^* in the order of $\Theta(SA)$ times as already discussed. Now, even though Π^* covers all other states, from our argument above we see that we need to play each $\pi \in \Pi^*$ in the order of $\Theta(S^2A^2)$ times to be able to determine all sub-optimal states. To solve this issue, we either have to increase the size of Π^* to include for example all policies visiting each column of the MDP or at the very least include changes of environments in the constraint set which make such policies optimal. This is clearly computationally feasible for the MDP in Figure 5-4, however, it is not clear how to proceed for general MDPs, without having to include exponentially many constraints. The following interesting questions arises: is it possible to come up with a relaxation of the LP, computable in polynomial time, with solution which is at most $o(SA)$ multiplicative constant away from what is optimal. We show in Section 5.7.3.2 that this is indeed possible for the special case of deterministic transition, tree-structured MDPs.

5.7 Proofs from Section 5.6

Let $N_{\psi,\pi}(k)$ be the random variable denoting the number of times policy π has been chosen by the strategy ψ. Let $N_{\psi,(s,a)}(k)$ be the number of times the state-action pair has been visited up to time k by the strategy ψ.

5.7.1 Proof of Theorem 5.6.1

We begin by formulating an LP characterizing the minimum regret incurred by any uniformly good algorithm ψ.

Theorem 5.7.1. *Let ψ be a uniformly good RL algorithm for Θ, that is, for all problem instances $\theta \in \Theta$ and exponents $\alpha > 0$, the regret of ψ is bounded as $\mathbb{E}[R_\theta(K)] \leq o(K^\alpha)$. Then, for any $\theta \in \Theta$, the regret of ψ satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[R_\theta(K)]}{\log{(K)}} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem

$$\begin{aligned} &\underset{\eta(\pi)\geq 0}{\textit{minimize:}} && \sum_{\pi\in\Pi} \eta(\pi)\left(v_\theta^* - v_\theta^\pi\right) \\ &\textit{subject to:} && \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1 \qquad \textit{for all} \ \ \lambda \in \Lambda(\theta)\end{aligned}, \tag{5.29}$$

where $\Lambda'(\theta) = \{\lambda \in \Theta \colon \Pi_\lambda^* \cap \Pi_\theta^* = \varnothing, KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*}) = 0\}$ are all environments that share no optimal policy with θ and do not change the rewards or transition kernel on π^*.

Proof. We can write the expected regret as $\mathbb{E}[R_\theta(K)] = \sum_{\pi\in\Pi} \mathbb{E}_\theta[N_{\psi,\pi}(K)](v_\theta^* - v_\theta^\pi)$. We will show that $\eta(\pi) = \mathbb{E}_\theta[N_{\psi,\pi}(K)]/\log(K)$ is feasible for the optimization problem in Equation 5.27. This is sufficient to prove the theorem. To do so we follow the techniques of Garivier et al. (2019). With slight abuse of notation, let $\mathbb{P}_\theta^{I_k}$ be the law of all trajectories up to episode k. We have

$$\begin{aligned} KL(\mathbb{P}_\theta^{I_{k+1}}, \mathbb{P}_\lambda^{I_{k+1}}) &= KL(\mathbb{P}_\theta^{Y_{k+1},I_k}, \mathbb{P}_\lambda^{Y_{k+1},I_k}) \\ &= KL(\mathbb{P}_\theta^{I_k}, \mathbb{P}_\lambda^{I_k}) + \mathbb{E}\left[\mathbb{E}_{\mathbb{P}_\theta^{\psi(I_k)}}\left[\log\left(\frac{\mathbb{P}_\theta^{\psi(I_k)}(Y_{k+1})}{\mathbb{P}_\lambda^{\psi(I_k)}(Y_{k+1})}\right) \,\middle|\, I_k\right]\right] \\ &= KL(\mathbb{P}_\theta^{I_k}, \mathbb{P}_\lambda^{I_k}) + \mathbb{E}\left[\sum_{\pi\in\Pi} \chi(\psi(I_k) = \pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi)\right].\end{aligned} \tag{5.30}$$

Iterating the argument we arrive at $\sum_{\pi\in\Pi} \mathbb{E}_\theta[N_{\psi,\pi}(K)] KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = KL(\mathbb{P}_\theta^{I_K}, \mathbb{P}_\lambda^{I_K})$ where $\mathbb{E}_\theta$ denotes expectation in problem instance θ. Next one shows that for any measurable $Z \in [0,1]$, it holds that $KL(\mathbb{P}_\theta^{I_K}, \mathbb{P}_\lambda^{I_K}) \geq kl(\mathbb{E}_\theta[Z], \mathbb{E}_\lambda[Z])$ where $kl(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ denotes the KL-divergence between two Bernoulli random variables p and q. This follows directly from Lemma 1 by Garivier et al. (2019). Finally we choose $Z = N_{\psi,\Pi_\lambda^*}(K)/K$ as the fraction of episodes where an optimal policy for λ was played (here we use the short-hand notation $N_{\psi,\Pi_\lambda^*}(K) = \sum_{\pi\in\Pi_\lambda^*} N_{\psi,\pi}(K)$). Evaluating the kl-term we have

$$kl\left(\frac{\mathbb{E}_\theta[N_{\psi,\Pi_\lambda^*}(K)]}{K}, \frac{\mathbb{E}_\lambda[N_{\psi,\Pi_\lambda^*}(K)]}{K}\right) \geq \left(1 - \frac{\mathbb{E}_\theta[N_{\psi,\Pi_\lambda^*}(K)]}{K}\right)\log\left(\frac{K}{K - \mathbb{E}_\lambda[N_{\psi,\Pi_\lambda^*}(K)]}\right) - \log(2)\,.$$

Since ψ is a uniformly good it follows that for any $\alpha > 0$, $K - \mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)] = o(K^\alpha)$. By assuming that $\Pi^*_\theta \cap \Pi^*_\lambda = \varnothing$, we get $\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)] = o(K)$. This implies that for K sufficiently large and all $1 \geq \alpha > 0$

$$kl\left(\frac{\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)]}{K}, \frac{\mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)]}{K}\right) \geq \log(K) - \log(K^\alpha) = (1-\alpha)\log(K) \xrightarrow{\alpha\to 0} \log(K)\,.$$

□

The set $\Lambda'(\theta)$ is uncountably infinite for any reasonable Θ we consider. What is worse the constraints of LP 5.27 will not form a closed set and thus the value of the optimization problem will actually be obtained on the boundary of the constraints. To deal with this issue it is possible to show the following.

Proof of Theorem 5.6.1. For the rest of this proof we identify $\Lambda'(\theta) = \{\lambda \in \Theta : \Pi^*_\lambda \cap \Pi^*_\theta = \emptyset, KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0, \forall \pi^*_\theta \in \Pi^*_\theta\}$ as the set from Theorem 5.7.1 and $\tilde{\Lambda}(\theta) = \{\lambda \in \Theta : v^{\pi^*_\lambda}_\lambda \geq v^{\pi^*_\theta}_\theta, \pi^*_\lambda \notin \Pi^*_\theta, KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0\}$. From the proof of Theorem 5.7.1 it is clear that we can rewrite $\Lambda'(\theta)$ as the union $\bigcup_{\pi\in\Pi} \Lambda_\pi(\theta)$, where $\Lambda_\pi(\theta) = \{\lambda \in \Theta : KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0, v^{\pi^*_\lambda} > v^{\pi^*_\theta}_\theta, \pi^*_\lambda = \pi\}$ is the set of all environments which make π the optimal policy. This implies that we can equivalently write LP 5.27 as

$$\begin{aligned} &\underset{\eta(\pi)\geq 0}{\text{minimize:}} && \sum_{\pi\in\Pi} \eta(\pi)\left(v^*_\theta - v^\pi_\theta\right) \\ &\text{subject to:} && \inf_{\lambda\in\Lambda_{\pi'}(\theta)} \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \geq 1 \qquad \text{for all} \quad \pi' \in \Pi \end{aligned}. \tag{5.31}$$

The above formulation now minimizes a linear function over a finite intersection of sets, however, these sets are still slightly inconvenient to work with. We are now going to try to make these sets more amenable to the proof techniques we would like to use for deriving specific lower bounds. We begin by noting that $\Lambda_\pi(\theta)$ is bounded in the following sense. We identify each λ with a vector in $[0,1]^{S^2A} \times [0,1]^{SA}$ where the first S^2A coordinates are transition probabilities and the last SA coordinates are the expected rewards. From now on we work with the natural topology on $[0,1]^{S^2A} \times [0,1]^{SA}$, induced by the ℓ_1 norm. Further, we claim that we can assume that

$KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda})$ is a continuous function over $\Lambda_{\pi'}(\theta)$. The only points of discontinuity are at λ for which the support of the transition kernel induced by λ does not match the support of the transition kernel induced by θ. At such points the $KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda}) = \infty$. This implies that such λ does not achieve the infimum in the set of constraints so we can just restrict $\Lambda_{\pi'}(\theta)$ to contain only λ for which $KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda}) < \infty$. With this restriction in hand the KL-divergence is continuous in λ.

Fix a π' and consider the set $\{\eta : \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda}) \geq 1\}$ corresponding to one of the constraints in LP 5.31. Denote $\tilde{\Lambda}_{\pi'}(\theta) = \{\lambda \in \Theta : KL(\mathbb{P}^{\pi^*_\theta}_{\theta}, \mathbb{P}^{\pi^*_\theta}_{\lambda}) = 0, v^{\pi^*_\lambda}_{\lambda} \geq v^{\pi^*_\theta}_{\theta}, \pi^*_\lambda \notin \Pi^*_\theta, \pi^*_\lambda = \pi'\}$. $\tilde{\Lambda}_{\pi'}(\theta)$ is closed as $KL(\mathbb{P}^{\pi^*_\theta}_{\theta}, \mathbb{P}^{\pi^*_\theta}_{\lambda})$ and $v^{\pi^*_\lambda}_{\lambda} - v^{\pi^*_\theta}_{\theta}$ are both continuous in λ. To see the statement for $v^{\pi^*_\lambda}_{\lambda}$, notice that this is the maximum over the continuous functions v^{π}_{λ} over $\pi \in \Pi$. Take any $\eta \in \Lambda_{\pi'}(\theta)$ and let $\{\lambda_j\}_{j=1}^{\infty}, \lambda_j \in \Lambda_{\pi'}(\theta)$ be a sequence of environments such that $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda_j}) \geq 1 + 2^{-j}$. If there is no convergent subsequence of $\{\lambda_j\}_{j=1}^{\infty}$ in $\Lambda_{\pi'}(\theta)$ we claim it is because of the constraint $v^{\pi^*_\lambda}_{\lambda} > v^{\pi^*_\theta}_{\theta}$. Take the limit λ of any convergent subsequence of $\{\lambda_j\}_{j=1}^{\infty}$ in the closure of $\Lambda_{\pi'}(\theta)$. Then by continuity of the divergence we have $0 = \lim_{j \to \infty} KL(\mathbb{P}^{\pi^*_\theta}_{\theta}, \mathbb{P}^{\pi^*_\theta}_{\lambda_j}) = KL(\mathbb{P}^{\pi^*_\theta}_{\theta}, \mathbb{P}^{\pi^*_\theta}_{\lambda})$, thus it must be the case that $v^{\pi^*_\lambda}_{\lambda} \leq v^{\pi^*_\theta}_{\theta}$. This shows that $\tilde{\Lambda}_{\pi'}(\theta)$ is a subset of the closure of $\Lambda_{\pi'}(\theta)$ which implies it is the closure of $\Lambda_{\pi'}(\theta)$, i.e., $\bar{\Lambda}_{\pi'}(\theta) = \tilde{\Lambda}_{\pi'}(\theta)$.

Next, take $\eta \in \{\eta : \min_{\lambda \in \bar{\Lambda}_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda}) \geq 1\}$ and let $\lambda_{\pi',\eta}$ be the environment on which the minimum is achieved. Such $\lambda_{\pi',\eta}$ exists because we just showed that $\bar{\Lambda}_{\pi'}(\theta)$ is closed and bounded and hence compact and the sum consists of a finite number of continuous functions. If $\lambda_{\pi',\eta} \in \Lambda_{\pi'}(\theta)$ then $\eta \in \{\eta : \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda}) \geq 1\}$. If $\lambda_{\pi',\eta} \notin \Lambda_{\pi'}(\theta)$ then $\lambda_{\pi',\eta}$ must be a limit point of $\Lambda_{\pi'}(\theta)$. By definition we can construct a convergent sequence of $\{\lambda_j\}_{j=1}^{\infty}, \lambda_j \in \Lambda_{\pi'}(\theta)$ to $\lambda_{\pi',\eta}$ such that $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda_j}) \geq 1$. This implies $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda_j}) \geq \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\lambda})$. Using the continuity of the KL term and taking limits, the above implies that the minimum upper bounds the infimum. Since we argued that $\Lambda_{\pi'}(\theta)$ is bounded and $\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda_j})$ is also bounded from below this implies $\bar{\Lambda}_{\pi'}(\theta)$ contains the infimum $\inf_{\lambda\in\Lambda_{\pi'}(\theta)}\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})$. This implies $\inf_{\lambda\in\Lambda_{\pi'}(\theta)}\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})\geq\min_{\lambda\in\bar{\Lambda}_{\pi'}(\theta)}\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})$, and so the infimum over $\Lambda_{\pi}(\theta)$ equals the minimum over $\bar{\Lambda}_{\pi}(\theta)$. Which finally implies that $\eta\in\{\eta:\inf_{\lambda\in\Lambda_{\pi'}(\theta)}\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})\geq 1\}$. This shows that LP 5.31 is equivalent to

$$\begin{aligned}
&\underset{\eta(\pi)\geq 0}{\text{minimize:}} && \sum_{\pi\in\Pi}\eta(\pi)\left(v^*_\theta-v^\pi_\theta\right)\\
&\text{subject to:} && \min_{\lambda\in\bar{\Lambda}_{\pi'}(\theta)}\sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})\geq 1 \qquad \text{for all}\ \ \pi'\in\Pi'
\end{aligned}$$

or equivalently that we can consider the closure of $\Lambda(\theta)$ in LP 5.27, $\bar{\Lambda}(\theta)=\{\lambda\in\Theta\colon v_\lambda^{\pi^*_\lambda}\geq v_\theta^{\pi^*_\theta},\pi^*_\lambda\notin\Pi^*_\theta,KL(\mathbb{P}_\theta^{\pi^*_\theta},\mathbb{P}_\lambda^{\pi^*_\theta})=0\}$ i.e. the set of environments which makes any π optimal without changing the environment on state-action pairs in π^*_θ. □

5.7.2 Proof of Theorem 5.6.3

Proof of Lemma 5.6.2. Let λ be the environment that is identical to θ except for the immediate reward for state-action pair for (s,a). Specifically, let $R_\lambda(s,a)$ so that $r_\lambda(s,a)=r_\theta(s,a)+\Delta$ with $\Delta=\mathrm{gap}_\theta(s,a)$. Since we assume that rewards are Gaussian, it follows that

$$\begin{aligned}
KL(\mathbb{P}^{\pi}_{\theta},\mathbb{P}^{\pi}_{\lambda})&=w^{\pi}_{\lambda}(s,a)KL(R_\theta(s,a),R_\lambda(s,a))\leq KL(R_\theta(s,a),R_\lambda(s,a))\\
&\leq\mathrm{gap}_\theta(s,a)^2
\end{aligned}$$

for any policy $\pi\in\Pi$. We now show that the optimal value function (and thus return) of λ is uniformly upper-bounded by the optimal value function of θ. To that end, consider their difference in any state s', which we will upper-bound by their difference in s as

$$\begin{aligned}
V^*_\lambda(s')-V^*_\theta(s')&\leq\chi(\kappa(s)>\kappa(s'))\mathbb{P}_\theta^{\pi^*_\lambda}(s_{\kappa(s)}=s|s_{\kappa(s')}=s')[V^*_\lambda(s)-V^*_\theta(s)]\\
&\leq V^*_\lambda(s)-V^*_\theta(s).
\end{aligned}$$

Further, the difference in s is exactly

$$\begin{aligned} V_\lambda^*(s) - V_\theta^*(s) &= r_\lambda(s,a) + \langle P_\theta(\cdot|s,a), V_\theta^* \rangle - V_\theta^*(s) \\ &= r_\theta(s,a) + \langle P_\theta(\cdot|s,a), V_\theta^* \rangle + \operatorname{gap}_\theta(s,a) - V_\theta^*(s) = 0. \end{aligned}$$

Hence, $V_\lambda^* = V_\theta^* \leq 1$ and thus $\lambda \in \Theta$. We will now show that there is a policy that is optimal in λ but not in θ. Let $\pi^* \in \Pi_\theta^*$ be any optimal policy for θ that has non-zero probability of visiting s and consider the policy

$$\tilde{\pi}(\tilde{s}) = \begin{cases} \pi^*(\tilde{s}) & \text{if } s \neq \tilde{s} \\ a & \text{if } s = \tilde{s} \end{cases}$$

that matches π^* on all states except s. We will now show that $\tilde{\pi}$ achieves the same return as π^* in λ. Consider their difference

$$\begin{aligned} v_\lambda^{\tilde{\pi}} - v_\lambda^{\pi^*} &\overset{(i)}{=} w_\lambda^{\tilde{\pi}}(s, \tilde{\pi}(s))[r_\lambda(s,\tilde{\pi}(s)) + \langle P_\lambda(\cdot|s,\tilde{\pi}(s)), V_\lambda^{\tilde{\pi}} \rangle] \\ &\quad - w_\lambda^{\pi^*}(s, \pi^*(s))[r_\lambda(s,\pi^*(s)) + \langle P_\lambda(\cdot|s,\pi^*(s)), V_\lambda^{\pi^*} \rangle] \\ &\overset{(ii)}{=} w_\lambda^{\pi^*}(s,\pi^*(s))[r_\lambda(s,\tilde{\pi}(s)) - r_\lambda(s,\pi^*(s)) + \langle P_\lambda(\cdot|s,\tilde{\pi}(s)) - P_\lambda(\cdot|s,\pi^*(s)), V_\lambda^{\pi^*} \rangle] \\ &\overset{(iii)}{=} w_\theta^{\pi^*}(s,\pi^*(s))[\Delta + r_\theta(s,\tilde{\pi}(s)) - r_\theta(s,\pi^*(s)) + \langle P_\theta(\cdot|s,\tilde{\pi}(s)) - P_\theta(\cdot|s,\pi^*(s)), V_\theta^* \rangle] \\ &\overset{(iv)}{=} w_\theta^{\pi^*}(s,\pi^*(s))[\Delta - \operatorname{gap}_\theta(s,\tilde{\pi}(s))] \end{aligned}$$

where (i) and (ii) follow from the fact that $\tilde{\pi}$ and π^* only differ on s and hence, their probability at arriving at s and their value for any successor state of s is identical. Step (iii) follows from the fact that λ and θ only differ on (s,a) which is not visited by π^*. Finally, step (iv) applies the definition of optimal value functions and value-function gaps. Since $\Delta = \operatorname{gap}_\theta(s,\tilde{\pi}(s))$, it follows that $v_\lambda^{\tilde{\pi}} = v_\lambda^{\pi^*} = v_\theta^{\pi^*} = v_\theta^*$. As we have seen above, the optimal value function (and return) is identical in θ and λ and, hence, $\tilde{\pi}$ is optimal in λ.

Note that the we can apply the chain of equalities above in the same manner to $v_\theta^{\tilde{\pi}} - v_\theta^{\pi^*}$ if we consider $\Delta = 0$. This yields

$$v_\theta^{\tilde{\pi}} - v_\theta^{\pi^*} = -w_\theta^{\pi^*}(s,\pi^*(s)) \operatorname{gap}_\theta(s,a) < 0$$

because $w_\theta^{\pi^*}(s, \pi^*(s)) > 0$ and $\mathrm{gap}_\theta(s, a) < 0$ by assumption. Hence $\tilde{\pi}$ is not optimal in θ, which completes the proof. □

Lemma 5.7.2 (Optimization problem over $\mathcal{S} \times \mathcal{A}$ instead of Π)**.** *Let optimal value $C(\theta)$ of the optimization problem Equation 5.27 in Theorem 5.6.1 is lower-bound by the optimal value of the problem*

$$\begin{aligned}
\underset{\eta(s,a)\geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a)\, \mathrm{gap}_\theta(s,a) \\
s.t. \quad & \sum_{s,a} \eta(s,a) KL(R_\theta(s,a), R_\lambda(s,a)) \\
& + \sum_{s,a} \eta(s,a) KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a)) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta)
\end{aligned} \tag{5.32}$$

Proof. First, we rewrite the objective of Equation 5.27 as

$$\sum_{\pi\in\Pi} \eta(\pi)(v_\theta^* - v_\theta^\pi) \overset{(i)}{=} \sum_{\pi\in\Pi} \eta(\pi) \sum_{s,a} w_\theta^\pi(s,a)\, \mathrm{gap}_\theta(s,a) = \sum_{s,a} \left(\sum_{\pi\in\Pi} \eta(\pi) w_\theta^\pi(s,a) \right) \mathrm{gap}_\theta(s,a)$$

where step (i) applies Lemma 5.5.1 proved in Section 5.5. Here, $w_\theta^\pi(s,a)$ is the probability of reaching s and taking a when playing policy π in MDP θ. Similarly, the LHS of the constraints of Equation 5.27 can be decomposed as

$$\begin{aligned}
& \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \\
& = \sum_{\pi\in\Pi} \eta(\pi) \sum_{s,a} w_\theta^\pi(s,a) \left(KL(R_\theta(s,a), R_\lambda(s,a)) + KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a)) \right) \\
& = \sum_{s,a} \left[\sum_{\pi\in\Pi} \eta(\pi) w_\theta^\pi(s,a) \right] \left(KL(R_\theta(s,a), R_\lambda(s,a)) + KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a)) \right)
\end{aligned}$$

where the first equality follows from writing out the definition of the KL divergence. Let now $\eta(\pi)$ be a feasible solution to the original problem Equation 5.27. Then the two equalities we just proved show that $\eta(s,a) = \sum_{\pi\in\Pi} \eta(\pi) w_\theta^\pi(s,a)$ is a feasible solution for the problem in Equation 5.32 with the same value. Hence, since Equation 5.32 is a minimization problem, its optimal value cannot be larger than $C(\theta)$, the optimal value of Equation 5.27. □

Proof of Theorem 5.6.3. Let $\bar{\Lambda}(\theta)$ be a set of all confusing MDPs from Lemma 5.6.2, that is, for every suboptimal (s,a), $\bar{\Lambda}(\theta)$ contains exactly one confusing MDP that differs with θ only in the immediate reward at (s,a). Consider now the relaxation of Theorem 5.6.1 from Lemma 5.7.2 and further relax it by reducing the set of constraints induced by $\Lambda(\theta)$ to only the set of constraints induced by $\bar{\Lambda}(\theta)$:

$$\begin{aligned} \underset{\eta(s,a)\geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a) \operatorname{gap}_\theta(s,a) \\ \text{s.t.} \quad & \sum_{s,a} \eta(s,a) KL(R_\theta(s,a), R_\lambda(s,a)) \geq 1 \qquad \text{for all} \ \ \lambda \in \bar{\Lambda}(\theta) \end{aligned}$$

Since all confusing MDPs only differ in rewards, we dropped the KL-term for the transition probabilities. We can simplify the constraints by noting that for each λ, only one KL-term is non-zero and it has value $\operatorname{gap}_\theta(s,a)^2$. Hence, we can write the problem above equivalently as

$$\begin{aligned} \underset{\eta(s,a)\geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a) \operatorname{gap}_\theta(s,a) \\ \text{s.t.} \quad & \eta(s,a) \operatorname{gap}_\theta(s,a)^2 \geq 1 \qquad \text{for all} \ \ (s,a) \in \mathcal{S} \times \mathcal{A} \text{ with } \ \operatorname{gap}_\theta(s,a) > 0 \end{aligned}$$

Rearranging the constraint as $\eta(s,a) \geq 1/\operatorname{gap}_\theta(s,a)^2$, we see that the value is lower-bounded by

$$\sum_{s,a} \eta(s,a) \operatorname{gap}_\theta(s,a) \geq \sum_{s,a\colon \operatorname{gap}_\theta(s,a)>0} \eta(s,a) \operatorname{gap}_\theta(s,a) \geq \sum_{s,a\colon \operatorname{gap}_\theta(s,a)>0} \frac{1}{\operatorname{gap}_\theta(s,a)},$$

which completes the proof. □

We note that because the relaxation in Lemma 5.7.2 essentially allows the algorithm to choose which state-action pairs to play instead of just policies, the final lower bound in Theorem 5.6.3 may be loose, especially in factors of H. However, it is unlikely that the $\operatorname{gap}_{\min}$ term arising in the upper bound of Simchowitz and Jamieson (2019) can be recovered. We conjecture that such a term can be avoided by algorithms, which do not construct optimistic estimators for the Q-function at each state-action pair but rather just work with a class of policies and construct only optimistic estimators of the return.

5.7.3 Lower bounds for deterministic MDPs

We will show that we can derive lower bounds in two cases:

1. We show that if the graph induced by the MDP is a tree, then we can formulate a finite LP which has value at most a polynomial factor of H away from the value of LP 5.27.

2. We show that if we assume that the value function for any policy is at most 1 and the rewards of each state-action pair are at most 1, then we can derive a closed form lower bound. This lower bound is also at most a polynomial factor of H away from the solution to LP 5.27.

We begin by stating a helpful lemma, which upper and lower bounds the KL-divergence between two environments on any policy π. Since we consider Gaussian rewards with $\sigma = 1/\sqrt{2}$ it holds that $KL(R_\theta(s,a), R_\lambda(s,a)) = (r_\theta(s,a) - r_\lambda(s,a))^2$. Further for any π and λ it holds that $KL(\theta(\pi), \lambda(\pi)) = \sum_{(s,a)\in\pi} KL(R_\theta(s,a), R_\lambda(s,a)) = \sum_{(s,a)\in\pi}(r_\theta(s,a) - r_\lambda(s,a))^2$. We can now show the following lower bound on $KL(\theta(\pi), \lambda(\pi))$.

Lemma 5.7.3. *Fix π and suppose λ is such that $\pi^*_\lambda = \pi$. Then $(v^* - v^\pi)^2 \geq KL(\theta(\pi), \lambda(\pi)) \geq \frac{(v^*-v^\pi)^2}{H}$.*

Proof. The second inequality follows from the fact that the optimization problem

$$\begin{aligned} \underset{\theta,\lambda\in\Lambda(\theta):\pi^*_\lambda=\pi}{\text{minimize:}} \quad & \sum_{(s,a)\in\pi} (r_\theta(s,a) - r_\lambda(s,a))^2 \\ \text{subject to:} \quad & \sum_{(s,a)\in\pi} r_\lambda(s,a) - r_\theta(s,a) \geq v^* - v^{\pi}, \end{aligned}$$

admits a solution at θ, λ for which $r_\lambda(s,a) - r_\theta(s,a) = \frac{v^*-v^\pi}{H}, \forall (s,a) \in \pi$. The first inequality follows from considering the optimization problem

$$\begin{aligned} \underset{\theta,\lambda\in\Lambda(\theta):\pi^*_\lambda=\pi}{\text{maximize}} \quad & \sum_{(s,a)\in\pi} (r_\theta(s,a) - r_\lambda(s,a))^2 \\ \text{s.t.} \quad & \sum_{(s,a)\in\pi} r_\lambda(s,a) - r_\theta(s,a) \geq v^* - v^\pi, \end{aligned}$$

and the fact that it admits a solution at θ, λ for which there exists a single state-action pair $(s,a) \in \pi$ such that $r_\theta(s,a) - r_\lambda(s,a) = v^* - v^\pi$ and for all other (s,a) it holds that $r_\lambda(s,a) = r_\theta(s,a)$. □

Using the above Lemma 5.7.3 we now show that we can restrict our attention only to environments $\lambda \in \Lambda(\theta)$ which make one of $\pi^*_{(s,a)}$ optimal and derive an upper bound on $C(\theta)$ which we will try to match, up to factors of H, later. Define the set $\tilde{\Lambda}(\theta) = \{\lambda \in \Lambda(\theta) : \exists (s,a) \in \mathcal{S} \times \mathcal{A}, \pi^*_\lambda = \pi^*_{(s,a)}\}$ and $\Pi^* = \{\pi \in \Pi, \pi \neq \pi^*_\theta : \exists (s,a) \in \mathcal{S} \times \mathcal{A}, \pi = \pi^*_{(s,a)}\}$. We have

Lemma 5.7.4. *Let $\tilde{C}(\theta)$ be the value of the optimization problem*

$$\begin{aligned} &\underset{\eta(\pi)\geq 0}{\text{minimize:}} && \sum_{\pi\in\Pi^*} \eta(\pi)(v^* - v^\pi) \\ &\text{subject to:} && \sum_{\pi\in\Pi^*} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \tilde{\Lambda}(\theta) \end{aligned}. \tag{5.33}$$

Then $\sum_{\pi\in\Pi^*} \frac{H}{v^*-v^\pi} \geq C(\theta) \geq \frac{\tilde{C}(\theta)}{H}$.

Proof. We begin by showing $C(\theta) \geq \frac{\tilde{C}(\theta)}{H}$ holds. Fix a $\pi \notin \Pi^*$ s.t. the solution of LP 5.27 implies $\eta(\pi) > 0$. Let $\lambda \in \tilde{\Lambda}(\theta)$ be a change of environment for which $KL(\theta(\pi), \lambda(\pi)) > 0$. We can now shift all of the weight of $\eta(\pi)$ to $\eta(\pi^*_\lambda)$ while still preserving the validity of the constraint. Further doing so to all $\pi^*_{(s,a)}$ for which $\pi^*_{(s,a)} \cap \pi \neq \emptyset$ will not increase the objective by more than a factor of H as $v^* - v^\pi \geq \frac{1}{H}\sum_{(s,a)\in\pi} v^* - v^{\pi^*_{(s,a)}}$. Thus, we have converted the solution to LP 5.27 to a feasible solution to LP 5.33 which is only a factor of H larger.

Next we show that $\sum_{\pi\in\Pi^*} \frac{H}{v^*-v^\pi} \geq C(\theta)$. Set $\eta(\pi) = 0, \forall \pi \in \Pi \setminus \Pi^*$ and set $\eta(\pi) = \frac{H}{(v^*-v^\pi)^2}, \forall \pi \in \Pi^*$. If π is s.t. $\eta(\pi) > 0$ then for any λ which makes π optimal it holds that

$$\begin{aligned} 1 &\leq \frac{H}{(v^* - v^{\pi^*_\lambda})^2} \times \frac{(v^* - v^{\pi^*_\lambda})^2}{H} \leq \frac{H}{(v^* - v^{\pi^*_\lambda})^2} KL(\theta(\pi^*_\lambda), \lambda(\pi^*_\lambda)) \\ &= \eta(\pi^*_\lambda) KL(\theta(\pi^*_\lambda), \lambda(\pi^*_\lambda)) \leq \sum_{\pi'\in\Pi} \eta(\pi') KL(\theta(\pi'), \lambda(\pi')), \end{aligned}$$

where the second inequality follows from Lemma 5.7.3. Next, if π is s.t. $\eta(\pi) = 0$ then for any λ which makes π optimal it holds that

$$\begin{aligned}
\sum_{\pi'\in\Pi}\eta(\pi')KL(\theta(\pi'),\lambda(\pi')) &\geq \sum_{(s,a)\in\pi^*_\lambda}\eta(\pi^*_{(s,a)})KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\
&= \sum_{(s,a)\in\pi^*_\lambda}\frac{H}{(v^*-v^{\pi^*_{(s,a)}})^2}KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\
&\geq \frac{H}{(v^*-v^{\pi^*_\lambda})^2}\sum_{(s,a)\in\pi^*_\lambda}KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\
&\geq \frac{H}{(v^*-v^{\pi^*_\lambda})^2}\sum_{(s,a)\in\pi^*_\lambda}KL(R_\theta(s,a),R_\lambda(s,a))\\
&= \frac{H}{(v^*-v^{\pi^*_\lambda})^2}KL(\theta(\pi^*_\lambda),\lambda(\pi^*_\lambda)) \geq 1,
\end{aligned}$$

where the second inequality follows from the fact that $v^{\pi^*_\lambda} \leq v^{\pi^*_{(s,a)}}, \forall (s,a)\in\pi^*_\lambda$. □

5.7.3.1 Proof of Theorem 5.6.4

Lemma 5.7.5. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by* 1*. Consider an MDP $\theta\in\Theta$ with deterministic transitions. Then, for any reachable state-action pair (s,a) that is not visited by any optimal policy, there exists a confusing MDP $\lambda\in\Lambda(\theta)$ with*

- *λ and θ only differ in the immediate reward at (s,a)*
- *$KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda) = w^\pi_\theta(s,a)(v^*_\theta - v_\theta^{\pi^*_{(s,a)}})^2$ for all $\pi\in\Pi$ where $v_\theta^{\pi^*_{(s,a)}} = \max_{\pi:\, w^\pi(s,a)>0} v^\pi_\theta$.*

Proof. Let $(s,a)\in\mathcal{S}\times\mathcal{A}$ be any state-action pair that is not visited by any optimal policy. Then $v_\theta^{\pi^*_{(s,a)}} = \max_{\pi:\, w^\pi(s,a)>0} v^\pi_\theta \leq v^*_\theta$ is strictly suboptimal in θ. Let $\tilde{\pi}$ be any policy that visits (s,a) and achieves the highest return $v_\theta^{\pi^*_{(s,a)}}$ in θ possible among such policies.

Define λ to be the MDP that matches θ except in the immediate reward at (s,a), which we set as $R_\lambda(s,a) = \mathcal{N}(r_\theta(s,a)+\Delta, 1/2)$ with $\Delta = v^*_\theta - v_\theta^{\pi^*_{(s,a)}}$. That is, the expected reward of λ in (s, a) is raised by Δ. For any policy π, it then holds

$$KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = w_\theta^\pi(s,a) KL(R_\theta(s,a), R_\lambda(s,a))$$
$$v_\lambda^\pi = w_\theta^\pi(s,a)\Delta + v_\theta^\pi$$

due to the deterministic transitions. Hence, while $v_\lambda^* = v_\theta^*$ and all optimal policies of θ are still optimal in λ, now policy $\tilde{\pi}$, which is not optimal in θ is optimal in λ.

By the choice of Gaussian rewards with variance 1/2, we have $KL(R_\theta(s,a), R_\lambda(s,a)) = (v_\theta^* - v_\theta^{\pi^*_{(s,a)}})^2$ and thus $KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = w_\theta^\pi(s,a)(v_\theta^* - v_\theta^{\pi^*_{(s,a)}})^2$ for all $\pi \in \Pi$.

It only remains to show that $\lambda \in \Theta$, i.e., that all immediate rewards and optimal value function is bounded by 1. For rewards, we have

$$r_\lambda(s,a) = r_\theta(s,a) + \Delta = r_\theta(s,a) + v_\theta^* - v_\theta^{\pi^*_{(s,a)}} = v_\theta^* - \underbrace{(v_\theta^{\pi^*_{(s,a)}} - r_\theta(s,a))}_{\geq 0} \leq v_\theta^* \leq 1$$

for (s, a) and for all other (s', a'), $r_\lambda(s', a') = r_\theta(s', a') \leq 1$. Finally, the value function at any reachable state is bounded by the optimal return $v_\lambda^* = v_\theta^* \leq 1$ and for any unreachable state, the optimal value function of λ is identical to the optimal value function of θ. Hence, $\lambda \in \Theta$. □

Proof of Theorem 5.6.4. The proof works by first relaxing the general LP 5.27 and then considering its dual. We now define the set $\check{\Lambda}(\theta)$ which consists of all changes of environment which make $\pi^*_{(s,a)}$ optimal by only changing the distribution of the reward at (s, a) by making it $v_\theta^* - v_\theta^{\pi^*_{(s,a)}}$ larger. Formally, the set is defined as

$$\check{\Lambda}(\theta) = \Big\{\lambda_{(s,a)} \colon \lambda \in \Lambda(\theta), KL(R_\theta(s,a), R_\lambda(s,a)) = (v_\lambda^* - v^{\pi^*_{(s,a)}})^2,$$
$$KL(R_\theta(s',a'), R_\lambda(s',a')) = 0, KL(P_\theta(s',a'), P_\lambda(s',a')) = 0, \forall (s',a') \neq (s,a)\Big\}.$$

This set is guaranteed to be non-empty (for any reasonable MDP) by Lemma 5.7.5. The relaxed LP is now give by

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize:}} \quad & \sum_{\pi\in\Pi} \eta(\pi)(v_\theta^* - v_\lambda^\pi) \\ \text{subject to:} \quad & \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1 \qquad \text{for all } \lambda \in \check{\Lambda}(\theta) \end{aligned}. \tag{5.34}$$

The dual of the above LP is given by

$$\begin{aligned}
&\underset{\mu(\lambda)\geq 0}{\text{maximize}} \quad && \sum_{\lambda\in\check{\Lambda}(\theta)} \mu(\lambda) \\
&\text{s.t.} && \sum_{\lambda\in\check{\Lambda}(\theta)} \mu(\lambda) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \leq v_\theta^* - v_\theta^\pi \qquad \text{for all } \pi \in \Pi.
\end{aligned} \tag{5.35}$$

By weak duality, the value of any feasible solution to Equation 5.35 produces a lower bound on $C(\theta)$ in Theorem 5.6.1. Let

$$\mathcal{X} = \{(s,a) \in \mathcal{S} \times \mathcal{A} \colon w_\theta^\pi(s,a) = 0 \text{ for all } \pi \in \Pi_\theta^* \text{ and } w_\theta^\pi(s,a) > 0 \text{ for some } \pi \in \Pi \setminus \Pi_\theta^*\}$$

be the set of state-action pairs that are reachable in θ but no optimal policy visits. Then consider a dual solution μ that puts 0 on all confusing MDPs except on the $|\mathcal{X}|$ many MDPs from Lemma 5.7.5. Since each such confusing MDP is associated with an $(s,a) \in \mathcal{X}$, we can rewrite μ as a mapping from $\mathcal{X}$ to $\mathbb{R}$ sending $(s,a) \to \lambda_{(s,a)}$. Specifically, we set

$$\mu(s,a) = \frac{1}{H}\left(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}\right)^{-1} \qquad \text{for all } (s,a) \in \mathcal{X}.$$

To show that this μ is feasible, consider the LHS of the constraints in Equation 5.35

$$\begin{aligned}
\sum_{\lambda\in\check{\Lambda}(\theta)} \mu(\lambda) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) &= \sum_{(s,a)\in\mathcal{X}} \frac{1}{H}\left(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}\right) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_{(s,a)}^\pi) \\
&= \sum_{(s,a)\in\mathcal{X}} \frac{1}{H}\left(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}\right)^{-1} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^{\pi^*_{(s,a)}})^2 \\
&= \sum_{(s,a)\in\mathcal{X}} \frac{1}{H} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^{\pi^*_{(s,a)}})
\end{aligned}$$

where the first equality applies our definition of μ and the second uses the expression for the KL-divergence from Lemma 5.7.5. By definition of $v_\theta^{\pi^*_{(s,a)}}$, we have $v_\theta^{\pi^*_{(s,a)}} \geq v_\theta^\pi$ for all policies π with $w_\theta^\pi(s,a) > 0$. Thus,

$$\begin{aligned}
\sum_{(s,a)\in\mathcal{X}} \frac{1}{H} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}) &\leq \sum_{(s,a)\in\mathcal{X}} \frac{1}{H} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^\pi) \\
&\leq v_\theta^* - v_\theta^\pi
\end{aligned}$$

where the second inequality holds because each policy visits at most H states. Thus proves that μ defined above is indeed feasible. Hence, its objective value

$$\sum_{\lambda \in \Lambda(\theta)} \mu(\lambda) = \sum_{(s,a)\in\mathcal{X}} \frac{1}{H}\left(v^*_\theta - v_\theta^{\pi^*_{(s,a)}}\right)$$

is a lower-bound for $C(\theta)$ from Theorem 5.6.1 which finishes the proof. □

5.7.3.2 Tree-structured MDPs

Even though Lemma 5.7.4 restricts the set of confusing environments from $\Lambda(\theta)$ to $\tilde{\Lambda}(\theta)$, this set could still have exponential or even infinite cardinality. In this section we show that for a type of special MDPs we can restrict ourselves to a finite subset of $\tilde{\Lambda}(\theta)$ of size at most SA.

Arrange $\pi^*_{(s,a)}, (s,a) \in \mathcal{S}\times\mathcal{A}$ according to the value functions $v^{\pi^*_{(s,a)}}$. Under this arrangement let $\pi_1 \succeq \pi_2 \succeq, \dots, \succeq \pi_m$. Let $\pi_0 = \pi^*_\theta$. We will now construct m environments $\lambda_1, \dots, \lambda_m$, which will constitute the finite subset. We begin by constructing λ_1 as follows. Let $\mathcal{B}_1$ be the set of all $(s_h, a_h) \in \pi_1$ and $(s_h, a_h) \notin \pi_0$. Arrange the elements in $\mathcal{B}_1$ in inverse dependence on horizon $(s_{h_1}, a_{h_1}) \preceq (s_{h_2}, a_{h_2}) \preceq \dots \preceq (s_{h_{H_1}}, a_{h_{H_1}})$, where $H_1 = |\mathcal{B}_1|$, so that $h_1 > h_2 >, \dots, h_{H_1}$. Let λ_1 be the environment which sets

$$\begin{aligned}
R_{\lambda_1}(s_{h_1}, a_{h_1}) &= \min(1, v^{\pi_0} - v^{\pi_1})\\
R_{\lambda_1}(s_{h_2}, a_{h_2}) &= \min(1, \max(R_\theta(s_{h_2}, a_{h_2}), R_\theta(s_{h_2}, a_{h_2}) + v^{\pi_0} - (v^{\pi_1} - R_\theta(s_{h_1}, a_{h_1})) - 1)))\\
&\vdots\\
R_{\lambda_1}(s_{h_i}, a_{h_i}) &= \min(1, \max(R_\theta(s_{h_i}, a_{h_i}), R_\theta(s_{h_i}, a_{h_i}) + v^{\pi_0} - (v^{\pi_1} - \sum_{\ell=1}^{i} R_\theta(s_{h_\ell}, a_{h_\ell})) - i))\\
&\vdots
\end{aligned}$$

Clearly λ_1 makes π_1 optimal and also does not change the value of any state-action pair which belongs to π_0 so it agrees with θ on π_0. Further $\pi_2, \pi_3, \dots, \pi_m$ are still suboptimal policies under λ_1. This follows from the fact that for any $i > 1$, $v^{\pi_1} > v^{\pi_i}$ and there exists (s,a) such that $(s,a) \in \pi_i$ but $(s,a) \notin \pi_1$ so $R_{\lambda_1}(s,a) = R_\theta(s,a)$. Further λ_1 only increases the rewards for state-action pairs in π_1 and hence $v^{\pi_1}_{\lambda_1} > v^{\pi_i}_{\lambda_1}$. Notice that there exists an index $\tilde{H}_1$ at which $R_{\lambda_1}(s_{h_{\tilde{H}_1}}, a_{h_{\tilde{H}_1}}) = v^{\pi_0} - (v^{\pi_1} - \sum_{\ell=1}^{\tilde{H}_1} R_\theta(s_{h_\ell}, a_{h_\ell})) - \tilde{H}_1) \geq R_\theta(a_{\tilde{H}_1}, s_{\tilde{H}_1})$. For this index it holds that for $h < \tilde{H}_1$, $R_{\lambda_1}(s_h, a_h) = 1$ and for $h > \tilde{H}_1$, $R_{\lambda_1}(s_h, a_h) = R_\theta(s_h, a_h)$.

Let

$$\mathcal{B}_i = \{(s,a) \in \pi_i : (s,a) \notin \bigcup_{\ell < i} \pi_\ell\}$$
$$\tilde{\mathcal{B}}_i = \{(s,a) \in \pi_i : (s,a) \in \bigcup_{\ell < i} \pi_\ell\}.$$

We first define an environment $\tilde{\lambda}_i$ on $(s,a) \in \tilde{\mathcal{B}}_i$ as follows. $R_{\lambda_i}(s,a) = R_{\lambda_\ell}(s,a)$, where $\ell < i$ is such that $(s,a) \in \mathcal{B}_\ell$. Let $v^{\pi_i}_{\tilde{\lambda}_i}$ be the value function of π_i with respect to $\tilde{\lambda}_i$.

Lemma 5.7.6. *It holds that* $v^{\pi_i}_{\tilde{\lambda}_i} \leq v^{\pi_0}$.

Proof. Let $\tilde{H}_i$ be the index for which it holds that for $\ell \leq \tilde{H}_i$, $(s_{h_\ell}, a_{h_\ell}) \in \pi_i \iff (s_{h_\ell}, a_{h_\ell}) \in \mathcal{B}_i$. Such a $\tilde{H}_i$ exists as there is a unique sub-tree $\mathcal{M}_i$, of maximal depth, for which it holds that if $\pi_j \bigcap \mathcal{M}_i \neq \emptyset \iff \pi_i \succeq \pi_j$. The root of this subtree is exactly at depth $H - h_{\tilde{H}_i}$. Let π_j be any policy such that $\pi_j \succeq \pi_i$ and $\exists (s_{h_{\tilde{H}_i}}, a_{h_{\tilde{H}_i}}) \in \pi_j$. By the maximality of $\mathcal{M}_i$ such a π_j exists. Because of the tree structure it holds that for any $h' > h_{\tilde{H}_i}$ if $(s_{h'}, a_{h'}) \in \pi_i \implies (s_{h'}, a_{h'}) \in \pi_j$ and hence $\tilde{\lambda}_i = \lambda_j$ up to depth $h_{\tilde{H}_i}$. Since π_i and π_j match up to depth $H - h_{\tilde{H}_i}$ and $\pi_j \succeq \pi_i$ it also holds that

$$\sum_{\ell \leq \tilde{H}_i} R_{\lambda_j}(s^{\pi_j}_{h_\ell}, a^{\pi_j}_{h_\ell}) \geq \sum_{\ell \leq \tilde{H}_i} R_\theta(s^{\pi_j}_{h_\ell}, a^{\pi_j}_{h_\ell}) \geq \sum_{\ell \leq \tilde{H}_i} R_\theta(s^{\pi_i}_{h_\ell}, a^{\pi_i}_{h_\ell}) = \sum_{\ell \leq \tilde{H}_i} R_{\tilde{\lambda}_i}(s^{\pi_i}_{h_\ell}, a^{\pi_i}_{h_\ell}).$$

Since π_j is optimal under λ_j the claim holds. □

For all $(s_{h_j}, a_{h_j}) \in \mathcal{B}_i$ we now set

$$R_{\lambda_i}(s_{h_j}, a_{h_j}) = \min(1, \max(R_\theta(s_{h_j}, a_{h_j}), R_\theta(s_{h_j}, a_{h_j}) + v^{\pi_0} - (v^{\pi_i}_{\tilde{\lambda}_i} - \sum_{\ell=1}^{j} R_{\tilde{\lambda}_i}(s_{h_\ell}, a_{h_\ell})) - j)), \tag{5.36}$$

and for all $(s_h, a_h) \in \tilde{\mathcal{B}}_i$ we set $R_{\lambda_i}(s_h, a_h) = R_{\tilde{\lambda}_i}(s_h, a_h)$. From the definition of $\tilde{\mathcal{B}}_i$ it follows that λ_i agrees with all λ_j for $j \leq i$ on state-action pairs in π_i. Finally we need to show that the construction in Equation 5.36 yields an environment λ_i for which π_i is optimal.

Lemma 5.7.7. *Under λ_i it holds that π_i is optimal.*

Proof. Let $\tilde{H}_i$ and π_j be as in the proof of Lemma 5.7.6. We now show that $\sum_{\ell \leq \tilde{H}_i} R_{\lambda_j}(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j}) \leq \sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i})$. We only need to show that $\sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}) \geq v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i}$. From Equation 5.36 we have $R_{\lambda_i}(s_{h_1}, a_{h_1}) = \min(1, v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i})$. If $R_{\lambda_i}(s_{h_1}, a_{h_1}) = v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i}$ then the claim is complete. Suppose $R_{\lambda_i}(s_{h_1}, a_{h_1}) = 1$. This implies $v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i} \geq 1 - R_\theta(s_{h_1}, a_{h_1})$. Next the construction adds the remaining gap of $v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i} + R_\theta(s_{h_1}, a_{h_1}) - 1$ to $R_\theta(s_{h_2}, a_{h_2})$ and clips $R_{\lambda_i}(s_{h_2}, a_{h_2})$ to 1 if necessary. Continuing in this way we see that if ever $R_{\lambda_i}(s_{h_j}, a_{h_j}) = R_\theta(s_{h_j}, a_{h_j}) + v^{\pi_0} - (v_{\tilde{\lambda}_i}^{\pi_i} - \sum_{\ell=1}^{j} R_{\tilde{\lambda}_i}(s_{h_\ell}, a_{h_\ell})) - j$ then $v^{\pi_0} - V_{\tilde{\lambda}_i}^{\pi_i} \leq \sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i})$. On the other hand if this never occurs, we must have $R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}) = 1 \geq R_{\lambda_j}(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j})$ which concludes the claim. □

Let $\widehat{\Lambda}(\theta) = \{\lambda_1, \ldots, \lambda_m\}$ be the set of the environments constructed above. We now show that the value of the optimization problem is not too much smaller than the value of Problem 5.27.

Theorem 5.7.8. *The value $\widehat{C}(\theta)$ of the LP*

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\textit{minimize:}} \quad & \sum_{\pi \in \Pi^*} \eta(\pi)(v^* - v^\pi) \\ \textit{subject to:} \quad & \sum_{\pi \in \Pi^*} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \widehat{\Lambda}(\theta)^{,} \end{aligned}$$

satisfies $\widehat{C}(\theta) \geq \frac{C(\theta)}{H^2}$ and $C(\theta) \geq \frac{\widehat{C}(\theta)}{H}$.

Proof. The inequality $C(\theta) \geq \frac{\widehat{C}(\theta)}{H}$ follows from Lemma 5.7.4 and the fact that the above optimization problem is a relaxation to LP 5.33.

To show the first inequality we consider the following relaxed LP

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize:}} \quad & \sum_{\pi\in\Pi} \eta(\pi)(v^* - v^\pi) \\ \text{subject to:} \quad & \sum_{\pi\in\Pi} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \widehat{\Lambda}(\theta) \end{aligned}.$$

Any solution to the LP in the statement of the theorem is feasible for the above LP and thus the value of the above LP is no larger. We now show that the value of the above LP is greater than or equal to $\frac{C(\theta)}{H^2}$. Fix $\lambda \in \widehat{\Lambda}(\theta)$. We show that for any $\lambda' \in \Lambda(\theta)$ such that $\pi^*_\lambda = \pi^*_{\lambda'}$ it holds that $KL(\theta(\pi), \lambda(\pi)) \leq H^2 KL(\theta(\pi), \lambda'(\pi)), \forall \pi \in \Pi$. This would imply that if η is a solution to the above LP, then $H^2\eta$ is feasible for LP 5.27 and therefore $\widehat{C}(\theta) \geq \frac{C(\theta)}{H^2}$.

Arrange $\pi \in \Pi : KL(\theta(\pi), \lambda(\pi)) > 0$ according to $KL(\theta(\pi), \lambda(\pi))$ so that

$$\pi_i \preceq \pi_j \iff KL(\theta(\pi_i), \lambda(\pi_i)) \geq KL(\theta(\pi_j), \lambda(\pi_j)).$$

Consider the optimization problem

$$\begin{aligned} \underset{\lambda'\in\Lambda(\theta)}{\text{minimize:}} \quad & KL(\theta(\pi_i), \lambda'(\pi_i)) \\ \text{subject to:} \quad & \pi^*_{\lambda'} = \pi^*_\lambda \end{aligned}.$$

If we let $\Delta_{\lambda'}(s_h, a_h), (s_h, a_h) \in \pi^*_\lambda$ denote the change of reward for (s_h, a_h) under environment λ', then the above optimization problem can be equivalently written as

$$\begin{aligned} \underset{\lambda'\in\Lambda(\theta)}{\text{minimize:}} \quad & \sum_{h=1}^{h_{\tilde{H}_i}} \Delta_{\lambda'}(s_h, a_h)^2 \\ \text{subject to:} \quad & \sum_{h=1}^{H} r(s_h, a_h) + \Delta_{\lambda'}(s_h, a_h) \geq v^* \end{aligned}.$$

It is easy to see that the solution to the above optimization problem is to set $r(s_h, a_h) + \Delta_{\lambda'}(s_h, a_h) = 1$ for all $h \in [h_{\tilde{H}_i} + 1, H]$ and spread the remaining mass of $v^* - \tilde{H}_i - (v^{\pi^*_\lambda} - \sum_{\ell=1}^{\tilde{H}_i}) R_\theta(s_{h_\ell}, a_{h_\ell})$ as uniformly as possible on $\Delta_{\lambda'}(s_h, a_h)$, $h \in [1, h_{\tilde{H}_i}]$. Notice that under this construction the solution to the above optimization problem and λ match for $h \in [h_{\tilde{H}_i} + 1, H]$. Since the remaining mass is now the same it now holds that for any λ', $\sum_{h=1}^{h_{\tilde{H}_i}} \Delta_{\lambda'}(s_h, a_h)^2 \geq \frac{1}{h^2_{\tilde{H}_i}} \sum_{h=1}^{h_{\tilde{H}_i}} \Delta_\lambda(s_h, a_h)^2$. This implies $KL(\theta(\pi_i), \lambda'(\pi_i)) \geq \frac{1}{\tilde{H}_i^2} KL(\theta(\pi), \lambda(\pi))$ and the result follows as $\tilde{H}_i \leq H, \forall i \in [H]$. □

5.7.4 Omitted proofs for Theorem 5.6.7

Proof of Lemma 5.6.5. Assume $n_k(s_{5,1}) \geq N/4$, then we have

$$\begin{aligned}\bar{Q}_k(s_{4,1}, 1) &= \frac{1}{4} + \epsilon + \sum_{i=4}^{6} b_k^{rw}(s_{i,1}, 1) + b_k(s_{i,1}) \\ &\leq \frac{1}{4} + \epsilon + 6c\sqrt{\frac{\log\left(MN/(4\delta)\right)}{N/4}} \leq \frac{1}{4} + \epsilon + \frac{48\epsilon}{\sqrt{n}}\,,\end{aligned}$$

where we assume ϵ is sufficiently small such that $b_k(s,a) \leq b_k^{rw}(s,a)$ for $n_k(s,a) \geq N/4$.

On the other hand, we have have with probability at least 1-δ, that $\forall k : \widehat{r}_k(s_{4,1}, 2) + b_k^{rw}(s_{4,1}, 2) \geq 1/12$. Hence conditioned under that event, we have

$$\begin{aligned}\bar{Q}_k(s_{4,1}, 2) &= \frac{1}{4} + b_k^{rw}(s_{4,1}, 2) + b_k(s_{4,1}, 2) + \max_{j \in \{2, \ldots n+1} \sum_{i=5}^{6} b_k^{rw}(s_{i,j}, 1) + b_k(s_{i,j}, 1) \\ &\geq \frac{1}{4} + c\sqrt{\frac{\log\left(MN/(n\delta)\right)}{N/n}} \geq \frac{1}{4} + 4\epsilon\,.\end{aligned}$$

The proof is completed for $n_0 = 48^2$. □

Proof of Lemma 5.6.6. First we split $\bar{Q}_k(s_{1,1}, 2)$ into the observed sum of mean rewards and bonuses from $s_{1,1}$ to $s_{5,2}$ and the value $\bar{V}_k(s_{5,2})$. Then we upper bound $\bar{Q}_k(s_{1,1}, 1)$ by $\bar{V}_k(s_{5,2})$ and the maximum observed sum of mean rewards and bonuses along the paths passing by $s_{3,j}$ for $j \in [n]$. Finally analogous to the proof of Lemma 5.6.5, it is straightforward show that the latter is always larger as long as the visitation count for $s_{2,2}$ exceeds $N/4$. □

Chapter 6

Discussion and conclusion

We now discuss several open problems which follow from the studied problems in this work. The discussion is followed by concluding remarks.

6.1 Future directions

Chapter 2: While the primary goal of the work presented in Chapter 2 was to improve on known policy regret bounds when side observations are present, at the core of the main approach was solving the problem of online learning with side observations and switching costs. We had assumed that the switching costs are constant and equal to one throughout the game. A natural direction for future work is to investigate the problem when switching costs can vary both throughout the rounds of the game and in between different actions. Is it possible to come up with strategies with regret which depends on the total sum of the switching costs, similarly to first or second order regret bounds in online learning games? Further, we assumed that the feedback graph is fixed and known before the start of the game. Can we give meaningful regret bounds if the feedback evolves throughout the rounds of the game and we only observe the additional losses but not the topology of the graph? Further, can we provide improved regret guarantees if additional feedback is not determined by a graph but rather by a metric space in which the accuracy of the feedback is dependent on the distance from the selected action. Finally, our lower bounds showed an instance of the online learning game in which the presented strategy is essentially optimal. The instance, however, was for a fixed feedback graph. Can we extend the lower bounds for arbitrary feedback graphs, that is can we create a lower bound instance for any feedback graph G such that any player strategy has to suffer regret at least $\tilde{\Omega}(\gamma(G)^{1/3}T^{2/3})$?

Chapter 3: The recent work of Cutkosky et al. (2020) and our experiments suggest that the upper regret bounds for both algorithms proposed in Chapter 3 are not tight. In particular we expect that there exists a strategy which obtains a regret bound of the order $O\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i}+R_{i^*}(T)\right)$. It is unclear, however, if such a regret bound is possible without a priori knowledge of the time horizon T. Further, the question of min-max regret bounds for the corralling problem has not yet been investigated. Therefore two natural directions are to derive instance dependent regret lower bounds in the stochastic setting with gap and derive min-max regret bounds in the general adversarial setting. Another fundamental question which our work fails to address is what would happen in the stochastic setting if there are multiple base algorithms containing the best arm. Is there a regret bound which interpolates between gap-dependent regret and worst case $\sqrt{T}$ regret in this setting? We also ask if corralling can be extended to policy regret algorithms so to create a strategy which enjoys a $o(T)$ regret bound for all memory bounded adversaries with $m \in o(T)$ simultaneously? Finally, recent work (Zhu and Nowak, 2021) has shown that model selection for linear stochastic bandits is impossible in the worst case. On the other hand the work of Foster et al. (2019) shows that under some mild assumptions model selection is possible. Algorithm 10 in Chapter 3 can also be used to solve the model selection problem under certain assumptions. A natural question is what other assumptions allow for model selection.

Chapter 4: The definition of Policy Equilibrium that was presented in Chapter 4 only holds for two-player games and constant (independent of time horizon) memory

m. It is non-trivial to extend the definition and the results beyond two-player games as the interpretation of the view of each player becomes more involved. In particular, it is not possible to reason that the opposing player is now choosing functions between action sets as there are multiple opposing players. Having players with different memory bounds complicates things further as there is no simple markovian structure to the utilities. Can we extend our results for multiplayer games and m which is sublinear in time horizon? It is possible to minimize policy regret against stronger competitors, beyond the best fixed action for the history dependent losses, similar to reductions between swap regret minimization and external regret minimization. We know that when players decide their actions according to swap regret minimizing algorithms then the average play converges to the set of Correlated Equilibrium. Can we show similar results for players who are able to minimize notions of policy regret induced by competing against a stronger type of memory bounded adversary? When faced with a new extended class of equilibria it is natural to revisit some standard questions in game theory pertaining to price of anarchy and price of stability. For example there exist games in which the worse CCE is no worse than the worst Nash. Is this also true about the worst PE? Further what is the social welfare of the best PE compared to that of the best Nash? Are there natural algorithms which let us achieve socially good PE?

Chapter 5: The results in Chapter 5 improve on prior work results about optimistic algorithms, however, we know from Theorem 5.6.7 and the work of Simchowitz and Jamieson (2019) that it is impossible for optimistic algorithms to achieve the information theoretic optimal regret. The very recent work of Xu et al. (2021) makes a step towards closing the gap to the information theoretic optimal bounds by proposing a model-free, action-elimination type algorithm which can avoid the dependence on $\frac{SA}{\text{gap}_{\min}}$. Unfortunately the proposed bounds suffer the same issues as described in Figure 5-2. Our Definition 5.16 of the averaged clipping thresholds does not really rely on optimism. Can we use this or a similar definition to show an improved regret bound for the algorithm proposed by Xu et al. (2021)? In general we were not able to show closed form lower bounds for non-deterministic MDPs or even derive a computationally tractable approach to approximating the solution of the semi-infinite LP governing the information theoretic optimal rates. We pose the following question: is it possible to find a constant or even a $o(SA)$ factor approximation to Problem 5.27, computable in polynomial time. Answering this question to the affirmative will result in a new algorithm with a regret bound which avoids the $\frac{SA}{\mathrm{gap}_{\min}}$ term. On the other hand a negative answer will show that, in general, there is no hope of improving on existing bounds.

Other open problems: We have already discussed several natural open problems following from Chapter 3 and Chapter 5. We ask if it is possible to do corralling in RL. One such example is the following. In Chapter 5 we discussed limitations of current instance dependent bounds for the finite horizon tabular setting and proposed a new notion of instance dependent bounds to tackle the problem. The new bounds hold for most model-based optimistic algorithms. As we already discussed, the very recent work of (Xu et al., 2021) tackles a slightly different problem with existing regret bounds by proposing a model-free algorithm with an instance dependent regret guarantee which has improved dependence on the size of the state-action space. Is it possible to corral a model-free algorithm such as StrongEuler and the algorithm in (Xu et al., 2021) to achieve a best of both worlds instance dependent guarantee, that is a regret bound which has the improved dependence on the size of the state-action space and only depends on the averaged gaps? While policy regret is the de facto notion of regret which is minimized in RL, as the comparator is the best policy for the MDP, it is natural to ask what would happen if the MDP evolves through time, depending on the policies which the player selects. Can we define and minimize a notion of policy regret in RL which captures the notion of an evolving MDP where the transition kernel and rewards at a given episode depend on prior policies selected by the player?

6.2 Conclusion

This thesis investigated four different problems related to a counterfactual notion of regret called policy regret. The problems span the topics of online learning with side information, corralling and model selection, algorithmic game theory, and reinforcement learning.

In Chapter 2 we presented an extensive analysis of policy regret minimization in the presence of graph feedback, a scenario relevant to several applications in practice. We gave a new algorithm whose regret guarantee only depends on the domination number of the feedback graph. We also presented a matching lower bound for a family of graphs that includes disjoint unions of star graphs. The technical tools introduced in our proofs are likely to help derive a lower bound for all graph families. Our algorithms were based on a reduction to the problem of online learning with feedback graphs and switching costs in the adversarial setting.

In Chapter 3 we presented an extensive analysis of the problem of corralling stochastic bandits. Our algorithms are applicable to a number of different contexts where this problem arises. There are also several natural extensions and related questions relevant to our study. One natural extension is the case where the set of arms accessible to the base algorithms admit some overlap and where the reward observed by one algorithm could serve as side-information to another algorithm. Another extension is the scenario of corralling online learning algorithms with feedback graphs. In addition to these and many other interesting extensions, our analysis was shown to exhibit a connection with the problem of model selection for linear contextual bandits (Foster et al., 2019; Cutkosky et al., 2020; Pacchiano et al., 2020b; Zhu and

Nowak, 2021).

In Chapter 4 we gave a new twist on policy regret by examining it in the game setting, where we introduced the notion of policy equilibrium and showed that it captures the behavior of no policy regret players. While our characterization is precise, we view this as only the first step towards truly understanding policy regret and its variants in the game setting. Further, we showed that coarse correlated equilibria are a strict subset of policy equilibria by showing that policy regret minimization is incompatible with external regret minimization. Finally, we leveraged stability of natural external regret minimization strategies to show that the average play of such strategies will also converge to a policy equilibrium.

In Chapter 5 we prove that optimistic algorithms such as StrongEuler, can suffer substantially less regret compared to what prior work had shown. We do this by introducing a new notion of gap, while greatly simplifying and generalizing existing analysis techniques. We further investigated the information-theoretic limits of learning episodic layered MDPs. We provide two new closed-form lower bounds in the special case where the MDP has either deterministic transitions or the optimal policy is supported on all states. These lower bounds suggest that our notion of gap better captures the difficulty of an episodic MDP for RL.

6.3 Other work

Outside of bandit algorithms and reinforcement learning which are the main focus of this thesis, I have worked on several problems in the intersection of representation learning and stochastic approximation. In (Arora et al., 2016, 2017), I investigate the two related problems of Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) from a learning perspective in which the observed data is coming from some unknown distribution and the goal is to minimize the population risk. We propose streaming algorithms based on stochastic mirror descent which enjoy nice convergence guarantees. In (Marinov et al., 2018), I investigate the problem of Online Principcal Component Analysis (PCA) in the presence of corrupted or missing data and propose two algorithms based on online mirror descent and Oja's algorithm which enjoy $O(\sqrt{T})$ regret guarantees. In (Ullah et al., 2018), we propose an algorithm for the problem of Kernel PCA with Random Fourier Features improving on all prior bounds and achieving near optimal statistical rates. In (Arora and Marinov, 2019), I investigate two convex relaxations for the streaming PCA problem and show that SGD on one of the relaxations can indeed obtain comparable convergence rates to Oja's algorithm, which is considered min-max optimal. Finally, in (Arora et al., 2020), we propose a simple differentially private algorithm which enjoys optimal rates both in terms of statistical complexity and stochastic first-order oracle complexity whenever the privacy parameter is inversely proportional to the number of samples.

Appendix A

Convex optimization

A version of this appendix appeared in the lecture notes for the Optimization for Machine Learning class (EN 601.481/681) in the Fall of 2018.

A.1 Convex sets

We begin by a review of basic terminology for convex sets.

Definition A.1.1 (Convex set)**.** A set $\mathcal{X} \subseteq \mathbb{R}^d$ is said to be convex if for all $\mathrm{x}, \mathrm{y} \in \mathcal{X}$ the line segment $[\mathrm{x}, \mathrm{y}]$ lies entirely in $\mathcal{X}$, i.e., $\{\alpha \mathrm{x} + (1-\alpha)\mathrm{y} : 0 \leq \alpha \leq 1\} \subseteq \mathcal{X}$.

Definition A.1.2 (Convex combination)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$. Then the point $\sum_{i=1}^k \alpha_i \mathrm{x}_i$ such that $\alpha_i \geq 0$ for all i, and $\sum_{i=1}^k \alpha_i = 1$ is called a ***convex combination*** of $\mathrm{x}_1, \ldots, \mathrm{x}_k$. If $\mathcal{X}$ is convex then $\sum_{i=1}^k \alpha_i \mathrm{x}_i \in \mathcal{X}$.

Definition A.1.3 (Convex hull)**.** The ***convex hull*** of a set of points $\mathcal{X} \subseteq \mathbb{R}^d$ is the minimal convex set containing $\mathcal{X}$, i.e., it is the intersection of all convex sets containing $\mathcal{X}$. It can be equivalently defined as the set of all (finite) convex combinations of points in $\mathcal{X}$:

$$\mathrm{conv}(\mathcal{X}) = \left\{ \sum_{i=1}^m \alpha_i \mathrm{x}_i : \mathrm{x}_i \in \mathcal{X}, \alpha_i \geq 0, \sum_{i=1}^m \alpha_i = 1 \right\}.$$

Definition A.1.4 (Probability simplex)**.** The d-dimensional ***probability simplex*** is denoted as Δ^{d-1} and is the convex set

$$\Delta^{d-1} = \{\mathrm{x} \in \mathbb{R}^d : \sum_{i=1}^{d} \mathrm{x}_i = 1, \mathrm{x}_i \geq 0 \forall i \in [d]\}.$$

A.2 Convex Functions

Before we begin our discussion of convex function we need the following useful definition of a Lipschitz function.

Definition A.2.1 (Lower semi-continuity)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is ***lower semi-continous*** at a point $\mathrm{x}_0 \in \mathcal{X}$ if $\liminf_{\mathrm{x}\to\mathrm{x}_0} f(\mathrm{x}) \geq f(\mathrm{x}_0)$. A function f is said to be lower semi-continuous if it is lower semi-continuous at every point in its domain.

Definition A.2.2 (Lipschitz continuity)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is L-***Lipschitz*** continuous with respect to a norm $\|\cdot\|$ on $\mathcal{X}$ if for all $\mathrm{x}, \mathrm{y} \in \mathcal{X}$ it holds that

$$|f(\mathrm{x}) - f(\mathrm{y})| \leq L\|\mathrm{x} - \mathrm{y}\|.$$

Next, we review the definition and basic properties of convex functions.

Definition A.2.3 (Convex function)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be ***convex*** if for all $\mathrm{x}, \mathrm{y} \in \mathcal{X}$ and all $\alpha \in [0, 1]$,

$$f(\alpha\mathrm{x} + (1-\alpha)\mathrm{y}) \leq \alpha f(\mathrm{x}) + (1-\alpha)f(\mathrm{y}).$$

We say that f is ***concave*** if $-f$ is convex.

Definition A.2.4 (Proper function)**.** A convex function $f : \mathcal{X} \to \mathbb{R}$ is ***proper*** if for all $\mathrm{x} \in \mathcal{X}$, $f(\mathrm{x}) > -\infty$ and further there exists at least one $\mathrm{y} \in \mathcal{X}$ s.t. $f(\mathrm{y}) < \infty$.

Definition A.2.5 (Subgradient, subdifferential)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, and let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. Then $\mathrm{g} \in \mathbb{R}^d$ is a ***subgradient*** of f at x_0 if and only if

$$f(\mathrm{x}) \geq f(\mathrm{x}_0) + \langle \mathrm{g}, \mathrm{x} - \mathrm{x}_0 \rangle \text{ for all } \mathrm{x} \in \mathcal{X}. \tag{A.1}$$

The set of all subgradients is called a ***subdifferential***: $\partial f(\mathrm{x}_0) = \{g \mid g \text{ is a subgradient of } f \text{ at } \mathrm{x}_0\}$.

Theorem A.2.1 (Convexity: non-empty sub-differential)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is convex if and only if there exists a subgradient at each* $\mathrm{x} \in \mathcal{X}$.

Further we have the following characterization of convex Lipschitz functions.

Theorem A.2.2 (Convexity: Lipschitz continous functions)**.** *A convex function $f : \mathcal{X} \to \mathbb{R}$ is L-**Lipschitz** continuous with respect to a norm $\|\cdot\|$ iff for all $\mathrm{x}_0 \in \mathcal{X}$ and all subgradients g at x_0 it holds that $\|\mathrm{g}\| \leq L$.*

Definition A.2.6 (Epigraph)**.** The ***epigraph*** of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$\text{epi } f = \{(\mathrm{x}, y) : \mathrm{x} \in \mathcal{X}, y \geq f(\mathrm{x})\}.$$

Claim A.2.3 (Convexity: epigraph)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is convex if and only if its epigraph is a convex set.*

Definition A.2.7 (Sublevel sets)**.** The α***-sublevel*** set of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$S_\alpha = \{\mathrm{x} \mid f(\mathrm{x}) \leq \alpha\}.$$

Claim A.2.4. *If $f : \mathcal{X} \to \mathbb{R}$ is a convex function, then S_α is a convex set for all α.*

Proof. Let $(\mathrm{x}, \mathrm{y}) \in S_\alpha$. Then, $f(\mathrm{x}) \leq \alpha$ and $f(\mathrm{y}) \leq \alpha$ by definition. This implies, due to convexity, $f(\lambda \mathrm{x} + (1 - \lambda)\mathrm{y}) \leq \lambda f(\mathrm{x}) + (1 - \lambda) f(\mathrm{y}) \leq \alpha$ for all $0 \leq \lambda \leq 1$. Thus, $\lambda \mathrm{x} + (1 - \lambda)\mathrm{y} \in S_\alpha$ for all $\lambda \in [0, 1]$. □

A.2.1 Differentiable functions.

When a function f is differentiable we can characterize convexity in the following way.

Theorem A.2.5 (Convexity: first order condition)**.** *A differentiable function $f : \mathcal{X} \to \mathbb{R}$ is convex if and only if $\mathcal{X}$ is convex and*

$$f(\mathrm{x}) \geq f(\mathrm{x}_0) + \nabla f(\mathrm{x}_0) \cdot (\mathrm{x} - \mathrm{x}_0) \textit{ for all } \mathrm{x}, \mathrm{x}_0 \in \mathcal{X}. \tag{A.2}$$

The first-order condition states that the linear approximation given by the first-order Taylor approximation gives a global underestimator of the function. Put differently, the local information, i.e., the function value and gradient at that point, provide global information about a convex function.

The first-order condition gives the following characterization of the globally optimal point: if the gradient of a convex function f vanishes at some point $\mathrm{x}_0 \in \mathcal{X}$, then x_0 must be a global minimizer of f:

$$\nabla f(\mathrm{x}_0) = 0 \qquad \Longrightarrow \qquad \mathrm{x}_0 \in \arg\min_{\mathrm{x} \in \mathcal{X}} f(\mathrm{x}).$$

The first-order condition states that the linear approximation given by the first-order Taylor approximation gives a global underestimator of the function. Put differently, the local information, i.e., the function value and gradient at that point, provide global information about a convex function.

Theorem A.2.6 (Convexity: second order condition)**.** *If f is a twice continuously differentiable on $\mathcal{X} \subseteq \mathbb{R}^d$. Then, f is convex if and only if $\nabla^2 f(\mathrm{x}) \succeq 0$ for all $\mathrm{x} \in \mathcal{X}$.*

Next, we show that a gradient (equivalently, a subgradient if f is not differentiable) defines a supporting hyperplane to sublevel sets.

Claim A.2.7. *Let $f(\mathrm{x}_0) = \alpha$. If $\nabla f(\mathrm{x}_0) \neq 0$, then $S_\alpha \subseteq \{\mathrm{x} \mid \langle \nabla f(\mathrm{x}_0), \mathrm{x} - \mathrm{x}_0 \rangle \leq 0\}$.*

Proof. Let $\mathrm{x} \in S_\alpha$. Then, $f(\mathrm{x}) \leq f(\mathrm{x}_0)$. Since f is convex, we have that $f(\mathrm{x}) \geq f(\mathrm{x}_0) + \langle \triangledown f(\mathrm{x}_0), \mathrm{x} - \mathrm{x}_0 \rangle$. Rearranging the terms we get,

$$\begin{aligned} \langle \triangledown f(\mathrm{x}_0), \mathrm{x} \rangle &\leq \langle \triangledown f(\mathrm{x}_0), \mathrm{x}_0 \rangle + (f(\mathrm{x}) - f(\mathrm{x}_0)) \\ &\leq \langle \triangledown f(\mathrm{x}_0), \mathrm{x}_0 \rangle, \end{aligned}$$

which implies that $\mathrm{x} \in \{\mathrm{x} \mid \langle \nabla f(\mathrm{x}_0), \mathrm{x} - \mathrm{x}_0 \rangle \leq 0\}$. □

This is a particularly important result, since if we are at x_0 and want to minimize f, then the gradient excludes the halfspace $\{\mathrm{x} \mid \langle \nabla f(\mathrm{x}_0), \mathrm{x} - \mathrm{x}_0 \rangle \geq 0\}$ from the search space.

A.2.2 Strict convexity, strong convexity, and smoothness

We assume that the function f is differentiable. Strict convexity and strong convexity can be defined for at points of non-differentiability through use of a sub-gradient.

Definition A.2.8 (Strict convexity)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is strictly convex iff $\forall \mathrm{x}, \mathrm{y} \in \mathcal{X}$ it holds that

$$f(\mathrm{y}) > f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{y} - \mathrm{x} \rangle.$$

Strict convexity is similar to convexity, however, requires that the inequality is strict in the lower bound. Strong convexity further requires that the function is lower bounded by a quadratic.

Definition A.2.9 (Strong convexity)**.** A differentiable function $f : \mathcal{X} \to \mathbb{R}$ is α-***strongly convex*** if for some $\alpha > 0$, and all $\mathrm{x}, \mathrm{y} \in \mathcal{X}$,

$$f(\mathrm{y}) \geq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{y} - x \rangle + \frac{\alpha}{2} \|\mathrm{y} - \mathrm{x}\|_2^2. \tag{A.3}$$

For twice differentiable functions, we can characterize these different notions in terms of a condition on the directional second derivatives, or equivalently as a condition on the eigenvalues of the Hessian, $\nabla^2 f(\mathrm{x})$. Recall, that a twice differentiable function is convex if and only if $\nabla^2 f(\mathrm{x}) \succeq 0$ for all $\mathrm{x} \in \mathcal{X}$. A function is strictly convex if and only if $\nabla^2 f(\mathrm{x}) \succ 0$, and strongly convex if and only if $\nabla^2 f(\mathrm{x}) \succeq \alpha \mathrm{I}$.

Next, we quickly review different equivalent conditions for strong convexity.

Proposition A.2.8 (Equivalent conditions for strong convexity)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. Then, a differentiable function $f : \mathcal{X} \to \mathbb{R}$, is α-strongly convex, for some $\alpha > 0$, if any of the following holds.*

1. $f(\mathrm{y}) \geq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{y} - x \rangle + \frac{\alpha}{2}\|\mathrm{y} - \mathrm{x}\|_2^2 \quad \forall \quad \mathrm{x}, \mathrm{y} \in \mathcal{X}.$

2. $h(\mathrm{x}) = f(\mathrm{x}) - \frac{\alpha}{2}\|\mathrm{x}\|^2$ *is convex.*

3. $\langle \nabla f(\mathrm{x}) - f(\mathrm{y}), \mathrm{x} - \mathrm{y} \rangle \geq \alpha\|\mathrm{x} - \mathrm{y}\|^2, \quad \forall \quad \mathrm{x}, \mathrm{y} \in \mathcal{X}.$

4. $f(\lambda\mathrm{x} + (1-\lambda)\mathrm{y}) \leq \lambda f(\mathrm{x}) + (1-\lambda)f(\mathrm{y}) - \frac{\lambda(1-\lambda)}{2}\alpha\|\mathrm{x} - \mathrm{y}\|^2 \quad \forall \quad \mathrm{x}, \mathrm{y} \in \mathcal{X}, \lambda \in [0, 1].$

Further if f is twice differentiable then it is α-strongly convex iff

6. $\nabla^2 f(\mathrm{x}) \succeq \alpha \mathrm{I}.$

Next, we discuss properties of smooth convex functions. Smoothness is a dual notion to strong convexity as we will shortly see. The notion intuitively states that the function f is upper bounded by a quadratic.

Definition A.2.10 (Smoothness)**.** A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β***-smooth*** if the gradient map $\nabla f \colon \mathbb{R}^d \to \mathbb{R}^d$ is β-Lipschitz, i.e, for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|\,. \tag{A.4}$$

The following equivalence holds.

Theorem A.2.9 (Quadratic upper bound for smooth functions)**.** *A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth iff for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$ it holds that*

$$f(\mathrm{y}) \leq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{y} - \mathrm{x} \rangle + \frac{\beta}{2}\|\mathrm{x} - \mathrm{y}\|^2.$$

Further all convex β-smooth functions satisfy for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$

$$f(\mathrm{y}) \geq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{y} - \mathrm{x} \rangle + \frac{\beta}{2}\|\nabla f(\mathrm{x}) - \nabla f(\mathrm{y})\|^2.$$

The second part of the above theorem is very similar to the strong convexity definition. This inequality can actually be seen as a corollary of a result which characterizes strong convexity and smoothness as dual properties. To state this result formally we need to introduce the ***Fenchel conjugate*** of a convex function f, which we do in Section A.3.3.

A.2.3 Jensen's Inequality

We now present a fundamental inequality for convex functions.

Theorem A.2.10 (Jensen's Inequality)**.** *Let X be a random variable taking values in a non-empty convex set $\mathcal{X} \subseteq \mathbb{R}^d$ with a finite expectation* $\mathrm{E}[X]$*, and f a measurable convex function defined over $\mathcal{X}$. Then,* $\mathrm{E}[X]$ *is in $\mathcal{X}$,* $\mathrm{E}[f(X)]$ *is finite, and the following inequality holds:*

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)].$$

A.3 Potential functions and Bregman divergence and Mirror descent

Mirror descent is a generalization of the popular gradient descent procedure for finding an approximate minimizer of a convex function. The gradient descent update for a function f at a point x_t can be summarized as follows

$$\mathrm{x}_{t+1} = \mathrm{x}_t - \eta \nabla f(\mathrm{x}_t),$$

where η is the step-size parameter. The motivation behind this update is that $-\nabla f(\mathrm{x}_t)$ is a descent direction for f if f is a convex function, thus selecting a sufficiently small step size η would result in an iterate x_{t+1} s.t. $f(\mathrm{x}_t) \geq f(\mathrm{x}_{t+1})$.

if we are working in a Banach space, i.e., with a geometry defined with respect to a norm (or a distance function) that is not induced by an inner product, then we cannot employ the gradient descent strategy. Instead, Nemirovsky and Yudin (1983) propose the following method. We can first map the point $\mathrm{x} \in \mathcal{X}$ to the dual space $\mathcal{X}^*$, then perform the gradient update in the dual space and finally map the resulting point back to the primal space. At each update, the new point in the primal space $\mathcal{X}$ might be outside the constraint set $\mathcal{C} \subset \mathcal{X}$ in which case it should be projected into the constraint set $\mathcal{C}$. We will define a new geometry using a function which we will refer to as the potential function $\Phi(\mathrm{x})$ and use Bregman projection based on Bregman divergence to define this geometry.

A.3.1 The Geometry of ℓ_p Norms

We begin by recalling the definition of the gradient. Given a function $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$, the gradient of f at x_0 is the linear operator $\nabla f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\lim_{\mathrm{x}\to\mathrm{x}_0,\ \mathrm{x}\in\mathcal{X}} \frac{|f(\mathrm{x}) - f(\mathrm{x}_0) - \nabla f[\mathrm{x} - \mathrm{x}_0]|}{\|\mathrm{x} - \mathrm{x}_0\|} = 0.$$

The gradient is a continuous linear functional on $\mathbb{R}^d$, i.e., ∇f lives in the dual space $\mathcal{X}^*$ of $\mathcal{X}$. A natural question to ask is how to identify elements of $\mathcal{X}$ with those in $\mathcal{X}^*$. The canonical identification of $\mathcal{X}$ and $\mathcal{X}^*$ is given as follows. Let $\{\mathrm{e}_i\}_{i=1}^d$ be the standard basis for $\mathcal{X}$. Construct a basis $\{\mathrm{e}_i^*\}_{i=1}^d$ for $\mathcal{X}^*$ such that $\mathrm{e}_i^*(\mathrm{e}_j) = \langle \mathrm{e}_i^*, \mathrm{e}_j\rangle = \delta_{i,j}$. In fact, this is the *natural* identification induced by the standard inner product on $\mathbb{R}^d$ (follows from the Riesz Representation Theorem). Indeed, this is what motivated the gradient descent update, i.e., identifying the descent direction with the negative gradient.

However, this also motivates the following question. If the standard inner product, which induces the Euclidean geometry on $\mathcal{X}$, gives us the canonical identification, what identification would a different geometry yield? For example, what happens if we consider the geometry induced by ℓ_p norms on $\mathcal{X}$? For simplicity, assume that $1 < p < \infty$, so that the derivative of the ℓ_p norm is continuously differentiable. Consider the ℓ_p-norm on $\mathcal{X}$. Then the dual norm induced on $\mathcal{X}^*$ is the ℓ_q-norm such that $\frac{1}{p} + \frac{1}{q} = 1$. Consider $\psi^* : \mathcal{X}^* \to \mathbb{R}$ that maps $\phi \in \mathcal{X}^*$ to $\psi^*(\phi) = \frac{1}{2}\|\phi\|_*^2$. This map is continuously differentiable and it induces the following identification $\nabla\psi^* : \mathcal{X}^* \to \mathcal{X}$ that maps $\phi \in \mathcal{X}^*$ to $\nabla\psi^*(\phi) \in \mathcal{X}$. We next show that the map $\nabla\psi^*$ is a one-to-one correspondence. First, note that by convexity of $\|\cdot\|_*$,

$$0 \geq \|\phi\|_* + \langle \nabla\|\phi\|_*, 0 - \phi\rangle \iff \langle \nabla\|\phi\|_*, \phi\rangle \geq \|\phi\|_*$$

$$2\|\phi\|_* \geq \|\phi\|_* + \langle \nabla\|\phi\|_*, 2\phi - \phi\rangle \implies \langle \nabla\|\phi\|_*, \phi\rangle \leq \|\phi\|_*$$

Then, by definition of the dual norm we know that $\nabla\|\phi\|_*$ is the maximizer of the linear functional ϕ over the $\|\cdot\|$-unit ball. We can also show that $\nabla\|\phi\|_*$ is the unique maximizer (Using Corollary 1.3 from Brezis (2010) and the remark that follows). Using the chain rule, we get $\nabla\psi^*(\phi) = \|\phi\|_*\nabla\|\phi\|_*$. So ϕ is mapped to the element which maximizes the linear functional ϕ over the ball of size $\|\phi\|_*$. It remains to be shown that this mapping is a bijection. Define $\psi : \mathcal{X} \to \mathbb{R}$ as $\psi(\mathrm{x}) = \frac{1}{2}\|\mathrm{x}\|^2$. By the exact same reasoning as for ψ^* we have that $\nabla\psi(\mathrm{x})$ is the unique maximizer of $\langle\phi, \mathrm{x}\rangle$ over $\|\phi\|_* \leq \|\mathrm{x}\|$. We only need to show that $\nabla\psi^*(\nabla\psi(\mathrm{x})) = \mathrm{x}$. Well $\nabla\psi^*(\nabla\psi(\mathrm{x})) = \arg\max_{\|\tilde{\mathrm{x}}\|=\|\nabla\psi(x)\|_*}\langle\tilde{\mathrm{x}}, \nabla\psi(\mathrm{x})\rangle = \arg\max_{\|\tilde{\mathrm{x}}\|=\|\mathrm{x}\|}\langle\tilde{\mathrm{x}}, \nabla\psi(\mathrm{x})\rangle$, since$\nabla\psi(\mathrm{x}) = \arg\max_{\|\phi\|_*=\|\mathrm{x}\|}\langle\phi, \mathrm{x}\rangle$. By corollary 1.4 from Brezis (2010) we know that $\langle\nabla\psi(\mathrm{x}), \mathrm{x}\rangle = \|\mathrm{x}\|^2$. On the other hand $\langle\tilde{\mathrm{x}}, \nabla\psi(\mathrm{x})\rangle \leq \|\tilde{\mathrm{x}}\|\|\nabla\psi(\mathrm{x})\| = \|\mathrm{x}\|^2$ for all $\|\tilde{\mathrm{x}}\| = \|\mathrm{x}\|$. By the remark after Corollary 1.3 Brezis (2010) x is the unique maximizer of $\max_{\|\tilde{\mathrm{x}}\|=\|\mathrm{x}\|}\langle\tilde{\mathrm{x}}, \nabla\psi(\mathrm{x})\rangle$.

We can now adapt our GD procedure to the identification induced by ϕ^* and ϕ in the following simple way:

$$\mathrm{x}_{t+1} = \nabla\psi^*(\nabla\psi(\mathrm{x}_t) - \eta\nabla f(\mathrm{x}_t)). \tag{A.5}$$

How can we adapt the analysis? Well we need to choose the appropriate potential function with which to track the progress of our algorithm. Let us try to see why we chose $\|\cdot - \mathrm{x}^*\|_2^2$ as our potential. Notice that by convexity we have that $2\langle\nabla f(\mathrm{x}), \mathrm{x}^* - \mathrm{x}\rangle \leq 0$, however, this is also the derivative of $\|\mathrm{x} - \mathrm{x}^*\|^2$ in the direction of $\nabla f(\mathrm{x})$. Thus $-\nabla f(\mathrm{x})$ is a descent direction for the smooth function $\|\mathrm{x} - \mathrm{x}^*\|^2$. Can we come up with a similar smooth function for $\langle\nabla f(\mathrm{x}), \mathrm{x}^* - \nabla\psi^*(\phi_\mathrm{x})\rangle$, where $\mathrm{x} = \nabla\psi^*(\phi_\mathrm{x})$. Well we can get the term $\mathrm{x}^* - \nabla\psi^*(\phi_\mathrm{x})$ as the derivative of $\langle\phi_\mathrm{x}, \mathrm{x}^*\rangle - \psi^*(\phi_\mathrm{x})$. Being a bit smarter we realize that

$$\langle\nabla f(\mathrm{x}), \mathrm{x}^* - \nabla\psi^*(\phi_\mathrm{x})\rangle = \left.\frac{\partial\psi^*(\phi_\mathrm{x} - t\nabla f(\mathrm{x})) - \langle\phi_\mathrm{x} - t\nabla f(\mathrm{x}), \mathrm{x}^*\rangle}{\partial t}\right|_{t=0},$$

and thus a good candidate for a potential function becomes $V(\phi) = \psi^*(\phi) - \langle\phi, \mathrm{x}^*\rangle$.

Let us see how we can analyze the update in A.5. First we use the fact that ψ^* is $\mathcal{L}$-smooth for some L. This follows from the fact ψ^* is the *Fenchel conjugate* of ψ and the fact that ψ is $\frac{1}{\mathcal{L}}$ strongly convex. Let $\phi_t = \nabla\phi(\mathrm{x}_t)$. The smoothness of ψ^* implies that V is also smooth with the same smoothness parameter L, since it is just a translation of ψ^*. We now have the following derivation:

$$\begin{aligned} V(\phi_{t+1}) = V(\phi_t - \eta\nabla f(\mathrm{x}_t)) &\leq V(\phi_t) - \eta\langle\nabla f(\mathrm{x}_t), \nabla\psi^*(\phi_t) - \mathrm{x}^*\rangle + \frac{L\eta^2}{2}\|\nabla f(\mathrm{x}_t)\|_*^2 \\ &\leq V(\phi_t) - \eta(f(\mathrm{x}_t) - f(\mathrm{x}^*)) + \frac{L\mathcal{L}\eta^2}{2} \\ &\implies f(\bar{\mathrm{x}}) - f(\mathrm{x}^*) \leq \frac{V(\phi_1) - V(\phi_{t+1}) + T\frac{L\mathcal{L}\eta^2}{2}}{2T\eta}. \end{aligned}$$

A.3.2 Mirror Maps

Let us dissect the above reasoning – there are two key parts to the proof. First the function ψ needs to be such that $\nabla\psi$ is a bijection between $\mathcal{X}$ and $\mathcal{X}^*$. Secondly we needed that this function is smooth so we can construct a smooth potential function with which to track the progress of our algorithm. We used properties of ℓ_p norms heavily when deriving these two statements, however, many other functions posses the required properties. We are going to call any function which posses the first property i.e. invertible continuous gradient a mirror map. Any mirror map is now going to induce an identification of the primal and dual spaces. In what follows the focus is only going to be on mirror maps ψ, which are α-strongly convex functions. We note that for the invertible property to hold we only need strict convexity, however, the strong convexity is going to allow us to construct a $\frac{1}{\alpha}$-smooth potential function with which to track progress.

A.3.3 The Fenchel Dual

It the short exercise above we saw that when $\psi \equiv \frac{1}{2}\|\cdot\|_p^2$, then $\nabla^{-1}\psi(\phi) = \nabla\psi^*(\phi)$ where $\psi^* \equiv \frac{1}{2}\|\cdot\|_q^2$, $\frac{1}{p} + \frac{1}{q} = 1$. In particular the inverse of the gradient operator was given by the gradient of the dual norm. We are now going to show how to use duality to get a similar result for general α-strongly convex functions ψ.

Definition A.3.1. The Fenchel dual of a proper function $\psi : \mathcal{X} \to \mathbb{R} \bigcup \{\infty\}$ is $\psi^* : \mathcal{X}^* \to \mathbb{R} \bigcup \{\infty\}$ such that:

$$\psi^*(\phi) = \sup_{\mathrm{x} \in \mathcal{X}} \langle \phi, \mathrm{x} \rangle - \psi(\mathrm{x}). \tag{A.6}$$

By definition ψ^* is convex and lower semi-continuous on $\mathcal{X}^*$. We now state two results for completeness without proof.

Lemma A.3.1 (Proposition 1.10 Brezis (2010))**.** *Assume $\psi : \mathcal{X} \to \mathbb{R}$ is convex and lower semi-continuous and that $\psi \not\equiv \infty$. Then $\psi^* \not\equiv \infty$. In particular ψ is bounded from below by an affine continuous function.*

Theorem A.3.2 (Fenchel–Moreau)**.** *Assume that ψ is convex, lower semi-continuous and $\psi \not\equiv \infty$. Then $\psi^{**} = \psi$.*

In the example we saw in the beginning of this section $\psi(\mathrm{x}) = \frac{1}{2}\|\mathrm{x}\|^2$ and $\psi^*(\mathrm{y}) = \frac{1}{2}\|y\|_*^2$. We also saw that the inverse of the gradient mapping $\nabla\psi$ was $\nabla\psi^*$. Next we are going to show that this is true not only for norms but for general ψ, which are α-strongly convex and continuously differentiable.

Lemma A.3.3. *If ψ is continuously differentiable and α-strongly convex then $\nabla\psi^*(\nabla\psi(\mathrm{x})) =$ x*

Proof. The proof is not much different from what we did in the case when ψ was the squared norm. First notice that since ψ is strongly convex then the maximizer of $\sup_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{y}, \mathrm{x} \rangle - \psi(\mathrm{x})$ exists and is unique. Let this maximizer be x^*. From first order optimality we know that $\nabla\psi(\mathrm{x}^*) = \mathrm{y}$. Now by definition we know that $\psi(\mathrm{x}^*) + \psi^*(\mathrm{y}) = \langle \mathrm{y}, \mathrm{x}^* \rangle$. On the other hand, using Fenchel-Moreau we have that $\psi^{**}(\mathrm{x}^*) + \psi^*(\mathrm{y}) = \langle \mathrm{y}, \mathrm{x}^* \rangle$ and thus y is a maximizer of $\sup_{\mathrm{y}\in\mathcal{X}^*} \langle \mathrm{y}, \mathrm{x}^* \rangle - \psi^*(\mathrm{y})$. This implies that $\nabla\psi^*(\mathrm{y}) = \mathrm{x}^*$ and thus $\nabla\psi^*(\nabla\psi(\mathrm{x}^*)) = \mathrm{x}^*$. □

We note that the above lemma also holds true if we do not assume that ψ is continuously differentiable or differentiable at all. We only need to assume ψ is lower semi-continuous, strictly convex and $\psi \not\equiv \infty$. Recall that in the convergence proof of mirror descent we used the fact that ψ^* is a smooth function. We now show that such an assumption is justified, provided that ψ is α-strongly convex.

Lemma A.3.4. *If ψ is continuously differentiable and α-strongly convex then ψ^* is $\frac{1}{\alpha}$-smooth.*

Proof. Let x_1 and x_2 be elements of the primal space which are maximizers of $\sup_{\mathrm{x}\in\mathcal{X}}\langle \mathrm{y}_i, \mathrm{x}\rangle - \psi(\mathrm{x})$ for $i = 1$ and $i = 2$ respectively. By strong convexity we have:

$$\psi(\mathrm{x}_2) - \psi(\mathrm{x}_1) - \langle \mathrm{y}_1, \mathrm{x}_2 - \mathrm{x}_1\rangle \geq \frac{\alpha}{2}\|\mathrm{x}_2 - \mathrm{x}_1\|^2$$
$$\psi(\mathrm{x}_1) - \psi(\mathrm{x}_2) - \langle \mathrm{y}_2, \mathrm{x}_1 - \mathrm{x}_2\rangle \geq \frac{\alpha}{2}\|\mathrm{x}_1 - \mathrm{x}_2\|^2.$$

Summing and using Holder's inequality we have

$$\|\mathrm{y}_1 - \mathrm{y}_2\|_*\|\mathrm{x}_1 - \mathrm{x}_2\| \geq \alpha\|\mathrm{x}_1 - \mathrm{x}_2\|^2.$$

Plugging in the fact $\mathrm{x}_i = \nabla\psi^*(\mathrm{y}_i)$ we conclude the proof. □

We note that the opposite relation between smoothness and strong-convexity also holds, i.e., if ψ is β-smooth, then ψ^* is $1/\beta$-strongly convex (Kakade et al., 2009). We now have all the ingredients to repeat the proof of convergence for mirror descent when the identification is induced by an α-strongly convex ψ.

A.3.4 Analysis in the primal space and Bregman divergence

Recall the potential function which tracked the progress of MD: $V(\mathrm{y}) = \psi^*(\mathrm{y}) - \langle \mathrm{y}, \mathrm{x}^*\rangle$. This function tracks the progress of the algorithm in the dual space in the sense that we bound $V(\nabla\psi(\mathrm{x}_{t+1})) - V(\nabla\psi(\mathrm{x}_{t+1}))$. Can we carry out the same analysis in the primal space i.e. is there a potential function $W : \mathcal{X} \to \mathbb{R}$ such that we can track the progress of the algorithm via $W(\mathrm{x}_{t+1}) - W(\mathrm{x}_t)$? We now show one such function as follows. Let $\mathrm{y} = \nabla\psi(\mathrm{x})$ so that $\psi^*(\mathrm{y}) + \psi(\mathrm{x}) = \langle \mathrm{y}, \mathrm{x}\rangle$. We have

$$V(\mathrm{y}) = \langle \mathrm{y}, \mathrm{x}\rangle - \psi(\mathrm{x}) - \langle \mathrm{y}, \mathrm{x}^*\rangle = \langle \nabla\psi(\mathrm{x}), \mathrm{x} - \mathrm{x}^*\rangle - \psi(\mathrm{x}).$$

Let us "translate" the analysis from the dual to the primal space using this new potential $W(\mathrm{x}) = \langle \nabla\psi(\mathrm{x}), \mathrm{x} - \mathrm{x}^*\rangle - \psi(\mathrm{x})$.

$$\begin{aligned} W(\mathrm{x}_{t+1}) - W(\mathrm{x}_t) &= \psi(\mathrm{x}_t) - \psi(\mathrm{x}_{t+1}) + \langle \nabla\psi(\mathrm{x}_{t+1}), \mathrm{x}_{t+1} - \mathrm{x}^*\rangle - \langle \nabla\psi(\mathrm{x}_t), \mathrm{x}_t - \mathrm{x}^*\rangle \\ &= \psi(\mathrm{x}_t) - \psi(\mathrm{x}_{t+1}) + \langle \nabla\psi(\mathrm{x}_t) - \eta\nabla f(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}^*\rangle - \langle \nabla\psi(\mathrm{x}_t), \mathrm{x}_t - \mathrm{x}^*\rangle \\ &= -\eta\langle \nabla f(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}^*\rangle + \psi(\mathrm{x}_t) - \psi(\mathrm{x}_{t+1}) + \langle \nabla\psi(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}_t\rangle \\ &\leq -\eta\langle \nabla f(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}^*\rangle - \frac{\alpha}{2}\|\mathrm{x}_{t+1} - \mathrm{x}_t\|^2 \\ &= \eta\langle \nabla f(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_t\rangle - \eta\langle \nabla f(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}_t\rangle - \frac{\alpha}{2}\|\mathrm{x}_{t+1} - \mathrm{x}_t\|^2 \\ &\leq f(x^*) - f(\mathrm{x}_t) + \eta\|\nabla f(\mathrm{x}_t)\|_*\|\mathrm{x}_t - \mathrm{x}_{t+1}\| - \frac{\alpha}{2}\|\mathrm{x}_t - \mathrm{x}_{t+1}\|^2 \\ &\leq f(x^*) - f(\mathrm{x}_t) + \frac{\eta^2}{2\alpha}\|\nabla f(\mathrm{x}_t)\|_*^2, \end{aligned}$$

where in the first inequality we have used the strong convexity of ψ in the second inequality we have used Holder's inequality and the third inequality follows from the fact that $2xy - y^2 \leq x^2$. To finish the analysis we only need to telescope the LHS and take the average on the RHS.

Even though we now have an analysis in the primal space the potential function we used seems a bit peculiar. If we didn't have the potential in the dual space it is very unlikely we come up with exactly W. Staring at the form of W we notice that we are a single term away from the *Bregman divergence* induced by ψ. In particular $W(x) + \psi(\mathrm{x}^*) = D_\psi(\mathrm{x}^*, \mathrm{x})$. Let us formally define the Bregman divergence induced by ψ.

Definition A.3.2. Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function. The Bregman divergence associated with ψ is:

$$D_{\psi}(\mathrm{x}_1, \mathrm{x}_2) := \psi(\mathrm{x}_1) - \psi(\mathrm{x}_2) - \langle \partial\psi(\mathrm{x}_2), \mathrm{x}_1 - \mathrm{x}_2 \rangle. \tag{A.7}$$

We now list some useful properties of the Bregman divergence of a strictly/strongly convex function ψ:

1. D_{ψ} is strictly convex in its first argument.

2. $D_{\psi}(\mathrm{x}, \mathrm{y}) \geq 0, \forall \mathrm{x}, \mathrm{y}$ and $D_{\psi}(\mathrm{x}, \mathrm{y}) = 0$ iff $\mathrm{x} = \mathrm{y}$.

3. In general $D_{\psi}(\mathrm{x}, \mathrm{y}) \neq D_{\psi}(\mathrm{y}, \mathrm{x})$ e.g. consider $\psi(\mathrm{x}) = \sum_i \mathrm{x}_i \log(\mathrm{x}_i)$.

4. In general $D_{\psi}(\mathrm{x}, \mathrm{y})$ is non-convex in its second argument e.g. $\psi(x) = -\log(x)$.

5. Linearity in ψ i.e. $D_{\psi + \alpha\phi}(\mathrm{x}, \mathrm{y}) = D_{\psi}(\mathrm{x}, \mathrm{y}) + \alpha D_{\phi}(\mathrm{x}, \mathrm{y})$.

6. $\frac{\partial D_{\psi}(\mathrm{x}, \mathrm{y})}{\partial \mathrm{x}} = \nabla\psi(\mathrm{x}) - \nabla\psi(\mathrm{y})$.

7. $D_{\psi}(\mathrm{x}, \mathrm{y}) + D_{\psi}(\mathrm{y}, \mathrm{z}) = D_{\psi}(\mathrm{x}, \mathrm{z}) + \langle \mathrm{x} - \mathrm{y}, \nabla\psi(\mathrm{z}) - \nabla\psi(\mathrm{y}) \rangle$.

8. If ψ is α-strongly convex with respect to $\|\cdot\|$ then $D_{\psi}(\mathrm{x}, \mathrm{y}) \geq \frac{\alpha}{2}\|\mathrm{x} - \mathrm{y}\|^2$.

9. $D_{\psi}(\mathrm{x}, \mathrm{y}) = D_{\psi^*}(\nabla\psi(\mathrm{y}), \nabla\psi(\mathrm{x}))$.

All of the above properties are easy to show by direct algebra and for the last property using Lemma A.3.3. Let us look at 2 standard examples for Bregman divergences. First consider the standard ℓ_2 norm $\|\cdot\|$. The strongly convex function inducing the Bregman divergence is given by $\psi(x) = \frac{1}{2}\|\mathrm{x}\|^2$. Now using the definition for divergence we have:

$$\begin{aligned} D_{\psi}(\mathrm{x}, \mathrm{y}) &= \frac{1}{2}\|\mathrm{x}\|^2 - \frac{1}{2}\|\mathrm{y}\|^2 - \langle \mathrm{y}, \mathrm{x} - \mathrm{y} \rangle \\ &= \frac{1}{2}\|\mathrm{x}\|^2 + \frac{1}{2}\|\mathrm{y}\|^2 - \langle \mathrm{y}, \mathrm{x} \rangle = \frac{1}{2}\|\mathrm{x} - \mathrm{y}\|^2. \end{aligned}$$

Since we begun our discussion with wanting a different geometry on $\mathcal{X}$ than the Euclidean one, let us look at what happens when we choose $\psi(\mathrm{x}) = \sum_{i=1}^{d} \mathrm{x}_i \log(\mathrm{x}_i)$.

We are also going to constrain our space to the set $C = \{\mathrm{x} \in \mathbb{R}^n_+ : \|\mathrm{x}\|_1 = 1\}$. The Bregman divergence in this case is known as Kullback-Leibler divergence and takes the form:

$$\begin{aligned} D_\psi(\mathrm{x}, \mathrm{y}) &= \sum_{i=1}^d \mathrm{x}_i \log(\mathrm{x}_i) - \sum_{i=1}^d \mathrm{y}_i \log(\mathrm{y}_i) - \sum_{i=1}^d (\mathrm{x}_i - \mathrm{y}_i) \log(\mathrm{y}_i) \\ &= \sum_{i=1}^d \mathrm{x}_i \log(\mathrm{x}_i / \mathrm{y}_i). \end{aligned}$$

We now show a very useful property for analyzing proximal methods which use Bregman divergence as the proximity map.

Lemma A.3.5. *Let f be a convex function such that $f \not\equiv \infty$. Suppose f has domain an open set containing the convex set C. Suppose ψ is α-strongly convex and let:*

$$\mathrm{x}^* = arg \min_{\mathrm{x} \in C} \{f(\mathrm{x}) + D_\psi(\mathrm{x}, \mathrm{x}_0)\}.$$

For any $\mathrm{y} \in C$ we have:

$$f(\mathrm{y}) + D_\psi(\mathrm{y}, \mathrm{x}_0) \geq f(\mathrm{x}^*) + D_\psi(\mathrm{x}^*, \mathrm{x}_0) + D_\psi(\mathrm{y}, \mathrm{x}^*).$$

Proof. By optimality for constraint optimization we have that there exists a subgradient d of $f(\mathrm{x}^*) + D_\psi(\mathrm{x}^*, \mathrm{x}_0)$ such that for all $\mathrm{x} \in C$ it holds that $\langle \mathrm{d}, \mathrm{x} - \mathrm{x}^* \rangle \geq 0$ or equivalently there exists a subgradient g of $f(\mathrm{x}^*)$ such that

$$\langle \mathrm{g} + \nabla\psi(\mathrm{x}^*) - \nabla\psi(\mathrm{x}_0), \mathrm{x} - \mathrm{x}^* \rangle \geq 0,$$

for all $\mathrm{x} \in C$. Using the fact that f is convex we have:

$$\begin{aligned} f(\mathrm{y}) \geq f(\mathrm{x}^*) + \langle \mathrm{g}, \mathrm{y} - \mathrm{x}^* \rangle &\geq f(\mathrm{x}^*) + \langle \nabla\psi(\mathrm{x}^*) - \nabla\psi(\mathrm{x}_0), \mathrm{y} - \mathrm{x}^* \rangle \\ &= f(\mathrm{x}^*) - \langle \nabla\psi(\mathrm{x}_0), \mathrm{x}^* - \mathrm{x}_0 \rangle + \psi(\mathrm{x}^*) - \psi(\mathrm{x}_0) + \langle \psi(\mathrm{x}_0), \mathrm{y} - \mathrm{x}_0 \rangle - \psi(\mathrm{y}) + \psi(\mathrm{x}_0) \\ &- \langle \nabla\psi(\mathrm{x}^*), \mathrm{y} - \mathrm{x}^* \rangle + \psi(\mathrm{y}) - \psi(\mathrm{x}^*) \\ &= f(\mathrm{x}^*) + D_\psi(\mathrm{x}^*, \mathrm{x}_0) - D_\psi(\mathrm{y}, \mathrm{x}_0) + D_\psi(\mathrm{y}, \mathrm{x}^*). \end{aligned}$$

□

As a direct corollary we get the General Pythagorean Theorem:

Theorem A.3.6. *If* x^* *is the projection of* x_0 *onto the convex set* C *with respect to the Bregman divergence induced by* ψ *i.e.*

$$\mathrm{x}^* = arg\min_{\mathrm{x}\in C} D_\psi(\mathrm{x}, \mathrm{x}_0),$$

then $D_\psi(\mathrm{y}, \mathrm{x}_0) \geq D_\psi(\mathrm{y}, \mathrm{x}^*) + D_\psi(\mathrm{x}^*, \mathrm{x}_0)$*.*

A.3.5 Mirror descent as proximal gradient descent

Recall that one motivation for the gradient descent update came from the following observation. If f is convex then $-\nabla f(\mathrm{x}_t)$ is the steepest direction of descent near an infinitesimal region around x_t. If we penalize moving away from x_t via a term which captures the geometry of the space, we are still likely to decrease the objective. This led to choosing the next step x_{t+1} as the minimizer of $\min_{\mathrm{x}\in\mathcal{X}}\langle\nabla f(\mathrm{x}_t), \mathrm{x}-\mathrm{x}_t\rangle + \frac{1}{\eta}\|\mathrm{x}-\mathrm{x}_t\|^2$. Even though we now should be convinced that this is not really the correct reasoning behind gradient descent, it is still natural to ask what would happen if we replaced the Euclidean norm squared by the Bregman divergence of some strongly convex ψ. After all from the properties we have seen so far Bregman divergence almost acts like a norm on $\mathcal{X}$. We now show what happens when x_{t+1} is chosen as the minimizer of $\langle\nabla f(\mathrm{x}_t), \mathrm{x}-\mathrm{x}_t\rangle + \frac{1}{\eta}D_\psi(\mathrm{x}, \mathrm{x}_t)$. It turns out that this step exactly recovers the mirror descent update.

Lemma A.3.7. *Let* $\mathrm{x}_{t+1} = arg\min_{\mathrm{x}\in\mathcal{X}}\{\langle\nabla f(\mathrm{x}_t), \mathrm{x}-\mathrm{x}_t\rangle + \frac{1}{\eta}D_\psi(\mathrm{x}, \mathrm{x}_t)\}$*. Then* $\mathrm{x}_{t+1} = \nabla\psi^*(\nabla\psi(\mathrm{x}_t) - \nabla f(\mathrm{x}_t))$*.*

Proof. Since $D_\psi(\mathrm{x}, \mathrm{x}_t)$ is strongly convex in x then $\langle\nabla f(\mathrm{x}_t), \mathrm{x}-\mathrm{x}_t\rangle + \frac{1}{\eta}D_\psi(\mathrm{x}, \mathrm{x}_t)$ is strongly convex in x and thus has a unique minimizer. Using first order optimality condition we have:

$$\begin{aligned}&\eta\nabla f(\mathrm{x}_t) + (\nabla\psi(\mathrm{x}_{t+1}) - \nabla\psi(\mathrm{x}_t)) = 0\\ \iff& \mathrm{x}_{t+1} = \nabla\psi^{-1}(\nabla\psi(\mathrm{x}_t) - \eta\nabla f(\mathrm{x}_t)) = \nabla\psi^*(\nabla\psi(\mathrm{x}_t) - \eta\nabla f(\mathrm{x}_t)).\end{aligned}$$

□

A.3.6 Online projected mirror descent

In this section we discuss how to obtain regret guarantees for the online convex optimization problem. Suppose instead of wanting to minimize a convex function $f(\mathrm{x}_t)$ and observing its gradient $\nabla f(\mathrm{x}_t)$ we are given a sequence of convex functions $\{f_t(\cdot)\}_{t=1}^T$ with gradients $\nabla f_t(\cdot)$. We will further assume that our iterates x_t are constraint to be in a convex set C. First we need to determine how the projection step looks like. It does not make sense to project with respect to Euclidean geometry any longer after all we came up with a different identification of primal and dual spaces exactly because we wanted to consider a different geometry on the primal space. We already know that the Bregman divergence plays the role of distance function in this new geometry and it is strongly convex with respect to the first argument so a good candidate for the projection of a point x_0 becomes $\widehat{\mathrm{x}} = \arg\min_{\mathrm{x}\in C} D_\psi(\mathrm{x},\mathrm{x}_0)$. This translates to the mirror descent update in the following way:

$$\begin{aligned}\mathrm{x}_{t+1/2} &= \nabla\psi^*(\nabla\psi(\mathrm{x}_t) - \eta\nabla f_t(\mathrm{x}_t))\\ \mathrm{x}_{t+1} &= \arg\min_{\mathrm{x}\in C} D_\psi(\mathrm{x},\mathrm{x}_{t+1/2}).\end{aligned}$$

Let us have a quick sanity check. From our discussion about how mirror descent can be interpreted as a proximal method the above update needs to be equivalent to $\mathrm{x}_{t+1} = \arg\min_{\mathrm{x}\in C}\{\eta g_t(\mathrm{x}) + D_\psi(\mathrm{x},\mathrm{x}_t)\}$, where $g_t(x) = \langle\nabla f_t(\mathrm{x}_t), \mathrm{x} - \mathrm{x}_t\rangle$. Using first order optimality for constraint convex optimization we have that $\mathrm{x}_{t+1} = \arg\min_{\mathrm{x}\in C} D_\psi(\mathrm{x},\mathrm{x}_{t+1/2})$ is equivalent to:

$$\langle\nabla\psi(\mathrm{x}_{t+1}) - \nabla\psi(\mathrm{x}_{t+1/2}), \mathrm{y} - \mathrm{x}_{t+1}\rangle = \langle\nabla\psi(\mathrm{x}_{t+1}) - \nabla\psi(\mathrm{x}_t) + \eta\nabla f_t(\mathrm{x}_t), \mathrm{y} - \mathrm{x}_{t+1}\rangle \geq 0, \forall \mathrm{y}\in C.$$

On the other hand $\langle \nabla\psi(\mathrm{x}_{t+1}) - \nabla\psi(\mathrm{x}_t) + \eta\nabla f_t(\mathrm{x}_t), \mathrm{y} - \mathrm{x}_{t+1}\rangle \geq 0, \forall \mathrm{y} \in C$ is exactly the optimality condition for $\min_{\mathrm{x}\in C}\{\eta g_t(\mathrm{x}) + D_\psi(\mathrm{x}, \mathrm{x}_t)\}$. We now present the stochastic projected mirror descent update:

$$\begin{aligned} \mathrm{x}_{t+1/2} &= \nabla\psi^*(\nabla\psi(\mathrm{x}_t) - \eta\nabla f_t(\mathrm{x}_t)) \\ \mathrm{x}_{t+1} &= \arg\min_{\mathrm{x}\in C} D_\psi(\mathrm{x}, \mathrm{x}_{t+1/2}). \end{aligned} \tag{A.8}$$

The above algorithm comes equipped with the following regret guarantee.

Theorem A.3.8. *After T iterations of A.8 with step size η we have:*

$$\sum_{t=1}^{T} f_t(\mathrm{x}_T) - f_t(\mathrm{x}^*) \leq \frac{D_\psi(\mathrm{x}_1, \mathrm{x}^*)^2}{2\eta} + \frac{\eta T\sigma^2}{\alpha},$$

where α is the strong convexity parameter of ψ and $\|\nabla f_t(\mathrm{x})\|_*^2 \leq \sigma^2, \forall t \in [T], \mathrm{x} \in C$, and x^* is any point in C.

Proof. Let us track the progress in the primal space using the Lyapunov function $D_\psi(\cdot, \mathrm{x}^*)$. Using the equivalence of the update to the proximal step together with lemma A.3.5 we have

$$\begin{aligned} \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_t\rangle + D_\psi(\mathrm{x}^*, \mathrm{x}_t) \geq{}& \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}_{t+1} - \mathrm{x}_t\rangle \\ &+ D_\psi(\mathrm{x}_{t+1}, \mathrm{x}_t) + D_\psi(\mathrm{x}^*, \mathrm{x}_{t+1}). \end{aligned}$$

Shuffling terms around this gives us:

$$\begin{aligned} D_\psi(\mathrm{x}^*, \mathrm{x}_{t+1}) - D_\psi(\mathrm{x}^*, \mathrm{x}_t) &\leq \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_{t+1}\rangle - D_\psi(\mathrm{x}_{t+1}, \mathrm{x}_t) \\ &= \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_t\rangle + \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}_t - \mathrm{x}_{t+1}\rangle - D_\psi(\mathrm{x}_{t+1}, \mathrm{x}_t) \\ &\leq \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_t\rangle + \eta\|\nabla f_t(\mathrm{x}_t)\|_*\|\mathrm{x}_t - \mathrm{x}_{t+1}\| - D_\psi(\mathrm{x}_{t+1}, \mathrm{x}_t) \\ &\leq \eta\langle\nabla f_t(\mathrm{x}_t), \mathrm{x}^* - \mathrm{x}_t\rangle + \frac{\eta^2}{2\alpha}\|\nabla f_t(\mathrm{x}_t)\|_*^2. \end{aligned}$$

Telescoping and using convexity we arrive at the conclusion of the theorem. □

Appendix B

Tools for lower bounds in bandit games

In this chapter we discuss two types of lower bounds for bandit problems. For both types of lower bounds we will fix a class of problem instances Θ. An instance $\theta \in \Theta$ describes precisely how the losses or rewards are generated over the T rounds of the bandit game. For example, in the stochastic MAB game θ can just be thought of as a distribution over the rewards of the arms at each round.

The first type of lower bound we consider is a min-max lower bound. Every regret bound depends both on the problem instance θ and the algorithm or policy which the player follows. Let us denote the policy of the player by π and suppose that this policy belongs to some fixed class of policies Π. This class, for example, could consist of all deterministic algorithms mapping observed rewards (or losses) to a new action. The goal of the min-max regret lower bound is to establish an inequality of the type:

$$\min_{\pi \in \Pi} \max_{\theta \in \Theta} \mathbb{E}[R_{\theta,\pi}(T)] \geq f(T),$$

where $f(T)$ is some function of the time horizon. The interpretation of such bounds is that there always exists a bandit problem θ such that no matter what strategy the player chooses to follow, she would necessarily incur regret at least $C(T)$ (in expectation).

The second type of lower bound is the so called instance dependent lower bounds.

These bounds are the counter-parts of the instance dependent regret upper bounds which we discussed for the stochastic MAB problem in Section 1.2.2.2. The instance dependent bounds read in the following way – for a fixed bandit instance θ and any policy π the expected regret $\mathbb{E}[R_{\theta,\pi}(T)]$ is bounded in the following way:

$$\lim_{T\to\infty} \frac{\mathbb{E}[R_{\theta,\pi}(T)]}{f(T)} \geq C(\theta),$$

where again $f(T)$ is some function of the horizon (most often $f(T) = \log(T)$), and $C(\theta)$ is some instance dependent constant. We note that while instance-dependent lower bounds might seem stronger than the min-max lower bounds, they are usually asymptotic in nature, that is they only hold in the limit as T goes to infinity, and further it is usually harder to derive such bounds.

B.1 Min-max lower bounds

The basic principle behind showing min-max lower bounds is to exhibit two instances of the bandit problem θ and θ' which are very close, in information theoretic terms, however, have different best arms. On one hand if θ and θ' are similar, information theory tells us that we would need many observations from θ and θ' to be able to tell the two instances apart. Let us consider the following simple example. Suppose that θ is the distribution $\mathcal{N}(0,1)$ and θ' is the distribution $\mathcal{N}(\epsilon,1)$. A sequence $(X_t)_{t=1}^T$ is sampled from either θ or θ'. How large does T need to be so that we can determine if the sequence was sampled from θ or θ'? More formally we are trying come up with a policy π which maps from the T samples to a binary space $\{0,1\}$, where 0 denotes that the samples came from θ and 1 denotes that the samples came from θ'. And so we are interested in π such that w.p. at least $1-\delta$ determines correctly the distribution, that is $\mathbb{P}(\pi(X_1,\ldots,X_T) = 0|\theta) \geq 1-\delta$ and $\mathbb{P}(\pi(X_1,\ldots,X_T) = 1|\theta') \geq 1-\delta$. For short, let us denote $\mathbb{P}_0(\cdot) = \mathbb{P}(\cdot|\theta)$ and $\mathbb{P}_1(\cdot) = \mathbb{P}(\cdot|\theta')$, and the event $\pi(X_1,\ldots,X_T) = 0$ as A. Answering the above question also implies that $\mathbb{P}_0(A) - \mathbb{P}_1(A) \geq 1-2\delta$. Intuitively, however, the difference $\mathbb{P}_1(A) - \mathbb{P}_2(A)$ should depend on ϵ and so a small ϵ would require more samples to distinguish between θ and θ'. The min-max lower bounds for multi-armed bandits will follow the same overall idea, that is to show that $\mathbb{P}_1$ and $\mathbb{P}_2$ are close for any event A.

To bound the difference between $\mathbb{P}_1(A) - \mathbb{P}_2(A)$ we are going to need some tools from information theory. We first give the definition of Kullback-Leibler (KL) divergence.

Definition B.1.1 (KL-divergence)**.** The KL-divergence for two (absolutely continuous) distributions $\mathbb{P}_1$ and p_2 on a sample space Ω is defined as

$$KL(\mathbb{P}_1, \mathbb{P}_2) = \mathbb{E}_{X \sim \mathbb{P}_1}\left[\log\left(\frac{\mathbb{P}_1(X)}{\mathbb{P}_2(X)}\right)\right] = \sum_{X \in \Omega} \mathbb{P}_1(X) \log\left(\frac{\mathbb{P}_1(X)}{\mathbb{P}_2(X)}\right).$$

Further, we define the KL-divergence between two Bernoulli random variables with parameters p_1 and p_2 respectively as

$$kl(p_1, p_2) = p_1 \log\left(\frac{p_1}{q_1}\right) + (1 - p_1) \log\left(\frac{1 - p_1}{1 - p_2}\right).$$

The KL-divergence enjoys two key properties which we will need. First, it decomposes over product distributions which follows from the chain rule, and second, it is an upper bound on the difference $|\mathbb{P}_1 - \mathbb{P}_2|$, which is known as Pinsker's inequality.

Theorem B.1.1 (Chain rule for relative entropy)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on Ω. Then*

$$KL(\mathbb{P}(X, Y), \mathbb{Q}(X, Y)) = KL(\mathbb{P}(X), \mathbb{Q}(X)) + KL(\mathbb{P}(Y|X), \mathbb{Q}(Y|X)).$$

Theorem B.1.2 (Pinsker's inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on Ω. Then*

$$\sup_{A \in \Omega} |\mathbb{P}(A) - \mathbb{Q}(A)| \leq \sqrt{\frac{1}{2} KL(\mathbb{P}, \mathbb{Q})}.$$

Using the above two theorems we can finally derive a bound on the difference between the distributions induced by θ and θ' and determine the number of samples to distinguish the two. First we use Pinsker's inequality to bound $(\mathbb{P}_1(A) - \mathbb{P}_2(A))^2 \leq \frac{1}{2}KL(\mathbb{P}_1, \mathbb{P}_2)$. Next we use the chain rule to write

$$KL(\mathbb{P}_1, \mathbb{P}_2) = \sum_{t=1}^{T} KL(\mathcal{N}(0,1), \mathcal{N}(\epsilon, 1)) = \frac{T\epsilon^2}{2}.$$

The above implies that there exists no policy which can distinguish θ from θ' with probability $1-\delta$ unless $T \geq \frac{2(1-\delta)^2}{\epsilon^2}$.

Standard MAB lower bound. We now show how to extend the above ideas to the bandit game. For the K-armed bandit game we define K environments $(\theta_i)_{i=1}^{K}$ s.t. θ_i is $\mathcal{N}(\mu_i, \mathrm{I})$ where $\mu_i \in \mathbb{R}^K$ is the vector with coordinate $j \neq i$ equal to $1/2$ and j-th coordinate equal to $1/2 - \epsilon$. Let $\mathbb{P}_i$ be the associated probability measure induced by the player's strategy over the T rounds of the game when interacting with losses generated by θ_i. Further, let $\mathbb{P}_0$ be the probability measure induced by the player's policy when interacting with a bandit game in which all losses are sampled from $\mathcal{N}(1/2, 1)$. Suppose the adversary selects which environment the player is going to face in the beginning of the game, uniformly at random from all θ_i's. If $N_i(T)$ denotes the random variable which is the number of times the player selected the i-th arm then we can write the expected regret of the player as

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{\mathbb{P}_i}[\epsilon(T - N_i(T))] = \epsilon T - \frac{\epsilon}{K}\sum_{i=1}^{K} \mathbb{E}_{\mathbb{P}_i}[N_i(T)].$$

Let us now bound $|\mathbb{E}_{\mathbb{P}_i}[N_i(T)] - \mathbb{E}_{\mathbb{P}_0}[N_i(T)]|$ using the same techniques as we did when trying to distinguish $\mathcal{N}(0,1)$ from $\mathcal{N}(\epsilon, 1)$. First we write

$$\mathbb{E}_{\mathbb{P}_i}[N_i(T)] = \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}_i}[i_t = i] = \sum_{t=1}^{T} \mathbb{P}_i(i_t = i),$$

where i_t denotes the random variable which denotes the arm selected by the player. We can decompose $\mathbb{E}_{\mathbb{P}_0}[N_i(T)]$ in the same way. Thus using Pinsker's inequality and AM-QM inequality the regret can be lower bounded as

$$\mathbb{E}[R(T)] \geq T\epsilon\,(1 - 1/K) - T\epsilon\sqrt{\sum_{i=1}^{K} \frac{KL(\mathbb{P}_0, \mathbb{P}_i)}{2K}} \geq \frac{T\epsilon}{2}\left(1 - \epsilon\sqrt{\frac{T}{K}}\right).$$

The above inequality reveals the trade-off in selecting ϵ by the adversary. On one hand the smaller ϵ is the harder it is to distinguish $\mathbb{P}_i$ from $\mathbb{P}_0$. On the other hand the regret from failing to distinguish the best arm also decreases with ϵ. Setting $\epsilon = \sqrt{\frac{K}{4T}}$ shows that the expected regret is lower bounded as $\mathbb{E}[R(T)] \geq \Omega(\sqrt{TK})$.

For a similar exposition and more details on min-max lower bounds we refer the reader to Slivkins (2019).

B.2 Instance dependent lower bounds

While min-max bounds provide a clear picture of what is possible to achieve in terms of regret minimization for the bandit problem they do not tell the full story. After all the player might not be faced with an adversary which has selected the worst possible problem instance. In fact, if ϵ, that is the gap between the best arm and other arms is large, the approach we present above is going to fail. Further, there exist algorithms for the stochastic MAB problem which achieve instance-dependent bounds which evaluate to $O(K \log(T))$ whenever ϵ is large. It is natural to ask if such algorithms are optimal. Next we show how to attempt such a lower bound.

Similarly to the min-max lower bounds we begin with a lower bound on the KL-divergence between the induced measures $\mathbb{P}$ and $\mathbb{P}'$ by the policy of the player acting on bandit instances θ and θ'. We well specify how to choose θ' given an instance θ in a bit. The key inequality states that

$$KL(\mathbb{P}, \mathbb{P}') \geq kl(\mathbb{E}_{\mathbb{P}}[X], \mathbb{E}_{\mathbb{P}'}[X]), \tag{B.1}$$

where X is any measurable random variable with respect to $\mathbb{P}$ and $\mathbb{P}'$ and kl is the KL-divergence between two Bernoulli random variables with the respective means. Using the chain rule we can also show that

$$KL(\mathbb{P}, \mathbb{P}') = \sum_{k=1}^{K} \mathbb{E}_{\theta_i}[N_k(T)] KL(\theta(k), \theta'(k)).$$

where we use $\theta(k)$ to denote the distribution of arm k under environment θ. Fix an arm k which is sub-optimal under θ. Let $X = N_k(T)/T$ and choose θ' so that it only differs from θ on arm k and arm k is optimal for θ' We have the following

$$\mathbb{E}_\theta[N_k(T)]KL(\theta(k), \theta'(k)) \geq \left(1 - \frac{\mathbb{E}_\theta[N_k(T)]}{T}\right) \log\left(\frac{T}{T - \mathbb{E}_{\theta'}[N_k(T)]}\right) - \log(2).$$

We claim that this is sufficient for our desired lower bound, at least for player strategies which enjoy instance-dependent guarantees. In particular, let π be such a strategy, that is for any $\alpha > 0$ the strategy has regret bounded by $o(T^\alpha)$ for large enough T on any instance problem instance. Then because k is sub-optimal for θ we have $\frac{\mathbb{E}_\theta[N_k(T)]}{T} \to 0$ and further $\log\left(\frac{T}{T - \mathbb{E}_{\theta'}[N_k(T)]}\right) \to \log(T)$. Let θ_k be the environment which minimizes the KL-divergence over all θ' which satisfy the above construction. Then what we have shown above is that the regret of the player is lower bounded as

$$\mathbb{E}[R(T)] \geq \sum_{k=1}^{K} \frac{\Delta_k \log(T)}{KL(\theta, \theta_k)},$$

for large enough T. The work of Garivier et al. (2018) shows such instance dependent bounds formally and further discusses lower bounds for small T as well.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In ***NIPS***, volume 11, pages 2312–2320, 2011.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In ***ICML***, pages 3–11. Citeseer, 1999.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. ***arXiv preprint arXiv:1612.06246***, 2016.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In ***Conference on learning theory***, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In ***Artificial intelligence and statistics***, pages 99–107. PMLR, 2013.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. ***J. ACM***, 64(5):30:1–30:24, 2017. doi: 10.1145/3088510. URL https://doi.org/10.1145/3088510.

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In ***Advances in Neural Information Processing Systems***, pages 974–982, 2016.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In ***Advances in Neural Information Processing Systems***, pages 1610–1618, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In ***Annual Conference on Learning Theory***, volume 40. Microtome Publishing, 2015.

Oren Anava, Elad Hazan, and Shie Mannor. Online convex optimization against adversaries with memory and application to statistical arbitrage. ***arXiv preprint arXiv:1302.6937***, 2013.

Raman Arora and Teodor Vanislavov Marinov. Efficient convex relaxations for streaming pca. In ***Advances in Neural Information Processing Systems***, pages 10496–10505, 2019.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In ***Proceedings of the 29th International Conference on Machine Learning***, pages 1747–1754, 2012a.

Raman Arora, Poorya Mianjy, and Teodor Marinov. Stochastic optimization for multiview representation learning using partial least squares. In ***International Conference on Machine Learning***, pages 1786–1794, 2016.

Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, and Nati Srebro. Stochastic approximation for canonical correlation analysis. In ***Advances in Neural Information Processing Systems***, pages 4775–4784, 2017.

Raman Arora, Michael Dinitz, Teodor Vanislavov Marinov, and Mehryar Mohri. Policy regret in repeated games. In ***Advances in Neural Information Processing Systems***, pages 6732–6741, 2018.

Raman Arora, Teodor V Marinov, and Mehryar Mohri. Bandits with feedback graphs and switching costs. ***arXiv preprint arXiv:1907.12189***, 2019.

Raman Arora, Teodor V Marinov, and Enayat Ullah. Private stochastic convex optimization: Efficient algorithms for non-smooth objectives. ***arXiv preprint arXiv:2002.09609***, 2020.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In ***International Conference on Artificial Intelligence and Statistics***, pages 2116–2124. PMLR, 2021.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. ***Theory of Computing***, 2012b.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In ***COLT***, pages 217–226, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. ***Theoretical Computer Science***, 410(19):1876–1902, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. ***Journal of Machine Learning Research***, 3(Nov):397–422, 2002.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In ***Advances in Neural Information Processing Systems***, pages 49–56, 2007.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. ***Machine learning***, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. ***SIAM Journal on Computing***, 2002b.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In ***Advances in Neural Information Processing Systems***, 2009.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. ***Econometrica: Journal of the Econometric Society***, pages 1–18, 1987.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In ***International Conference on Machine Learning***, pages 263–272, 2017.

Peter L Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In ***Conference on Learning Theory***, 2008.

Kush Bhatia and Karthik Sridharan. Online learning with dynamics: A minimax perspective. ***arXiv preprint arXiv:2012.01705***, 2020.

Avrim Blum and Yishay Mansour. From external to internal regret. ***Journal of Machine Learning Research***, 8:1307–1324, 2007.

Béla Bollobás and Ernest J Cockayne. Graph-theoretic parameters concerning domination, independence, and irredundance. ***Journal of Graph Theory***, 3(3):241–249, 1979.

Haim Brezis. ***Functional analysis, Sobolev spaces and partial differential equations***. Springer Science & Business Media, 2010.

Sébastien Bubeck. ***Bandits games and clustering foundations***. PhD thesis, INRIA Nord Europe (Lille, France), 2010.

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

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In ***Algorithmic Learning Theory***, pages 111–127. PMLR, 2018.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In ***Conference On Learning Theory***, pages 508–528. PMLR, 2019.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. In ***The 2014 ACM International Conference on Measurement and Modeling of Computer Systems***, SIGMETRICS '14, pages 289–300. ACM, 2014a.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Stochastic bandits with side observations on networks. In ***The 2014 ACM international conference on Measurement and modeling of computer systems***, pages 289–300, 2014b.

Stephane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In ***UAI***, 2012.

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

Xi Chen and Xiaotie Deng. Settling the complexity of two-player nash equilibrium. In ***2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)***, pages 261–272. IEEE, 2006.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In ***Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics***, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Vasek Chvatal. A greedy heuristic for the set-covering problem. ***Mathematics of operations research***, 4(3):233–235, 1979.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In ***International Conference on Machine Learning***, pages 811–819, 2016.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In ***Advances in Neural Information Processing Systems***, pages 1763–1771, 2017.

Ashok Cutkosky, Abhimanyu Das, and Manish Purohit. Upper confidence bounds for combining stochastic bandits. ***arXiv preprint arXiv:2012.13115***, 2020.

Christoph Dann. ***Strategic Exploration in Reinforcement Learning - New Algorithms and Learning Guarantees***. PhD thesis, Carnegie Mellon University, 2019.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform pac bounds for episodic reinforcement learning. In ***Advances in Neural Information Processing Systems***, pages 5713–5723, 2017.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC reinforcement learning with rich observations. ***arXiv preprint arXiv:1803.00606***, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. ***International Conference on Machine Learning***, 2019.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. ***SIAM Journal on Computing***, 39(1):195–259, 2009.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In ***Proceedings of the forty-sixth annual ACM symposium on Theory of computing***, pages 459–467. ACM, 2014.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. ***arXiv preprint arXiv:2002.07125***, 2020.

Lars Engebretsen and Jonas Holmerin. Clique is hard to approximate within n 1-o (1). In ***International Colloquium on Automata, Languages, and Programming***, pages 2–12. Springer, 2000.

Paul Erdöos, Ralph Faudree, and Edward T Ordman. Clique partitions and clique coverings. ***Discrete Mathematics***, 72(1-3):93–101, 1988.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In ***Computational Learning Theory***, 2002.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In ***Proceedings of the seventh annual conference on Computational learning theory***, pages 88–97. ACM, 1994.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In ***2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)***, pages 115–122. IEEE, 2010.

Dean P Foster and Rakesh V Vohra. A randomization rule for selecting forecasts. ***Operations Research***, 41(4):704–709, 1993.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. ***Games and Economic Behavior***, 21(1-2):40, 1997.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. ***arXiv preprint arXiv:1906.00531***, 2019.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. ***Advances in Neural Information Processing Systems***, 33, 2020a.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. ***arXiv preprint arXiv:2010.03104***, 2020b.

David A Freedman. On tail probabilities for martingales. ***the Annals of Probability***, pages 100–118, 1975.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. ***Journal of computer and system sciences***, 55(1):119–139, 1997.

Drew Fudenberg and David K Levine. Conditional universal consistency. ***Games and Economic Behavior***, 29(1-2):104–130, 1999.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In ***Proceedings of the 24th annual conference on learning theory***, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. ***Mathematics of Operations Research***, 44(2):377–399, 2018.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. ***Mathematics of Operations Research***, 44(2):377–399, 2019.

Sascha Geulen, Berthold Vöcking, and Melanie Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In ***COLT***, pages 132–143, 2010.

Wayne Goddard and Michael A. Henning. Independent domination in graphs: A survey and recent results. ***Discrete Mathematics***, 313(7):839–854, 2013.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. ***SIAM journal on control and optimization***, 35(3):715–743, 1997.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. ***Econometrica***, 68(5):1127–1150, 2000.

Johan Hastad. Clique is hard to approximate within $n^{1-\epsilon}$. ***Acta Mathematica***, 182 (1):105–142, 1999.

Elad Hazan and Satyen Kale. Computational equivalence of fixed points and no regret algorithms, and convergence to equilibria. In ***Advances in Neural Information Processing Systems***, pages 625–632, 2008.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. ***Journal of Machine Learning Research***, 12(4), 2011.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. ***arXiv preprint arXiv:2011.11566***, 2020.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In ***Conference On Learning Theory***, pages 3395–3398, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In ***International Conference on Machine Learning***, pages 1704–1713, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? ***arXiv preprint arXiv:1807.03765***, 2018.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. ***arXiv preprint arXiv:2006.05606***, 2020.

Sham Kakade. ***On the sample complexity of reinforcement learning***. PhD thesis, University College London, 2003.

Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. ***Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf***, 2(1), 2009.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. ***Journal of Computer and System Sciences***, 71(3):291–307, 2005.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In ***International conference on algorithmic learning theory***, pages 199–213. Springer, 2012.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. ***Canadian Journal of Mathematics***, 12:363–366, 1960.

Tomáv*s* Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In ***Advances in Neural Information Processing Systems***, pages 613–621, 2014.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In ***Advances in Neural Information Processing Systems***, pages 1840–1848, 2016.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. ***Advances in applied mathematics***, 6(1):4–22, 1985.

Tor Lattimore. The pareto regret frontier for bandits. ***arXiv preprint arXiv:1511.00048***, 2015.

Tor Lattimore and Csaba Czepesvari. ***Bandit Algorithms***. Cambridge University Press, 2018.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In ***International Conference on Machine Learning***, pages 5662–5670. PMLR, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more:

high-probability data-dependent regret bounds for adversarial bandits and mdps. ***arXiv preprint arXiv:2006.08040***, 2020a.

Chung-Wei Lee, Haipeng Luo, and Mengxiao Zhang. A closer look at small-loss bounds for bandits with graph feedback. In ***Conference on Learning Theory***, pages 2516–2564. PMLR, 2020b.

Ehud Lehrer. A wide range no-regret theorem. ***Games and Economic Behavior***, 42(1):101–115, 2003.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In ***Proceedings of the 19th international conference on World wide web***, pages 661–670, 2010.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. ***Machine learning***, 2(4):285–318, 1988.

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

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In ***32nd AAAI Conference on Artificial Intelligence***, 2018.

Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. In ***Conference on Learning Theory***, pages 979–986. PMLR, 2018.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption robust exploration in episodic reinforcement learning. ***arXiv preprint arXiv:1911.08689***, 2019.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In ***Advances in Neural Information Processing Systems***, pages 684–692, 2011.

Teodor Vanislavov Marinov, Poorya Mianjy, and Raman Arora. Streaming principal component analysis in noisy setting. In ***International Conference on Machine Learning***, pages 3413–3422, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. ***arXiv preprint arXiv:0907.3740***, 2009.

Neri Merhav, Erik Ordentlich, Gadiel Seroussi, and Marcelo J Weinberger. On sequential strategies for loss functions with memory. ***IEEE Transactions on Information Theory***, 48(7):1947–1958, 2002.

Mehryar Mohri and Scott Yang. Conditional swap regret and conditional correlated equilibrium. In ***Advances in Neural Information Processing Systems***, pages 1314–1322, 2014.

Mehryar Mohri and Scott Yang. Online learning with transductive regret. In ***Advances in Neural Information Processing Systems***, pages 5220–5230, 2017.

John Nash. Non-cooperative games. ***Annals of mathematics***, pages 286–295, 1951.

John F Nash et al. Equilibrium points in n-person games. ***Proceedings of the national academy of sciences***, 36(1):48–49, 1950.

Arkadij Semenovic Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Arkadii Semenovich Nemirovsky and David Borisovich Yudin. ***Problem complexity and method efficiency in optimization***. John Wiley &amp; Sons, Inc., Panstwowe Wydawnictwo Naukowe (PWN), 1983.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. ***Algorithmic game theory***. Cambridge university press, 2007.

Maillard Odalric and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In ***Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics***, pages 570–578. JMLR Workshop and Conference Proceedings, 2011.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In ***Advances in Neural Information Processing Systems***, pages 8874–8882, 2018.

Francesco Orabona and Dávid Pál. Scale-free online learning. ***Theoretical Computer Science***, 716:50–69, 2018.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In ***Advances in Neural Information Processing Systems***, pages 3003–3011, 2013.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. ***arXiv preprint arXiv:2012.13045***, 2020a.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. ***arXiv preprint arXiv:2003.01704***, 2020b.

Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. ***Journal of the ACM (JACM)***, 55(3):1–29, 2008.

Martin Puterman. ***Markov Decision Processes: Discrete Stochastic Dynamic Programming***. Wiley-Interscience, 1994.

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

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. ***J. Mach. Learn. Res.***, 16(1):155–186, 2015.

Anshuka Rangi and Massimo Franceschetti. Online learning with feedback graphs and switching costs. In ***The 22nd International Conference on Artificial Intelligence and Statistics***, pages 2435–2444, 2019.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. ***Journal of the ACM (JACM)***, 62(5):32, 2015.

Ankan Saha, Prateek Jain, and Ambuj Tewari. The interplay between stability and regret in online learning. ***arXiv preprint arXiv:1211.6158***, 2012.

Antoine Salomon and Jean-Yves Audibert. Deviations of stochastic bandit regret. In ***International Conference on Algorithmic Learning Theory***, pages 159–173. Springer, 2011.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In ***Conference on Learning Theory***, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In ***International Conference on Machine Learning***, pages 1287–1295. PMLR, 2014.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. ***arXiv preprint arXiv:1905.03814***, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. ***arXiv preprint arXiv:1904.07272***, 2019.

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

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In ***Advances in Neural Information Processing Systems***, pages 1505–1512, 2008.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. ***Biometrika***, 25(3/4):285–294, 1933.

Aristide Tossou, Christos Dimitrakakis, and Devdatt Dubhashi. Thompson sampling for stochastic bandits with graph feedback. In ***31st AAAI Conference on Artificial Intelligence***, 2017.

Enayat Ullah, Poorya Mianjy, Teodor Vanislavov Marinov, and Raman Arora. Streaming kernel pca with $\tilde{o}(\sqrt{n})$ random features. In ***Advances in Neural Information Processing Systems***, pages 7311–7321, 2018.

Michal Valko. ***Bandits on graphs and structures***. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2016.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In ***Conference On Learning Theory***, pages 1263–1291. PMLR, 2018.

Yifan Wu, András György, and Csaba Szepesvari. Online learning with gaussian payoffs and side observations. In ***Advances in Neural Information Processing Systems 28***, pages 1360–1368. Curran Associates, Inc., 2015a.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with Gaussian payoffs and side observations. In ***NIPS***, pages 1360–1368, 2015b.

Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. ***arXiv preprint arXiv:2102.04692***, 2021.

Kunhe Yang, Lin F Yang, and Simon S Du. Q-learning with logarithmic regret. ***arXiv preprint arXiv:2006.09118***, 2020.

A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. ***https://arxiv.org/abs/1901.00210***, 2019.

Yinglun Zhu and Robert Nowak. Pareto optimal model selection in linear bandits. ***arXiv preprint arXiv:2102.06593***, 2021.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In ***Advances in neural information processing systems***, pages 1583–1591, 2013.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In ***The 22nd International Conference on Artificial Intelligence and Statistics***, pages 467–475. PMLR, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. ***Journal of Machine Learning Research***, 22(28):1–49, 2021.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In ***Proceedings of the 20th international conference on machine learning (icml-03)***, pages 928–936, 2003.