Politecnico di Torino

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

Doctoral Dissertation
National Doctoral Program in Artificial Intelligence (37^{th} cycle)

Leveraging Structured Feedback in Online Learning

By

Khaled Eldowa
******

Supervisor(s):
Prof. Nicolò Cesa-Bianchi, Supervisor
Prof. Marcello Restelli, Co-Supervisor

Politecnico di Torino
2025

Declaration

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

Khaled Eldowa
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Acknowledgements

My gratitude and thanks go to my family for their love and support, my supervisors Nicolò and Marcello for their guidance, and all members of LAILA over the past three years for their valued company. I would also like to thank the two reviewers of this thesis, Satyen Kale and Ronald Ortner, for their comments and feedback.

Abstract

The field of online learning provides a general framework for modeling sequential decision-making problems. In many of its applications, the learner only receives partial feedback upon making a decision. One extreme of this phenomenon, modeled by the multi-armed bandit problem, is when the learner only gets feedback concerning the chosen decision, which provides no information about the resulting outcome had they chosen a different decision. However, it is often the case that realistic problems are endowed with more nuanced structures, emerging from the properties of the decision space and the form of the loss function, that allow leakage of information between the elements of the decision space. In this dissertation, four such structures are studied; namely, bandits with mediator feedback, online learning with graph feedback, bandits with expert advice, and online convex Markov decision processes. New algorithms are presented with proven regret guarantees, complemented in most cases with impossibility results. The aim of these findings is to characterize to what extent the structure of the problem (and the resulting information leakage) can be leveraged by the learner towards optimizing their performance. In particular, for bandits with mediator feedback, we introduce a new information-theoretic measure for the complexity of the decision set and achieve new regret bounds featuring this quantity. Additionally, we prove nearly matching lower bounds for certain classes of instances. For online learning with graph feedback and bandits with expert advice, we provide improved upper and lower regret bounds that contribute towards a refined characterization of the minimax regret rate in both settings. Finally, we study online episodic Markov decision processes with convex objectives in the adversarial setting, and present new algorithms enjoying sub-linear regret guarantees in this problem.

Publications

This dissertation is based on the following publications and preprints.

- Chapter 2 is based on the following journal publication:
  Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 25(353):1–36, 2024
  which subsumes an earlier workshop publication that appeared as:
  Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information-theoretic regret bounds for bandits with fixed expert advice. In *2023 IEEE Information Theory Workshop (ITW)*, pages 30–35, 2023a
- Chapter 3 is based on the following conference publication:
  Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 46122–46133. Curran Associates, Inc., 2023b
- Chapter 4 is based on the following preprint (to appear in the Journal of Artificial Intelligence Research):
  Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *arXiv preprint arXiv:2406.16802*, 2024
- Chapter 5 is partly based on the following preprint (to appear in the Proceedings of the 42nd International Conference on Machine Learning):
  Bianca Marin Moreno, Khaled Eldowa, Pierre Gaillard, Margaux Brégère, and Nadia Oudjane. Online episodic convex reinforcement learning. *arXiv preprint arXiv:2505.07303*, 2025
  Part of this work was done while the author was visiting the research team THOTH at the Inria centre at the University Grenoble Alpes in the spring of 2024.

Additionally, the following conference publication is part of the author's Ph.D. research but is out of the scope of the dissertation:

Khaled Eldowa and Andrea Paudice. General tail bounds for non-smooth stochastic mirror descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3205–3213. PMLR, 2024

Contents

1 Introduction **1**

1.1 Facing Experts and Bandits . . . 3

1.2 Richer Structures: An Overview of The Models Studied in This Work . . . 7

1.2.1 Bandits with Mediator Feedback . . . 7

1.2.2 Bandits with Graph Feedback . . . 8

1.2.3 Bandits with Expert Advice . . . 8

1.2.4 Convex Markov Decision Processes with Bandit Feedback . . . 10

1.3 Online Convex Optimization as a Common Solution Strategy . . . 11

2 Information Capacity Regret Bounds for Bandits with Mediator Feedback **14**

2.1 Introduction . . . 14

2.1.1 Prior Results . . . 16

2.1.2 Contributions . . . 18

2.1.3 Additional Related Works . . . 19

2.2 Preliminaries . . . 20

2.2.1 Information Theory Background . . . 21

2.2.2 Problem Setting . . . 24

2.3 The Policy Set Capacity: An Improved Regret Bound for EXP4 . . . 25

2.3.1 The Policy Set Capacity . . . 26

2.3.2 A Regret Bound for EXP4 in Terms of the Capacity . . . 27

2.3.3 Examples . . . 30

2.3.4 A Generalization for Time-Varying Policy Distributions 32
2.4 Best-of-Both-Worlds Bounds 35
2.5 Lower Bounds 40
2.5.1 The Two Policies Case 41
2.5.2 ε-Greedy Policies 43
2.5.3 The Multitask Bandits Structure 44
2.6 An Impossibility Result Under Linear Bandit Feedback 45
2.7 The Full-Information Case 48
2.8 Conclusions and Future Directions 49

3 On the Minimax Regret for Online Learning with Feedback Graphs 51
3.1 Introduction 51
3.1.1 Additional related work 53
3.2 Problem Setting 54
3.3 FTRL with Tsallis Entropy Regularization for Feedback Graphs 55
3.4 Adapting to Arbitrary Sequences of Graphs 60
3.5 An Improved Lower Bound via Multitask Learning 62
3.6 Conclusions and Future Work 65
3.7 Comparison with Chen et al. (2024) 65

4 Improved Regret Bounds for Bandits with Expert Advice 67
4.1 Introduction 67
4.2 Preliminaries 69
4.3 q-FTRL for Bandits with Expert Advice 70
4.4 An Improved Instance-Based Regret Bound 72
4.5 A Lower Bound for Restricted Expert Advice via Feedback Graphs 81
4.6 Conclusion 84

5 Online Convex Markov Decision Processes with Bandit Feedback 85
5.1 Introduction 85

5.1.1 Problem Setting . . . 87
5.1.2 Related Works . . . 87
5.1.3 Contributions and Road map . . . 89
5.2 On the Representation of Occupancy Measures . . . 90
5.3 Entropic Regularization and Ellipsoidal Gradient Estimation . . . 97
5.4 Self-Concordant Regularization and (Dikin) Ellipsoidal Gradient Estimation 105
5.5 Future Directions . . . 114

References **115**

Appendix A Supplementary Material for Chapter 2 **126**
A.1 Proof of Theorem 2.7 . . . 126
A.2 Proof of Theorem 2.9 . . . 128

Appendix B Supplementary Material for Chapter 3 **130**
B.1 Auxiliary Results . . . 130
B.2 Proofs of Section 3.3 . . . 133
B.3 Proofs of Section 3.4 . . . 136
B.4 Proof of the Lower Bound . . . 138

Appendix C Supplementary Material for Chapter 4 **147**
C.1 Auxiliary Results . . . 147

Chapter 1

Introduction

Online learning is an area of machine learning that addresses sequential-decision making problems. The basic scenario it studies is that of a learner repeatedly interacting with an environment for a series of rounds by performing an action at each round and accordingly suffering a loss—or in more optimistic formulations, gaining a reward—that encodes the desirability of the resulting outcome. A primary premise in this family of problems is that the effect of an executed action, and in particular, the loss it brings about in a given round, is generated via a mechanism that is, at least partially, unknown to the learner. Naturally, the learner aims to minimize the losses accumulated throughout the interaction period. The extent to which they succeed at this is usually measured by comparing their performance against that of a benchmark strategy that relies on a deeper knowledge of the environment's workings. In particular, the learner would like that their *regret*—the difference between the sum of the losses they suffer and the sum of those incurred by said benchmark strategy—grows sub-linearly in the number of rounds. This would imply that the learner's per-round performance asymptotically matches the benchmark. This ability to adapt to the behavior of the environment, so as to track the benchmark strategy with limited prior knowledge, defines a broad notion of *learning* in this domain.

Owing to its generality, this framework can accommodate a diverse range of practical applications. Imagine, for instance, a trader, deciding daily on how to allocate their capital over a number of assets. The trader's gain on a certain day can be formulated as a function of the factor by which their wealth increases, which depends on the (apriori unknown) ratio of closing to opening price for each asset. In the absence of any assumptions on how the market behaves, this is known as the universal portfolio selection problem (Cover, 1991), where the target benchmark is usually set as the best (in hindsight) static allocation of assets over the time horizon of interest.

Or imagine a web retailer, offering a product to a sequence of customers. This retailer, unaware of the maximum price each customer would be willing to pay for this product (i.e., their valuations), must decide on a (non-negotiable) price to propose to each customer as they arrive. This is an instance of the dynamic pricing problem (Kleinberg and Leighton, 2003; Misra et al., 2019). Assuming the customers' valuations are independently drawn from an unknown distribution belonging to a given family, the retailer might aim to match their performance, in terms of the expected accumulated revenue in a given period of time, with that of the strategy that chooses in every round the price that generates the maximum expected revenue. If an extra hint concerning a customer's valuation is available beforehand, a prudent (or greedy) retailer could strive to exploit this and set as their benchmark the best policy (within a certain set) that maps this contextual information to a price decision.

While sharing a common character, these two examples readily illustrate some of the possible variety in the formulation of an online learning problem. One key component is the structure of the decision set; in the dynamic pricing problem, a decision is a price within an admissible interval, while in portfolio selection, a decision can be seen as a probability distribution over the assets. A second component concerns the form of the loss (or reward) function and how it changes across rounds, whether in a stochastic or an arbitrary manner. Another crucial component, in part related to the previous one, is the feedback received by the learner after executing the action. It is reasonable to assume that the trader in the first example gets to observe the return of all the assets at the end of the day; hence, they could infer the performance of any allocation in hindsight. Whereas in dynamic pricing, if the product was sold to a customer, the retailer cannot infer what would have been the outcome had they posted a higher price, since they never observe the valuation of the customer. Nor can they, in general, infer whether a failed sale could have gone through at a specific lower price. Lastly, the target benchmark and the availability of contextual information are likewise important aspects.

The literature on online learning is mostly themed around studying how a variation in one or more of these components affects the inherent difficulty of the problem. This work follows suit; we study a few online learning models, each having its own peculiar characteristics in terms of the structure of the decision set, the form of the loss function, and the type of received feedback—henceforth referred to collectively as the feedback structure of the problem. Before introducing these models, we briefly present two simple and fundamental online learning problems that will serve as reference models to be extended or modified in the sequel. Moreover, through these two examples, we introduce

some of the basic methodology adopted in this work and provide a simple demonstration of how the structure of the problem can affect its difficulty.

1.1 Facing Experts and Bandits

In the problem of *prediction with expert advice*, one of the earliest online learning problems studied in the literature (Littlestone and Warmuth, 1989, 1994; Cesa-Bianchi et al., 1993, 1997), the learner's task in every round is to predict the next element in a sequence, say of numbers between 0 and 1. Upon doing so, they incur a loss that is a function of the prediction and the true outcome, say the absolute value of their difference. Before making a decision, the learner receives some recommendations as to what to predict from a number of experts, possibly endowed with extra knowledge. Without placing any assumptions on how the sequence evolves, the benchmark is set as the best performing expert in hindsight. A more general formulation due to Freund and Schapire (1997), nevertheless referred to in the literature by the same name, abstracts away the form of the loss function and the nature of the elements in the sequence. Instead, the problem is reduced to choosing which expert to follow in a given round, given only the knowledge of the losses incurred by each expert in the previous rounds.

Formally, let K and T be two positive integers denoting the number of experts and the time horizon. For round $t \in [T]$, let $\ell_t \colon [K] \to [0,1]$ be the loss function such that $\ell_t(i)$ denotes the loss of the i-th expert at round t.[1] Let $I_t \in [K]$ denote the learner's (possibly randomized) pick of the expert to follow at round t, upon choosing which, the loss function ℓ_t is revealed. We assume the sequence of loss functions $(\ell_t)_{t\in[T]}$ is fixed before the start of the game, but otherwise arbitrary. The (expected) regret of the learner is defined as

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t(i)\,,$$

where the expectation is taken over the randomization in the decisions of the learner. The dependence of R_T on the sequence of loss functions is suppressed from the notation for simplicity, but as we recall, a successful learner should, at the very least, guarantee that R_T is sub-linear in T for any sequence of loss functions. That is, we are interested here in the *worst-case regret* of the learner, $\max_{(\ell_t)_{t\in[T]}} R_T$. Accordingly, an apt indicator of the inherent difficulty of the problem is the best achievable worst-case regret for any strategy, known in game-theoretic terms as the *minimax regret*. A strategy in this context can be

[1] For a positive integer n, $[n]$ denotes the set $\{1,\dots,n\}$.

seen as a sequence of functions mapping the interaction history so far $(I_1, \ell_1, \ldots, I_{t-1}, \ell_{t-1})$ to a probability distribution over the experts, from which I_t is to be sampled.

Now, how do we achieve sub-linear regret? It turns out it is very easy to fail at this; any deterministic strategy cannot guarantee sub-linear regret in the worst case. Fix any such strategy and partition the experts into two non-empty sets S_1 and S_2. With $S^t \in \{S_1, S_2\}$ denoting the subset to which I_t belongs, consider a sequence of loss functions where $\ell_t(i) = 1$ for $i \in S^t$ and $\ell_t(i) = 0$ otherwise. Then, $\sum_{t=1}^T \ell_t(I_t) = T$ and $\min_{i\in[K]} \sum_{t=1}^T \ell_t(i) \leq T/2$, implying that R_T is linear in T. Going further, an interesting implication of this reasoning is that setting

$$I_t \in \operatorname*{arg\,min}_{i\in[K]} \sum_{s=1}^{t-1} \ell_s(i)$$

is not a sound strategy in general; it fails (i.e., suffers linear regret) on the loss sequence where in odd rounds, $\ell_t(i) = 0$ for $i \in S_1$ and $\ell_t(i) = 1$ for $i \in S_2$, with the opposite assignment in even rounds. It fails on this loss sequence even if ties are broken randomly. This update rule, known as *follow the leader* (FTL), is perhaps the most immediate approach in the way of attempting to track the benchmark strategy. But as illustrated above, it leaves one vulnerable to an "adversarial" loss sequence, abruptly changing across rounds.

To evade this, a natural idea is to randomize the choice of expert at each round, since it is allowed to do so, with the probability of picking an expert being correlated with their relative performance so far. One such strategy is to sample $I_t \sim p_t$, where

$$p_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)}{\sum_{j\in[K]} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(j)\right)}$$

for every $i \in [K]$, with $\eta > 0$ being a tunable parameter. As it turns out, this exponential weights update, known as the *Hedge* algorithm (Freund and Schapire, 1997), is equivalent (see, e.g., Shalev-Shwartz, 2012, Section 2.6.1) to the rule that sets[2]

$$p_t = \operatorname*{arg\,min}_{p\in\Delta_K} \left\langle p, \eta \sum_{s=1}^{t-1} \ell_s \right\rangle + \phi(p), \tag{1.1}$$

where $\Delta_K := \{x \in \mathbb{R}^K : \sum_{i=1}^K x = 1 \text{ and } x_i \geq 0 \; \forall i \in [K]\}$ is the probability simplex in R^K, and for $p \in \mathbb{R}^K$, $\phi(p) := \sum_{i\in[K]} p(i) \log p(i)$ is the negative Shannon entropy of p. This strategy is an instance of the *follow the regularized leader* (FTRL) family (Abernethy et al., 2008), with ϕ as the regularizer. With a suitable choice of η, the worst-case expected regret of this strategy is guaranteed to be at most of order $\sqrt{T \log K}$ (Freund and Schapire, 1997). This

[2] We sometimes treat a function $f\colon [K] \to \mathbb{R}$ as a vector in R^K with $f(i)$ being its i-th coordinate.

bound features a sub-linear dependence on T and a mild sub-logarithmic dependence on K, and is, in fact, unimprovable; this is the minimax regret for this problem, up to constants (Cesa-Bianchi et al., 1997).

As to why this approach works, a simple manipulation, irrespective of the definition of ϕ, shows that (Shalev-Shwartz, 2012, Lemma 2.3)

$$\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \ell_t \rangle \leq \frac{\phi(\mathbf{e}_{i^*}) - \phi(p_1)}{\eta} + \sum_{t=1}^{T} \langle p_t - p_{t+1}, \ell_t \rangle$$

where $\mathbf{e}_{i^*} \in \mathbb{R}^K$ is the indicator vector for $i^* \in \arg\min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i)$, and the expectation of the left-hand side is equal to that of the regret. The first term, known as the *bias* term, describes how far the value of ϕ at the comparator $\mathbf{e}_{i^*}$ is from that at our initial guess p_1 (the minimizer of ϕ). The second term, the *stability* term, measures the difference between each two consecutive decisions in the direction of the loss function at the earlier round. The particular choice of the negative entropy as a regularizer[3] guarantees that $\langle p_t - p_{t+1}, \ell_t \rangle \leq \eta \|\ell_t\|_\infty^2 \leq \eta$ (Shalev-Shwartz, 2012, Lemma 2.10). Thus, picking a smaller η ensures that the decisions do not drift much across rounds, leading to a smaller stability term; yet, at the same time, too small an η would magnify the effect of the bias term on the regret. Given that $\phi(\mathbf{e}_{i^*}) - \phi(p_1) = \log K$, the best choice of η at this junction is $\sqrt{\log(K)/T}$, which yields a $2\sqrt{T \log K}$ regret bound holding for any sequence of losses.

The appeal of the FTRL rule is that it makes explicit the learner's attempt to balance between the optimistic urge to follow the best expert so far, resolved that the trend will continue, and the pessimistic urge of hedging one's bets as much as possible since, after all, the loss functions of past rounds need not have any relationship with those of future rounds. Indeed, If the learner is made aware that the loss of each expert is sampled independently at each round from a certain distribution (of unknown mean), then just following the leader can in certain cases be the optimal behavior (Kotłowski, 2018). Here, instead, adopting the FTRL update with a regularizer suited to the geometry of the decision set (effectively, the probability simplex) and an optimally tuned (in a worst case sense) learning rate parameter yielded a minimax-optimal strategy.

The *multi-armed bandit* problem, in its non-stochastic formulation (Auer et al., 1995, 2002a), is identical to the experts problem except that the learner does not observe the loss function ℓ_t at the end of a round, only $\ell_t(I_t)$, the loss suffered by the chosen expert. Also, terminology-wise, one usually uses the term "arms", or actions, instead of experts.

[3] Thanks, in particular, to it being 1-strongly convex in the $\|\cdot\|_1$ norm over the set Δ_k (Shalev-Shwartz, 2012, Example 2.5).

The more limited feedback provided in the bandit problem adds another incentive for hedging one's choices: that of *exploration*, which becomes necessary to gain sufficient information about the quality of the actions. The stochastic multi-armed bandit problem, where the loss of each arm follows a fixed distribution (with unknown mean), is perhaps the most emblematic online learning problem, with a rich literature that goes as far back as the works of Thompson (1933) and Robbins (1952). It presents in a simple and explicit manner the exploration-exploitation dilemma, this tension between attempting to minimize short-term losses by choosing the best-performing action so far and seeking the potential long-term benefit of exploring seemingly sub-optimal but poorly-estimated actions.

As we are mostly concerned in this work with the worst-case (in terms of loss sequences) regret, we focus here on the non-stochastic variant. Using the FTRL algorithm presented in the experts problem requires knowledge of the entire loss function at each round. This can be dealt with using loss estimators, a standard approach for extending a strategy from a full-information setting to a partial information one. With $p_t(i)$ denoting, similarly to before, the probability of picking action i at round t given the interaction history so far $(I_1, \ell_1(I_1), \dots, I_{t-1}, \ell_{t-1}(I_{t-1}))$, the inverse-probability-weighted loss estimate for action i at round t is given by $\widehat{\ell}_t(i) = \mathbb{I}\{i = I_t\}\ell_t(i)/p_t(i)$, which is an unbiased estimate. Using the FTRL rule in (1.1) with these loss estimates recovers in its exponential weights form the EXP3 algorithm of Auer et al. (2002a), which can be shown with a more careful analysis and a suitably chosen learning rate to satisfy a regret bound of order $\sqrt{KT\log K}$. The cost of having partial access to the loss functions is paid for in the analysis through the dependence of the stability term on the variance of the loss estimates, which leads to a worse dependence on the number of actions.

A slightly better bound, of order $\sqrt{KT}$, is achievable via a similar algorithm that uses the negative 1/2-Tsallis entropy $\psi_{1/2}(p) \coloneqq 2\left(1 - \sum_{i\in[K]}\sqrt{p(i)}\right)$ as the regularizer in place of the negative Shannon entropy (Audibert and Bubeck, 2009; Abernethy et al., 2015). This bound is minimax-optimal up to constants as it is matches a lower bound derived by Auer et al. (2002a), which asserts that this worsening in the dependence of the regret on the number of actions from $\sqrt{\log K}$ to $\sqrt{K}$ is unavoidable when moving from experts to bandits.

1.2 Richer Structures: An Overview of The Models Studied in This Work

While able to model scenarios where the learner receives partial feedback, the structure of the bandit problem can in many cases be too rigid in that observing the loss of one action provides no information about that of another action. This is not the case in the dynamic pricing problem since playing one action reveals the losses assigned to a portion of the action space. While in the portfolio selection problem, even if one does not observe the returns of all the assets at the end of the trading day, the observed outcome of adopting a certain allocation of assets carries some information about the quality of similar allocations. We see in both examples how the structure of the problem can give rise to more nuanced types of feedback beyond simple full information or bandit feedback. Understanding to what extent such subtle forms of *information leakage* can be exploited by the learner is thus of both theoretical and practical interest. In the following, we briefly describe the abstract models that we study in this work, each inducing a distinctive feedback structure, and outline our contributions to the understanding of their theoretical properties.

1.2.1 Bandits with Mediator Feedback

In this model, the elements of the decision set, referred to as *policies*, are probability distributions over a set of *outcomes*. In every round, a loss function assigns a loss to each outcome, while the loss of a policy is accordingly taken as the expected loss of an outcome sampled from its distribution. When the learner picks a policy in a given round, an outcome is sampled from its distribution and revealed to the learner along with its assigned loss. No assumption is placed on how the loss function evolves over time, and the benchmark is set as the best performing policy in hindsight over a given horizon of rounds.

The extreme case when the distribution of each policy is entirely concentrated on a unique outcome is nothing but a standard bandit problem with the policies as arms. But if the supports of the distributions do overlap, then observing the loss assigned to the outcome sampled in one round provides some information about the losses of all policies placing positive probability over this outcome. Hence, a high degree of similarity between the policies' distributions facilitates this kind of information leakage, resulting, as one would hope, in an easier problem. We attempt to answer in Chapter 2 the question of how

much the structure of the policy set affects the best achievable regret. We introduce a new complexity measure for the policy set, called the *policy set capacity*, and prove new upper and lower bounds on the regret featuring this quantity in different settings. Our results improve upon previous works (e.g., McMahan and Streeter, 2009; Papini et al., 2019) where similar models were studied.

1.2.2 Bandits with Graph Feedback

This model (Mannor and Shamir, 2011; Alon et al., 2017) interpolates between full information and bandit feedback in a direct manner. In the simplest formulation of this model, a problem instance is characterized by an undirected graph whose vertices are the actions. The learner, who is aware of this graph, observes in each round the loss of the selected action and those of all its neighbors in the graph. Clearly, an empty graph recovers the bandit problem, while a fully-connected graph (a clique) recovers the experts problem. For intermediate cases, a bound of order $\sqrt{\alpha T \log K}$, achieved by Mannor and Shamir (2011), is the best known regret guarantee, where α is the independence number of the graph, a measure of its dis-connectivity. They also proved a similar lower bound of order $\sqrt{\alpha T}$, which asserts that α is a key quantity in characterizing the difficulty of this problem. This too remained a state-of-the-art result until now.

While the minimax regret is known in both the full information and bandit cases up to constants ($\sqrt{T \log K}$ and $\sqrt{KT}$ respectively), an abrupt factor logarithmic in K separates the aforementioned best known upper and lower bounds as soon as we move away from these two cases. We make considerable progress towards closing this gap in Chapter 3, improving this characterization from both ends by proving more refined bounds that interpolate more smoothly between the minimax regret at the two extremes. Precisely, we achieve an improved regret bound of order $\sqrt{\alpha T(1+\log(K/\alpha))}$, and complement it with a $\sqrt{\alpha T \log_\alpha K}$ lower bound holding for a certain family of instances. For a related family, the latter bound has been improved in the subsequent work of Chen et al. (2024) to match the former.

1.2.3 Bandits with Expert Advice

This model modifies the standard bandit problem by setting a different benchmark; instead of the best action, the learner competes with the best strategy within a given set. This increases the modeling power of the bandit problem significantly, allowing

the adoption of meaningful notions of optimality in complicated scenarios such as non-stationary environments (Luo et al., 2018) or environments where the set of available actions varies over time (Kleinberg et al., 2010). The prediction of each strategy in a given round, taking the form of a probability distribution over the actions, is revealed to the learner at the beginning of the round as the "advice" of an "expert". Auer et al. (2002a) introduced this framework, proposed the EXP4 algorithm, and showed that its regret is of order at most $\sqrt{KT\log N}$, where N is the number of experts. This logarithmic dependence on N affords one to use this algorithm to compete with a very large number of strategies while incurring a mild cost in terms of regret.

This problem is at least as hard as a full-information game played on the experts set, where $\sqrt{T\log N}$ regret is optimal, but not harder than a bandit game played on the same set, where $\sqrt{NT}$ regret is optimal. We firstly ask whether the $\sqrt{\log N}$ factor in the bound of Auer et al. (2002a) is always necessary. A partial answer is provided by a $\sqrt{KT\log_K N}$ lower bound proved by Seldin and Lugosi (2016) when $N \geq K$. Analogously to our result in the graph feedback setting, a special case of a general result by Kale (2014) shows that a bound of order $\sqrt{\min\{K,N\}T(1+\log(N/\min\{K,N\}))}$ is achievable. In Chapter 4, for a harder setting where the latter bound is still achievable, we prove a matching lower bound of order $\sqrt{KT\log(N/K)}$ for a family of instances where $N > K$. We obtain this result as a consequence of a reduction from bandits with graph feedback combined with the recent results of Chen et al. (2024).

We then ask whether the dependence on K in the bound of Auer et al. (2002a) (or $\min\{K,N\}$ in the bound of Kale, 2014) is improvable in favorable instances where the recommendations of the experts are similar. One should notice that the bandits with expert advice problem exhibits a form of mediator feedback; the recommendations of the experts can be seen as the policies, and the arms of the bandit problem can be seen as the outcomes. With this in mind, an extension of the mediator feedback problem that allows for the policies' distributions to vary between rounds is considered in chapter 2. There, we improve upon the bound of Auer et al. (2002a) by achieving another of order $\sqrt{\sum_{t=1}^{T}\mathcal{C}_t\log N}$, where $\mathcal{C}_t$ is the capacity of the policies' distributions (or the experts' recommendations) at round t, which measures the dis-similarity between the policies, and is never larger than $\min\{K,N\}$. This bound, however, always features the $\sqrt{\log N}$ factor, causing it in some cases to fall short of Kale's (2014) bound. We design in Chapter 4 a strategy achieving a regret guarantee that combines the merits of the two aforementioned bounds, constituting thus the best loss-sequence-independent bound to date for this classic problem.

1.2.4 Convex Markov Decision Processes with Bandit Feedback

In a slight departure from the type of environments described so far, the last model we study concerns Markov decision processes (MDPs; Puterman, 2014), but will nevertheless be addressed through an online learning lens. An episodic MDP can be characterized by the number of steps within an episode, a set of states, a set of actions, a per-step transition kernel that determines the probability of transitioning to a certain state having executed a certain action, and a per-step loss (or reward) function over state-action pairs. In one formulation of the problem, an agent interacts with this environment for an episode by selecting an action at every step; incurring a loss that depends on the current step, the current state and the executed action; and transitioning into a new state that is drawn from the transition kernel of the current step.

The agent's choice of action at each step is typically dictated by some Markovian policy that it adopts, meaning that the probability of selecting a certain action depends only on the current state (and possibly also on the current step within the episode). The standard goal is to minimize the expected sum of losses incurred by the agent over the course of an episode. When the loss function and the transition kernels are unknown apriori, the problem transitions into the realm of reinforcement learning (RL; Sutton and Barto, 2018), which has flourished in recent years with successful applications in fields such as finance (Hambly et al., 2023) and robotics (Zhao et al., 2020).

Differently from the models we discussed so far, the state at which an agent finds itself (and which affects future losses) depends on the actions it had executed in the past. However, as typical in the regret minimization literature in RL (e.g., Azar et al., 2017; Jin et al., 2020), the primary concern in the online learning problem we consider here is not the agent's regret within an episode; instead, we consider each episode of interaction as one round in a game between a learner and an environment. In this game, the learner chooses in each round a policy to be executed in an MDP whose loss function (and in some formulations, also its transition kernels) is determined on a per-round basis by the environment. Hence, the learner's decision space is a set of policies, typically the class of Markovian policies, and the benchmark defining the regret is the best policy within this set, determined in hindsight for a certain horizon of rounds.

Notice that a policy, together with the MDP, induces in each step of the episode a distribution over the states and actions. This sequence of per-step state-action distributions induced by a policy is referred to as its occupancy measure. Notice then that the expected sum of losses within an episode is a linear function in the occupancy measure of the adopted policy; it can be seen as the sum of an inner product at every step between the

state-action distribution induced by the policy and a vector representing the loss assigned to every state-action pair at that step.

A more general formulation sets the objective in each episode as a convex function of the occupancy measure of the policy. This formulation has been recently studied in a number of different settings (see, e.g., Hazan et al., 2019; Geist et al., 2022; Mutti et al., 2023; Moreno et al., 2024a), its appeal being that it subsumes important problems like imitation learning and pure exploration via maximization of the entropy of the state distribution, which cannot be expressed via the classic linear objective. In Chapter 5, we contribute to this line of work by studying for the first time a bandit feedback model for the online convex-objective MDP problem. We assume the MDP's transition dynamics are known and fixed throughout the rounds, but allow the convex objective function at each round to be chosen adversarially. We address the problem by adapting techniques from the bandit convex optimization literature (see, e.g., Lattimore, 2024) to this setting.

1.3 Online Convex Optimization as a Common Solution Strategy

Online convex optimization (OCO), as coined by Zinkevich (2003), is a simple and general online learning framework. We refer to the monographs by Shalev-Shwartz (2012), Hazan (2021), and Orabona (2023) for comprehensive treatments of this subject. Given a convex domain $S \subseteq \mathbb{R}^d$, the learner in this framework decides at every round on a prediction $w_t \in S$, observes a convex function $f_t \colon S \to \mathbb{R}$, and suffers $f_t(w_t)$ as a loss. The default objective is to minimize the following familiar notion of regret:

$$\sum_{t=1}^{T} f_t(w_t) - \min_{w \in S} \sum_{t=1}^{T} f_t(w) .$$

While this framework readily subsumes problems such as online portfolio selection, it does not immediately model the problems we described in the last section. The primary reason being that in none of the problems we described does the learner observe the entire loss function. In some cases too, the decision set is not convex. Nonetheless, OCO algorithms and analysis techniques are the standard tools used in the literature to address online learning problems with adversarial losses, even under partial information.

We have already seen this in our brief discussion on the (non-stochastic) bandit problem in Section 1.1. We have seen how the popular Exp3 algorithm of Auer et al.

Algorithmic Scheme 1.1 Reducing a generic bandit problem to an OCO problem

input: online problem with decision set V and observation set $\mathcal{O}$
initialization: choose a surrogate convex set $\widehat{V} \in \mathbb{R}^d$ and an OCO algorithm $\mathcal{A}$ to be executed on $\widehat{V}$ as a decision set
for $t = 1, \ldots, T$ **do**
  receive a prediction $\widehat{v}_t$ from $\mathcal{A}$
  map $\widehat{v}_t$ to a distribution p_t over the elements of V
  draw $v_t \in V$ from P_t and execute it in the environment
  receive an observation $o_t \in \mathcal{O}$
  map v_t, $\widehat{v}_t$, and o_t to a (possibly random) $g_t \in \mathbb{R}^d$
  send the loss function $f_t(\cdot) := \langle \cdot, g_t \rangle$ to $\mathcal{A}$
end for

(2002a) can be seen to employ an instance of the FTRL algorithm (an OCO method) at its core, with the same being true for the algorithm achieving the minimax regret bound. There, FTRL was executed on the simplex Δ_k as the domain, in place of the non-convex action set $[K]$, and on the estimated losses $\widehat{\ell}_t$, substituting for the inaccessible loss vectors ℓ_t. Although seemingly disparate, all the problems we described in the previous section will be addressed in a similar manner, via morphing them into OCO problems.

We provide in the following a brief informal description of the main ingredients of this reduction, we refer to Section 6.3 in (Hazan, 2021) for a more thorough exposition. Suppose that V denotes the (possibly non-convex) decision set in our online problem with partial information, $\ell_t \colon V \to \mathbb{R}$ denotes the loss function at round t, $\mathcal{L}$ denotes the set of all such possible loss functions, and that $\mathcal{O}$ denotes the set of possible observations provided in any given round upon choosing an element of V. For example, in the bandit problem, $V = [K]$, $\mathcal{L}$ is set of all functions mapping $[K]$ to $[0,1]$, and the elements of the observation set take the form $(i, \ell(i))$ where $i \in [K]$ and $\ell \in \mathcal{L}$. We summarize the reduction to OCO in Algorithmic Scheme 1.1. Taking the EXP3 algorithm as an example, we choose $\widehat{V}$ as the simplex Δ_K and $\mathcal{A}$ as FTRL with the negative entropy regularizer. At round t, the prediction $\widehat{v}_t \in \Delta_K$ of FTRL readily corresponds to a distribution over $[K]$, from which action v_t is sampled. Subsequently, the loss $\ell_t(v_t)$ is observed, and we choose $g_t(i) = \mathbb{I}\{i = v_t\}\ell_t(i)/\widehat{v}_t(i)$.

Let $\phi \colon X \to \mathbb{R}$ be a convex function to be used a regularizer, where $X \supseteq V$. In this work, we will either use FTRL as the OCO routine, yielding that

$$\widehat{v}_t \in \operatorname*{arg\,min}_{w \in \widehat{V}} \left\langle w, \eta \sum_{s=1}^{t-1} g_s \right\rangle + \phi(w)\,,$$

or the related online mirror descent algorithm (OMD; Orabona, 2023, Chapter 6), which takes the form

$$\widehat{v}_t \in \operatorname*{arg\,min}_{w \in \widehat{V}} \left\langle w, \eta g_{t-1} \right\rangle + D_\phi(w, \widehat{v}_{t-1}),$$

where $\eta > 0$ is the learning rate, and for $a, b \in \widehat{V}$ such that ϕ is differentiable at b, $D_\phi(a,b) := \phi(a) - \phi(b) - \langle a - b, \nabla\phi(b)\rangle$ is the Bregman divergence based on ϕ between a and b. Via Lemmas 6.16 and 7.14 in (Orabona, 2023), both algorithms satisfy for any $w \in \widehat{V}$ that

$$\sum_{t=1}^{T} \langle \widehat{v}_t - w, g_t \rangle \le \frac{\phi(w) - \phi(\widehat{v}_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} g_t^\intercal (\nabla^2 \phi(z_t))^{-1} g_t,$$

for some z_t on the line segment between $\widehat{v}_t$ and $\widetilde{v}_{t+1} := \operatorname{arg\,min}_{w \in X} \left\langle w, \eta g_t \right\rangle + D_\phi(w, \widehat{v}_t)$, assuming $\widehat{v}_t$ and $\widetilde{v}_{t+1}$ belong to the interior of X, where ϕ is required to be twice differentiable with a positive definite Hessian. In the case of Exp3, the expectation of the left hand-side of the above inequality (with w taken as the indicator vector in $\mathbb{R}^K$ corresponding to the best action) coincides with the regret in the bandit problem, which through this result can be easily shown to be bounded by $(\log K)/\eta + \eta K T/2$, leading to the aforementioned $\sqrt{2KT\log K}$ regret bound with the best tuning of η.

We see then that with $v^* \in \operatorname{arg\,min}_{v \in V} \sum_{t=1}^{T} \ell_t(v)$, our task is reduced to choosing $\widehat{V}$, the regularizer ϕ, the sampling scheme of v_t, and the loss (or gradient) estimate g_t such that we can guarantee reasonable bounds on the quantities

$$\frac{\phi(w^*) - \phi(\widehat{v}_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[g_t^\intercal (\nabla^2 \phi(z_t))^{-1} g_t \right] \quad \text{and} \quad \mathbb{E}\left[\langle \widehat{v}_t - w^*, g_t \rangle - (\ell_t(v_t) - \ell_t(v^*)) \right] \ \forall t$$

for some $w^* \in \widehat{V}$. Through this approach, the structure of the problem can be favorable in so much as it facilitates the application of a reduction with convenient properties in terms of the behavior of the quantities above. We shall see in the coming chapters how a prudent instantiation of this general scheme, coupled with a careful analysis, will lead us in many cases to achieving near-optimal regret bounds.

Chapter 2

Information Capacity Regret Bounds for Bandits with Mediator Feedback*

2.1 Introduction

The multi-armed bandit problem, as mentioned in the introductory chapter, fails to model situations in which, aside from observing the loss of the played action, the learner could—in the same round—obtain information concerning the losses of other actions. This information leakage, as we referred to it, could be a result of prior knowledge of an underlying structure for the losses, or owing to more explicit side observations. Regardless of form, such extra information can lead to more efficient learning in the face of large (or even infinite) action sets. A prominent example of structured losses is exhibited by the (adversarial) linear bandit problem, in which the action set is a subset of $\mathbb{R}^d$ and the loss assigned to an action in a given round is the inner product between the action and a common latent loss vector associated with that round. For this setting, Bubeck et al. (2012) provide an algorithm achieving nearly optimal regret bounds of order $\sqrt{dT\log K}$ for finite action sets and $d\sqrt{T\log T}$ for compact actions sets, improving drastically in many cases upon what is achievable in the standard bandit problem. On the other hand, a simple form of side observations is modeled by the problem of bandits with graph feedback, which was briefly introduced in Section 1.2.2. There, we saw that algorithms exploiting the provided side observations can enjoy improved regret guarantees (compared to the standard bandit case) depending on the structure of the graph.

*This chapter is based on (Eldowa, Cesa-Bianchi, Metelli, and Restelli, 2024).

The subject of this chapter is a bandit model where the information leakage results from a combination of side observations and a structured assignment of losses. In the basic template of this model, the learner is faced with a policy set as their decision space, whose elements are each associated with a probability distribution over a common space of outcomes. At each round, a (latent) loss map associates each outcome with a numerical loss, and we define the loss incurred by a policy as the expected loss of an outcome sampled from its distribution. Upon choosing a policy, an outcome is drawn from its distribution, and the learner subsequently observes *both* the outcome and its associated loss. Depending on the problem, the loss map could be fixed over the rounds or changing in a stochastic or adversarial manner. The regret in this framework is defined as the difference between the (expected) cumulative loss of the learner and that of the optimal policy in hindsight. Our main goal is to understand how the structure of the policy set, in particular, how the similarity between the policies affects the achievable regret.

One aspect of this problem reminiscent of linear bandits is that having chosen a policy, the learner's expected loss is linear in the policy's distribution, seen as a vector in the simplex. The distinction, however, is that the learner does not observe this quantity, as would be the case under linear bandit feedback; instead, one outcome, sampled via the chosen policy, and its assigned loss are observed. Later on, we shall expand further on the difference between the two feedback models and how it impacts the difficulty of regret minimization.

This framework has been studied in the works of Papini et al. (2019) and Metelli et al. (2021), where it was referred to as the *mediator feedback* model. The name here highlights the role of the outcomes as an extra layer of feedback "mediating" between the chosen policy and the observed loss, thereby allowing additional information gain regarding other policies. This feedback model arises in these two works from a certain formulation of the online policy optimization problem in episodic reinforcement learning. An instance of their setting is characterized by a Markov decision process and a set of policies that map states to distributions over actions. At every round of interaction, the learner selects a policy through which they interact with the MDP for a fixed horizon. Naturally, the learner observes both the sampled trajectory and the accumulated reward, and aims to compete with the best policy from the given policy set. In this case, the trajectories are the outcomes over which each policy induces a probability distribution.

The feedback structure of the mediator framework is shared with the more classical problem of bandits with expert advice (Auer et al., 2002a), which we briefly described in Section 1.2.3. To recall, this problem is a variation of the standard bandit problem, where

at the beginning of every round, the learner receives "advice" from each of a number of "experts" in the form of a probability distribution over the actions. The benchmark is then taken as the best performing expert in hindsight. Here, the actions of the bandit instance play the role of the outcomes, which are sampled from the distributions provided by the experts, playing the role of the policies. Exactly fitting this problem into the mediator feedback framework requires restricting the learner to only access the actions by way of sampling from (a mixture of) the experts' distributions, though this requirement is already satisfied by most state-of-the-art approaches. A more important distinction of the expert advice problem is the incorporation of a contextual element in that the distributions recommended by the experts can vary from round to round. Given our stated goal of studying the extent to which the similarity between the available distributions can be exploited, the addition of this contextual element is somewhat orthogonal to the main focus of this chapter, though it will still be briefly treated.

2.1.1 Prior Results

The EXP4 algorithm was proposed in the work of Auer et al. (2002a) to address the bandits with expert advice problem, and remains to date an important benchmark in the contextual bandit literature. It was shown to enjoy a regret bound of order $\sqrt{KT\log N}$, where N denotes the number of experts and K the number of actions. This was later complemented in the work of Seldin and Lugosi (2016) with a lower bound of order $\sqrt{KT\log N/\log K}$, which, ignoring the discrepancy in the log factors, essentially asserts the optimality of the bound of Auer et al. (2002a). However, this lower bound concerns an instance where the number of experts is exponential in the number of actions, and the experts' distributions are deterministic. The former is not surprising; if the number of experts is small compared to the arms, one can always achieve a regret of order $\sqrt{NT}$ by playing a minimax optimal bandit algorithm directly over the experts, entirely casting aside the structure of the problem. More importantly, the bound of Auer et al. (1995) does not reflect one's expectation that the problem should become easier if the recommended distributions are more similar, and the lower bound does not address this question.

Via an elaborate modification of EXP4, McMahan and Streeter (2009) did address these very two points. For a fixed set of expert recommendations, their algorithm achieves a bound of order $\sqrt{\mathcal{S}T\log N}$, where $\mathcal{S}$, formally defined in Section 2.3, is a notion of effective size of the set of recommended distributions (i.e., the policy set). It satisfies $\mathcal{S} \leq \min\{K, N\}$, but can be smaller depending on the similarity between the distributions—or the "agreement between the experts". Specifically, it reaches its smallest value, $\mathcal{S} = 1$, when

all the distributions are identical. One issue with this bound is that when the number of experts is small (say, only two), the fact that $\mathcal{S} \geq 1$ means that no substantial improvement is achieved over the $\sqrt{NT}$ bound, no matter how similar the distributions are. Indeed, $\mathcal{S} - 1$ is arguably a more apt metric as it can shrink arbitrarily if the distributions are similar enough. In particular, it reduces to the total variation distance when there are only two distributions. Nevertheless, even if the bound were to scale with this quantity, one may ask whether this is the best achievable dependence on the structure of the policy set. We note in passing that the bound of McMahan and Streeter (2009) can also be achieved by plain EXP4 in the general case (see Lattimore and Szepesvári, 2020, Theorem 18.3), where it takes the form $\sqrt{\sum_t \mathcal{S}_t \log N}$ with $\mathcal{S}_t$ measuring the (dis)agreement between the experts at the t-th round.

While these results concern the adversarial regime, where no statistical constraints are placed on the losses, similar mediator feedback problems have been studied in the stochastic regime, where the losses are drawn at every round from a fixed distribution. This includes the aforementioned works of Papini et al. (2019) and Metelli et al. (2021), and that of Sen et al. (2018), who consider a stochastic variant of the expert advice problem. Unlike the worst-case flavour (in terms of the dependence of the loss map) of the bounds in the adversarial regime, the results in these works are generally instance-based; the bounds enjoy a logarithmic dependence on the time horizon, but degrade in harder instances where suboptimal policies are difficult to discern. Still, the dependence of these bounds on the policy set structure is largely independent of the loss map as it is primarily represented via diameter-like quantities measuring the pairwise maximum "distance" between the distributions according to some dissimilarity measure, mainly the chi-squared divergence or related quantities. Comparing $\mathcal{S} - 1$ with this chi-squared "diameter", neither quantity uniformly dominates the other, though the former is always bounded, while the latter need not be.

In the online learning and bandits literature, best-of-both-worlds (BOBW) algorithms (Bubeck and Slivkins, 2012) address the adversarial and stochastic regimes simultaneously without prior knowledge of the nature of the environment. They guarantee regret (poly) logarithmic in the time horizon when the faced environment is stochastic, while retaining sub-linear regret against general environments. Of particular relevance to our setting is the recent work of Dann et al. (2023), where they obtain the first BOBW guarantees for bandits with expert advice (or contextual bandits) using EXP4 as a black-box decision rule within a cascade of two meta-algorithms. For stochastic environments, the bound is of order $K \log T \log N / \Delta$, where Δ denotes the minimum sub-optimality gap for the experts, whereas the traditional bound of $\sqrt{KT \log N}$ is guaranteed for all environments.

As apparent, these bounds feature the usual coarse dependence on the number of actions and are therefore unable to reflect the affinity of the recommended distributions.

2.1.2 Contributions

We study in this chapter a generic mediator feedback setting with finite policy and outcome sets. We propose a new complexity (or effective size) measure for the policy set, which we refer to as the chi-squared capacity of the policy set, or simply the policy set capacity. This quantity (defined in Section 2.3) can be interpreted as the information capacity of a certain hypothetical communication channel induced by the policy set, albeit we define the mutual information between the input and output of said channel in terms of the chi-squared divergence in lieu of the standard Kullback–Leibler divergence. The channel referred to here is one whose input alphabet is the policy set and output alphabet is the outcome set, while its transition matrix is defined such that conditioned on an input policy, the output is drawn from the policy's distribution. Besides its more natural information-theoretic interpretation, this notion of capacity is never larger than either $\mathcal{S}-1$ or the maximum pairwise chi-squared divergence. Moreover, when the policy set consists of only two policies, the capacity reduces to a (symmetric) divergence measure of the same order as the squared Hellinger distance and the triangular discrimination.

In Section 2.3, we consider the adversarial regime and provide an improved regret bound of order $\max\{\sqrt{\mathcal{C}T\log N}, \log N\}$ for EXP4, where $\mathcal{C}$ denotes the capacity. Unlike prior results, the horizon-dependent term in this bound can shrink arbitrarily if the distributions are similar enough. We then extend this result to the case when the policies' distributions can vary between rounds, thus addressing the general bandits with expert advice problem. In particular, we show that the same algorithm run with an adaptive learning rate enjoys a bound essentially of order $\sqrt{\sum_t \mathcal{C}_t \log N}$, where $\mathcal{C}_t$ is the capacity of the policies' distributions at round t. This bound is obtained as an implication of a stronger history-dependent bound, where the complexity of the policy set in a given round is represented through the mutual (chi-squared-)information between the chosen policy and the drawn outcome, conditioned on the events up to the previous round.

Still considering the EXP4 algorithm, we provide best-of-both-worlds bounds in Section 2.4. In the stochastic regime, we show that the algorithm enjoys a bound of order $\mathcal{C}\log T\log(NT)/\Delta$, where Δ is the minimum sub-optimality gap for the policies. Simultaneously, the algorithm is shown to retain a worst-case bound of order $\sqrt{\mathcal{C}T\log T\log N}$. The former bound follows from a more general guarantee that we provide for the adver-

sarially corrupted stochastic regime, an intermediate regime commonly considered in BOBW works (see, e.g., Dann et al., 2023; Ito et al., 2022). The proof of this result builds upon the techniques developed in (Ito et al., 2022) for proving BOBW bounds for a related algorithm in the setting of online learning with graph feedback.

We complement these results in Section 2.5 by proving worst-case lower bounds for three families of policy sets. These lower bounds scale with the policy set capacity in the same manner as the regret bound we provided for EXP4 in the adversarial setting, asserting its optimality up to factors logarithmic in the number of policies. Additionally, in Section 2.6, we prove another lower bound through which we aim to compare the studied feedback model with that of linear bandits. As alluded to before, the policies' distributions can be treated as a set of arms belonging to the simplex under an alternative linear bandit formulation of the problem, where the learner observes the inner product between the chosen policy and the latent loss map. Considering a particular family of policy sets, we prove that under linear bandit feedback, one must incur regret of order at least $\sqrt{NT}$ against any policy set in this family, even as the capacity approaches zero. This shows that the attainability of regret guarantees that improve as policies become more similar is a distinctive feature of mediator feedback, a key aspect of which is that the observed loss is attributed to a single outcome sampled via the chosen policy and revealed to the learner.

Finally, in Section 2.7, we consider a full-information variant of the problem, where the learner observes the entire loss map at every round. This can be seen as a generalization of the prediction with expert advice problem (Cesa-Bianchi et al., 1997). For this setting, we show that a simple Online Mirror Descent strategy enjoys a regret bound of order $\sqrt{\mathcal{C}_{\mathrm{KL}} T}$, where $\mathcal{C}_{\mathrm{KL}}$ is an altered form of the policy set capacity, based on the more standard KL-divergence. This quantity is no larger than $\log N$, and can be interpreted as an information radius of the policy set.

2.1.3 Additional Related Works

Krishnamurthy et al. (2020) study a contextual bandit problem with continuous actions, where the learner competes with a set of competitor policies mapping states (contexts) to actions. Instead of placing a smoothness assumption on the loss function, they opt for minimizing a notion of smoothed regret. More precisely, they fix a smoothing kernel that maps each action to a distribution over action. Accordingly, they obtain a new competitor class of smoothed policies that map states to distributions over actions by composing the original policies with the smoothing kernel. This has the effect of forcing a favourable

structure on the policy set. Indeed, they obtain (via EXP4) a bound of order $\sqrt{\kappa T \log N}$ in the adversarial regime, where N is the number of policies and κ, called the kernel complexity, is defined as the largest possible density assigned by the smoothing kernel with respect to some base probability measure. This latter quantity upper bounds the continuous analogue of $\mathcal{S}$ for any given context. Similar smoothed benchmarks are studied in (Majzoubi et al., 2020) and (Zhu and Mineiro, 2022) under realizability assumptions.

In its dependence on the mutual information between policies and outcomes, one of the regret bound we provide for EXP4 (see Theorem 2.2) bears some superficial resemblance to the PAC-Bayesian results in (Seldin et al., 2011). For a stochastic contextual bandit problem, Seldin et al. (2011) prove bounds on the per-round instantaneous regret that depend on the mutual information between the observed state and the chosen action. This quantifies the complexity of the adopted decision rule, which is traded off against its empirical regret measured according to past observations. Another notable mention is the information-theoretic analysis of Russo and Van Roy (2016) for Thompson sampling in Bayesian bandit problems, which results in bounds scaling with the mutual information between the faced environment and the optimal action, or some satisficing benchmark (Arumugam and Van Roy, 2021; Russo and Van Roy, 2022). This measures the amount of information that needs to be acquired about the environment to identify the target action.

Another related line of work concerns the best arm identification (BAI) problem, a variant of the MAB problem where the learner's aim is to find the optimal arm efficiently. Reddy et al. (2023) and Poiani et al. (2023) study the BAI problem with the added constraint that the learner can only sample arms via a number of given stochastic policies. As the objective remains identifying the optimal arm, the manner in which the structure of the policy set affects the achievable regret is fundamentally different from our setting. Indeed, in their problem, a more diverse policy set can be advantageous to the learner.

2.2 Preliminaries

In this section, we start by reviewing some concepts from information theory that will be referenced throughout the rest of the chapter. Then, we lay down a formal statement of the main problem setting.

2.2.1 Information Theory Background

Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and define the limits $f(0) = \lim_{x\to 0^+} f(x)$ and $f'(\infty) = \lim_{x\to\infty} f(x)/x$ (either of which could be infinite). If P and Q are two distributions (probability mass functions) on a common finite set Ω, the f-divergence (Ali and Silvey, 1966; Csiszár, 1967; Polyanskiy and Wu, 2023, Section 7.1) between them is defined as:

$$D_f(P \,\|\, Q) := \sum_{x\in\Omega} Q(x) f\Big(\frac{P(x)}{Q(x)}\Big),$$

with the understanding that $0f(0/0) = 0$ and $0f(a/0) = \lim_{x\to 0^+} xf(a/x) = af'(\infty)$ for $a > 0$. Notable properties of f-divergences include joint convexity in P and Q, non-negativity, and the fact that $D_f(P \,\|\, P) = 0$. Examples for f-divergences used in this work include

- $f(x) = (1/2)|x-1| \longrightarrow$ total variation distance:

$$\delta(P,Q) := \frac{1}{2}\sum_x |P(x) - Q(x)| = 1 - \sum_x \min\{P(x), Q(x)\}.$$

- $f(x) = (1/2)(\sqrt{x}-1)^2 \longrightarrow$ squared Hellinger distance:

$$H^2(P,Q) := \frac{1}{2}\sum_x (\sqrt{P(x)} - \sqrt{Q(x)})^2.$$

- $f(x) = (x-1)^2/(x+1) \longrightarrow$ triangular discrimination (also known as the Vincze–Le Cam divergence, see Sason and Verdú, 2016):

$$\Delta(P,Q) := \sum_x \frac{(P(x) - Q(x))^2}{P(x) + Q(x)}.$$

- $f(x) = x\log x \longrightarrow$ KL-divergence:

$$D(P \,\|\, Q) := \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

- $f(x) = (x-1)^2 \longrightarrow$ chi-squared divergence:

$$\chi^2(P \,\|\, Q) := \sum_x Q(x)\Big(\frac{P(x)}{Q(x)} - 1\Big)^2 = \sum_x \frac{P(x)^2}{Q(x)} - 1.$$

Let X and Y be two discrete random variables mapping Ω to the finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. The quantity defined as

$$D_f(P_{Y|X} \,\|\, Q_{Y|X} \,|\, P_X) \coloneqq \sum_{x \in \mathcal{X}} P_X(x) D_f(P_{Y|X=x} \,\|\, Q_{Y|X=x})$$

is known as the conditional f-divergence, where a summand corresponding to some $x \in \mathcal{X}$ is set to zero if $P_X(x) = 0$. An immediate property of f-divergences is that if $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$, then

$$D_f(P_{X,Y} \,\|\, Q_{X,Y}) = D_f(P_{Y|X} \,\|\, Q_{Y|X} \,|\, P_X)\,. \tag{2.1}$$

When jointly distributed according to $P_{X,Y}$, the mutual information between X and Y based on a given f-divergence (or their mutual f-information) is defined as (Polyanskiy and Wu, 2023, Section 7.8):

$$I_f(X;Y) \coloneqq D_f(P_{X,Y} \,\|\, P_X P_Y)\,,$$

which is the divergence between their joint distribution and the product of the marginals; hence, it is non-negative and symmetric in X and Y. Moreover, when f is strictly convex at unity (as is the case for the above examples), we have that $I_f(X;Y) = 0$ if and only if X and Y are independent (Makur and Zheng, 2020). Via (2.1), it holds that[1]

$$I_f(X;Y) = D_f(P_{Y|X} \,\|\, P_Y \,|\, P_X) = \textstyle\sum_x P_X(x) D_f(P_{Y|X=x} \,\|\, \sum_{x'} P_X(x') P_{Y|X=x'})\,.$$

The previous identity justifies the overloaded notion $I_f(P_X, P_{Y|X}) \coloneqq I_f(X;Y)$ formulating I_f as a function of P_X and the kernel $P_{Y|X}$.

When the f-divergence of choice is the KL-divergence, we obtain the standard mutual information, denoted simply as $I(X;Y)$, whereas $I_{\chi^2}(X;Y)$ will denote the mutual information based on the chi-squared divergence. The latter is bounded from above by $\min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$ considering that

$$\begin{aligned} I_{\chi^2}(X;Y) &= \sum_{x:\, P_X(x)>0} P_X(x) \sum_{y:\, P_{Y|X=x}(y)>0} P_{Y|X=x}(y) \frac{P_{Y|X=x}(y)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X=x'}(y)} - 1 \\ &\leq \sum_{x:\, P_X(x)>0} P_X(x) \sum_{y:\, P_{Y|X=x}(y)>0} P_{Y|X=x}(y) \frac{P_{Y|X=x}(y)}{P_X(x) P_{Y|X=x}(y)} - 1 \leq |\mathcal{X}| - 1\,, \end{aligned}$$

[1] By symmetry, this also holds when the roles of X and Y are exchanged.

which holds with equality if the distributions $\{P_{Y|X=x}\}_{x\in\mathcal{X}}$ have disjoint supports and P_X has full support. Symmetrically, we also have that $I_{\chi^2}(X;Y) \leq |\mathcal{Y}| - 1$. On the other hand, as $D(P\,\|\,Q) \leq \log(\chi^2(P\,\|\,Q)+1)$ (Polyanskiy and Wu, 2023, Section 7.6), we have that $I(X;Y) \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$. In particular, $I(X;Y) = \log|\mathcal{X}|$ holds if the distributions $\{P_{Y|X=x}\}_{x\in\mathcal{X}}$ have disjoint supports and P_X is uniform. Another distinction between the two quantities concerns their behaviour as functions of P_X. While for a fixed kernel $P_{Y|X}$, $I(P_X, P_{Y|X})$ is continuous in P_X (Cover and Thomas, 2006, Section 7.3), the same does not hold in general for I_{χ^2}. To see this, consider a simple instance where $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $P_{Y|X=0}(0) = 0.5$, and $P_{Y|X=1}(0) = 1$. If $P_X(0) = \varepsilon$ for some $\varepsilon \in (0,1]$, then

$$I_{\chi^2}(P_X, P_{Y|X}) = \frac{2-(3/2)\varepsilon}{2-\varepsilon} - \frac{1}{2} =: g(\varepsilon)\,,$$

which satisfies $\lim_{\varepsilon\to 0^+} g(\varepsilon) = 1/2$ even though $I_{\chi^2}(P_X, P_{Y|X}) = 0$ at $\varepsilon = 0$ by definition since X and Y become independent. Moreover, since g is decreasing in the interval $(0,1]$, I_{χ^2} as a function of P_X attains no maximum.

Maximizing the standard mutual information $I(P_X, P_{Y|X})$ in P_X gives rise to what we will refer to as the (KL-)information capacity of the kernel $P_{Y|X}$, denoted as:

$$\mathcal{C}_{\mathrm{KL}}(P_{Y|X}) \coloneqq \max_{P_X\in\mathcal{P}_\mathcal{X}} I(P_X, P_{Y|X})\,,$$

where $\mathcal{P}_\mathcal{X}$ is the set of possible distributions over the elements of $\mathcal{X}$. More practically, $\mathcal{C}_{\mathrm{KL}}(P_{Y|X})$ is better known as the information capacity of the discrete memoryless stationary channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition matrix $P_{Y|X}$ (Cover and Thomas, 2006, Chapter 7; Polyanskiy and Wu, 2023, Chapter 19). This quantity has an operational significance as it quantifies the highest rate per channel use at which information can be reliably sent (Cover and Thomas, 2006, Theorem 7.7.1; Polyanskiy and Wu, 2023, Theorem 19.9). Analogously, we define the χ^2-capacity of $P_{Y|X}$ as:

$$\mathcal{C}_{\chi^2}(P_{Y|X}) \coloneqq \sup_{P_X\in\mathcal{P}_\mathcal{X}} I_{\chi^2}(P_X, P_{Y|X})\,,$$

where we use the supremum in place of the maximum as the latter might not exist per the counterexample provided earlier.

2.2.2 Problem Setting

Let $\mathcal{X} := \{1, \dots, K\}$ denote a set of $K \geq 2$ outcomes, and let $\Theta := \{\theta_1, \dots, \theta_N\} \subset \Delta_K$ denote a policy set consisting of $N \geq 2$ distributions over the outcomes, where, as before, Δ_K is the probability simplex in $\mathbb{R}^K$. Hence, for an outcome $x \in \mathcal{X}$ and policy $\theta \in \Theta$, $\theta(x)$ denotes the probability assigned to x by θ. We consider a mediator feedback problem where a learner, possessing full knowledge of the policy set Θ, plays a sequential game with an unknown environment for T rounds. From the environment's characteristic distribution,[2] a latent sequence of loss vectors $(\ell_t)_{t=1}^T$ is drawn at the beginning of the game, where $\ell_t \in [0,1]^K$ maps each outcome to a loss at the t-th round. Ensuingly, the learner sequentially interacts with the environment by selecting at each round t a policy $\vartheta_t \in \Theta$, possibly at random, and subsequently observing the pair $(X_t, \ell_t(X_t))$, where X_t is a random outcome distributed according to ϑ_t. Slightly overloading the notation, we let $\ell_t(\theta)$ denote the expected value (conditioned on ℓ_t) of $\ell_t(X_t)$ had the learner picked policy θ at round t; that is, $\ell_t(\theta) := \sum_{x \in \mathcal{X}} \theta(x) \ell_t(x)$. The learner's objective is to minimize their regret, defined as:

$$R_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\vartheta_t)\right] - \min_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\theta)\right],$$

where the expectation is taken over both the learner's and the environment's randomization. We use a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to define all random variables. For round $t \in [T]$, let $\mathcal{H}_t := (\vartheta_s, X_s, \ell_s(X_s))_{s=1}^t$ denote the interaction history up to the end of round t, and let $\mathcal{F}_t := \sigma(\mathcal{H}_t)$ denote the σ-algebra generated by $\mathcal{H}_t$. Accordingly, we define $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and $\mathbb{P}^t(\cdot) := \mathbb{P}(\cdot \mid \mathcal{F}_{t-1})$, with $\mathcal{F}_0$ being the trivial σ-algebra. Analogously to (Russo and Van Roy, 2016), we define $I^t(X;Y)$ and $I^t_{\chi^2}(X;Y)$ as the mutual information and the mutual chi-squared-information between (discrete) random variables X and Y with $\mathbb{P}^t$ as the base measure. Notice that these quantities are random variables owing to their dependence on the history.

[2] This formulation subsumes the standard (non-adaptive) adversarial case, where an environment is characterized by a deterministic sequence of loss vectors.

2.3 The Policy Set Capacity: An Improved Regret Bound for EXP4

Before defining the policy set capacity, we provide some context by briefly reviewing some quantities used in related works to describe the richness of the policy set. McMahan and Streeter (2009) introduce the quantity

$$\mathcal{S}(\Theta) := \sum_{x \in \mathcal{X}} \max_{\theta \in \Theta} \theta(x) .$$

It is easily verified that $1 \leq \mathcal{S}(\Theta) \leq \min\{K, N\}$, where the lower bound is attained in the limit case when all the policies are identical, and the upper bound is attained either when the policies have disjoint supports or when each outcome is matched with a policy entirely concentrated on that outcome. To get a finer sense of this quantity, we define $\mathcal{V}(\Theta) := \mathcal{S}(\Theta) - 1$. Notice then that when there are only two policies, that is, $\Theta = \{\theta_1, \theta_2\}$, $\mathcal{V}$ reduces to the total variation distance between the two distributions:

$$\mathcal{V}(\{\theta_1, \theta_2\}) = \sum_x \max\{\theta_1(x), \theta_2(x)\} - 1 = 1 - \sum_x \min\{\theta_1(x), \theta_2(x)\} = \delta(\theta_1, \theta_2) .$$

More generally, one can obtain the somewhat coarse bound:

$$\mathcal{V}(\Theta) \leq \min_{\alpha \in \Delta_K} \sum_\theta \delta(\theta, \alpha) ,$$

where the distribution minimizing the right-hand side acts as the geometric median of Θ in terms of the total variation distance. This inequality follows from Theorem II.1 in (Guntuboyina, 2011), which for any f-divergence and any $\alpha \in \Delta_K$, provides the (implicit) bound:

$$f(\mathcal{S}(\Theta)) + (N-1) f\Big(\frac{N - \mathcal{S}(\Theta)}{N-1}\Big) \leq \sum_\theta D_f(\theta \,\|\, \alpha) .$$

Another relevant quantity is the chi-squared "diameter" of the policy set:

$$d_{\chi^2}(\Theta) := \max_{\theta, \theta' \in \Theta} \chi^2(\theta \,\|\, \theta') ,$$

which is featured in the regret bounds of Sen et al. (2018) and Papini et al. (2019), see Section 2.4. Though it has no general upper bound, d_{χ^2} can be smaller than $\mathcal{V}$ as shown in the examples section below. Sen et al. (2018) also obtain bounds in terms of another

diameter-like quantity based on the logarithm of (one plus) the f-divergence with $f(x) = x\exp(x-1)-1$, though d_{χ^2} is never larger.

2.3.1 The Policy Set Capacity

Let ϑ and X be two random variables taking values respectively over Θ and $\mathcal{X}$ such that $\mathbb{P}_{X|\vartheta=\theta}(x) = \theta(x)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. Then, we define the (chi-squared) capacity of the policy set as:

$$\mathcal{C}(\Theta) \coloneqq \mathcal{C}_{\chi^2}(\mathbb{P}_{X|\vartheta}),$$

which does not depend on the distribution of ϑ. More explicitly, if we define

$$Q_\tau(\Theta) \coloneqq I_{\chi^2}(\tau, \mathbb{P}_{X|\vartheta}) = \textstyle\sum_\theta \tau(\theta)\chi^2\big(\theta \,\big\|\, \sum_{\theta'}\tau(\theta')\theta'\big),$$

for some distribution $\tau \in \mathcal{P}_\Theta$;[3] then,

$$\mathcal{C}(\Theta) = \sup_{\tau\in\mathcal{P}_\Theta} Q_\tau(\Theta).$$

As alluded to before, this definition inspires an interpretation of the policy set as inducing a stationary, memoryless channel defined via the kernel $\mathbb{P}_{X|\vartheta}$. Intuitively, $\mathcal{C}(\Theta)$ can be seen to measure the dependency between ϑ and X maximised over the prior distribution of ϑ. Hence, the more dissimilar the distributions are, the larger this quantity.

Since $\mathcal{C}$ is based on I_{χ^2}, it satisfies $0 \le \mathcal{C}(\Theta) \le \min\{K, N\} - 1$, which is the same range as that of $\mathcal{V}$. In particular, much like $\mathcal{V}$, the upper bound is attained either when the policies have disjoint supports or when each outcome is matched with a policy entirely concentrated on that outcome. Moreover, $\mathcal{C}(\Theta) = 0$ if and only if X and ϑ are independent no matter how ϑ is distributed, which requires the policies to be identical. More distinctively, it holds in general that $\mathcal{C}(\Theta) \le \min\{\mathcal{V}(\Theta), d_{\chi^2}(\Theta)\}$. On the one hand, the (joint) convexity of the chi-squared divergence implies that

$$\mathcal{C}(\Theta) \le \sup_{\tau\in\mathcal{P}_\Theta} \textstyle\sum_{\theta,\theta'} \tau(\theta)\tau(\theta')\chi^2(\theta\,\|\,\theta') \le \max_{\theta,\theta'} \chi^2(\theta\,\|\,\theta') = d_{\chi^2}(\Theta).$$

[3] Inline with previous notation, $\mathcal{P}_\Theta$ denotes the set of possible distributions over the policies.

Algorithm 2.1 EXP4 (fixed policy set)

input: sequence of learning rates $(\eta_t)_{t=1}^T$
initialize: $\forall \theta \in \Theta$, $\widehat{\ell}_0(\theta) = 0$
for $t = 1, \dots, T$ **do**
  draw $\vartheta_t \sim p_t$, where $p_t(\theta) = \frac{\exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta))}{\sum_{\theta'} \exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta'))}$
  draw $X_t \sim \vartheta_t$, and observe loss $\ell_t(X_t)$
  $\forall \theta \in \Theta$, set $\widehat{\ell}_t(\theta) = \frac{\theta(X_t)}{\sum_{\theta'} p_t(\theta')\theta'(X_t)} \ell_t(X_t)$
end for

On the other hand, we have that

$$\begin{aligned}
\mathcal{C}(\Theta) &= \sup_{\tau \in \mathcal{P}_\Theta} \sum_\theta \tau(\theta) \Big(\sum_x \frac{\theta(x)^2}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \Big) \\
&= \sup_{\tau \in \mathcal{P}_\Theta} \sum_x \frac{\sum_\theta \tau(\theta)\theta(x)^2}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \\
&\le \sup_{\tau \in \mathcal{P}_\Theta} \sum_x \max_{\theta''} \theta''(x) \frac{\sum_\theta \tau(\theta)\theta(x)}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \\
&= \sum_x \max_\theta \theta(x) - 1 = \mathcal{S}(\Theta) - 1 = \mathcal{V}(\Theta) \,.
\end{aligned}$$

2.3.2 A Regret Bound for EXP4 in Terms of the Capacity

EXP4, detailed in Algorithm 2.1, adopts a simple and natural approach for tackling mediator feedback problems. Its choice of policy in a given round is drawn from a running distribution over the policies taking an exponential weights form. There, each policy θ is weighted according to a proxy of the sum of its losses so far, where an importance-weighted estimator $\widehat{\ell}_t(\theta)$ replaces the inaccessible $\ell_t(\theta)$. The following theorem provides a regret bound for EXP4 that scales with the policy set capacity. This result improves upon the $\sqrt{\mathcal{S}(\Theta) T \log N}$ bound, seemingly the best available worst-case bound for the considered setting. Further, we instantiate the capacity in the ensuing discussion for three families of policy sets for which the bound of this theorem will be shown to be near-optimal in Section 2.5. While the proposed learning rate schedule requires exact knowledge of the capacity, this requirement will be lifted in Theorem 2.2, which also addresses the case when the policies' distributions can vary between rounds.

Theorem 2.1. *Algorithm 2.1 with* $\eta_t = \min\Big\{1, \sqrt{\frac{\log N}{e\mathcal{C}(\Theta) t}}\Big\}$ *satisfies*

$$R_T \le 2 \max\Big\{\sqrt{e\mathcal{C}(\Theta) T \log N}, \log N\Big\} \,.$$

Proof. Let $\theta^* \in \arg\min_{\theta\in\Theta} \mathbb{E}\sum_{t=1}^{T} \ell_t(\theta)$. For a policy θ, we define a shifted version of the loss at time t as $\zeta_t(\theta) := \sum_x \big(\theta(x) - \psi_t(x)\big)\ell_t(x)$, where $\psi_t(x) := \sum_\theta p_t(\theta)\theta(x)$. Thus, $\zeta_t(\theta) = \ell_t(\theta) - \sum_{\theta'} p_t(\theta')\ell_t(\theta')$. Notice that for any two policies θ and θ', $\zeta_t(\theta) - \zeta_t(\theta') = \ell_t(\theta) - \ell_t(\theta')$. Hence, $R_T = \mathbb{E}\sum_t \big(\ell_t(\vartheta_t) - \ell_t(\theta^*)\big) = \mathbb{E}\sum_t \big(\zeta_t(\vartheta_t) - \zeta_t(\theta^*)\big)$. Next, we define $\widehat{\zeta}_t(\theta)$ as an estimate of the shifted loss of θ at time t:

$$\widehat{\zeta}_t(\theta) := \big(\theta(X_t) - \psi_t(X_t)\big)\frac{\ell_t(X_t)}{\psi_t(X_t)} = \widehat{\ell}_t(\theta) - \ell_t(X_t)\,. \tag{2.2}$$

For convenience, we will sometimes treat the distribution p_t as a vector belonging to the simplex $\Delta_N \subset \mathbb{R}^N$, where its i-th coordinate $p_t(i)$ denotes $p_t(\theta_i)$ for each $i \in [N]$. Analogously, the functions $\widehat{\ell}_t$ and $\widehat{\zeta}_t$ will sometimes be handled as vectors in $\mathbb{R}^N$. Notice that p_t and ψ_t are measurable with respect to $\mathcal{F}_{t-1}$, and that ℓ_t is independent of ϑ_t and X_t conditioned on $\mathcal{F}_{t-1}$. Hence, it holds that $\mathbb{E}_t\zeta_t(\vartheta_t) = \mathbb{E}_t\sum_\theta p_t(\theta)\zeta_t(\theta)$, and that $\mathbb{E}_t\widehat{\zeta}_t(\theta) = \mathbb{E}_t\zeta_t(\theta)$ for any fixed $\theta \in \Theta$. Consequently, thanks to the tower rule and the linearity of expectation, we have that

$$\mathbb{E}\sum_t \big(\zeta_t(\vartheta_t) - \zeta_t(\theta^*)\big) = \mathbb{E}\sum_t \langle p_t - \mathbf{e}_{\theta^*}, \zeta_t\rangle = \mathbb{E}\sum_t \langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle\,,$$

where $\mathbf{e}_{\theta^*} \in \mathbb{R}^N$ is the indicator vector for θ^*.

It is well known (see Shalev-Shwartz, 2012, Section 2.7) that for every round t, the definition of p_t in Algorithm 2.1 is equivalent to

$$p_t = \operatorname*{arg\,min}_{p\in\Delta_N} \eta_t\Big\langle \sum_{s=1}^{t-1} \widehat{\ell}_s, p\Big\rangle - H(p)\,, \tag{2.3}$$

where $H(p) := \sum_{i=1}^{N} p(i)\log(1/p(i))$ is the Shannon entropy of p. Note that for any $p \in \Delta_N$,

$$\Big\langle \sum_{s=1}^{t-1} \widehat{\zeta}_s, p\Big\rangle = \Big\langle \sum_{s=1}^{t-1} \widehat{\ell}_s, p\Big\rangle - \sum_{s=1}^{t-1} \ell_s(X_s)\,.$$

Hence, by adding constant terms (i.e., not depending on p) to the objective function in (2.3) and changing the scaling, we can arrive at the following alternative characterization of p_t for $t \in [T+1]$:

$$p_t = \operatorname*{arg\,min}_{p\in\Delta_N} \Big\langle \sum_{s=1}^{t-1} \widehat{\zeta}_s, p\Big\rangle + \frac{1}{\eta_t}\big(\log N - H(p)\big)\,,$$

which is equivalent to the update rule of the FTRL algorithm when executed on the losses $(\widehat{\zeta}_t)_{t\in[T]}$ with Δ_N as the decision set and a sequence of regularizers $(\phi_t)_{t\in[T+1]}$ where

$$\phi_t(p) = \frac{1}{\eta_t}\big(\log N - H(p)\big) \qquad \forall p \in \Delta_N\,,$$

which is the negative Shannon entropy normalized to the range $[0, \log N]$ and scaled by the learning rate. Let $D_{\phi_t}(\cdot;\cdot)$ be the Bregman divergence based on ϕ_t, and set $\eta_{T+1} = \eta_T$. We can then use Lemma 7.14 in (Orabona, 2023) to obtain the following regret bound for FTRL on the estimated shifted losses:

$$\textstyle\sum_t \langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle \le \dfrac{\log N}{\eta_T} + \dfrac{1}{2}\sum_t \eta_t \langle z_t, \widehat{\zeta}_t^2\rangle\,,$$

where z_t lies on the line segment between p_t and $\widetilde{p}_{t+1} = \arg\min_{u\in\mathbb{R}^N_{\ge 0}} \langle \widehat{\zeta}_t, u\rangle + D_{\phi_t}(u; p_t)$. By its definition, it is easy to show that $\widetilde{p}_{t+1}(i) = p_t(i)\exp(-\eta_t\widehat{\zeta}_t(i))$ for every $i \in [N]$. Notice that $\eta_t\widehat{\zeta}_t(i) \ge -\eta_t\ell_t(X_t) \ge -\eta_t \ge -1$ since $\widehat{\ell}_t$ is non-negative, $\eta_t \in (0,1]$, and $\ell_t(X_t) \le 1$. Hence, it holds for every $i \in [N]$ that $\widetilde{p}_{t+1}(i) \le e\, p_t(i)$, implying that $\langle z_t, \widehat{\zeta}_t^2\rangle \le e\langle p_t, \widehat{\zeta}_t^2\rangle$.

Overall, we have shown that

$$R_T \le \frac{\log N}{\eta_T} + \frac{e}{2}\textstyle\sum_t \eta_t \mathbb{E}\sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\,. \tag{2.4}$$

Now, for every $\theta \in \Theta$ and $t \in [T]$, we have that

$$\begin{aligned}
\mathbb{E}_t\widehat{\zeta}_t(\theta)^2 &= \mathbb{E}_t\big(\theta(X_t) - \psi_t(X_t)\big)^2 \frac{\ell_t(X_t)^2}{\psi_t(X_t)^2} \\
&\le \mathbb{E}_t \frac{\big(\theta(X_t) - \psi_t(X_t)\big)^2}{\psi_t(X_t)^2} \\
&= \mathbb{E}_t \sum_x \frac{\big(\theta(x) - \psi_t(x)\big)^2}{\psi_t(x)^2}\mathbb{I}\{x = X_t\} \\
&= \sum_x \frac{\big(\theta(x) - \psi_t(x)\big)^2}{\psi_t(x)} \\
&= \sum_x \psi_t(x)\Big(\frac{\theta(x)}{\psi_t(x)} - 1\Big)^2 = \chi^2(\theta\,\|\,\psi_t) = \chi^2\big(\theta \,\big\|\, \textstyle\sum_{\theta'} p_t(\theta')\theta'\big)\,,
\end{aligned}$$

where the third equality holds since $\mathbb{E}_t\mathbb{I}\{x = X_t\} = \mathbb{P}^t(x = X_t) = \psi_t(x)$. This implies that

$$\mathbb{E}_t\textstyle\sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2 \le \sum_\theta p_t(\theta)\chi^2\big(\theta \,\big\|\, \sum_{\theta'} p_t(\theta')\theta'\big) = Q_{p_t}(\Theta)\,. \tag{2.5}$$

Consequently, we arrive at the following bound:

$$R_T \le \frac{\log N}{\eta_T} + \frac{e}{2}\mathbb{E}\sum_t \eta_t Q_{p_t}(\Theta) \le \frac{\log N}{\eta_T} + \frac{e}{2}\mathcal{C}(\Theta)\sum_t \eta_t .$$

If $T \ge \frac{\log N}{e\mathcal{C}(\Theta)}$, then $\eta_T = \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)T}}$ and $R_T \le 2\sqrt{e\mathcal{C}(\Theta)T\log N}$, where we have used that $\eta_t \le \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)t}}$ for $t \in [T]$ and that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$. Otherwise, if $T < \frac{\log N}{e\mathcal{C}(\Theta)}$, then $\eta_1 = \cdots = \eta_T = 1$ and $R_T \le \log N + \frac{e\mathcal{C}(\Theta)T}{2} \le 2\log N$. □

Compared to the standard analysis of EXP4 (see, e.g., Bubeck and Cesa-Bianchi, 2012, Section 4.2) or the analysis of McMahan and Streeter (2009), the main nuance in the proof is the use of a more refined bound on the second moment of the estimated losses, thanks to which, the dependence on the capacity occurs naturally in the regret bound. Moreover, carrying out the analysis in terms of the shifted losses (introduced in the proof) causes the horizon-dependent term in the regret bound to feature $\mathcal{C}(\Theta)$ directly instead of $1+\mathcal{C}(\Theta)$, which can lead to a substantial improvement whenever $\mathcal{C}(\Theta) \ll 1$. This distinction is particularly relevant, as discussed in the introduction, when the number of policies is very small (e.g., just two), as in that case, a bound scaling with $1+\mathcal{C}(\Theta)$ would not improve much on the trivial worst-case bound of $\sqrt{NT}$, no matter how benign the policy set is. Interestingly, all prior works seem to suffer from this shortcoming.

2.3.3 Examples

We now examine the quantity $\mathcal{C}(\Theta)$ for a selection of policy set structures and compare it with related quantities.

Two Policies

We start with the case when the policy set consists of only two policies, i.e., $\Theta = \{\theta_1, \theta_2\}$. As mentioned before, we have that $\mathcal{V}(\Theta) = \delta(\theta_1, \theta_2)$, while $d_{\chi^2} = \max\{\chi^2(\theta_1 \,\|\, \theta_2), \chi^2(\theta_2 \,\|\, \theta_1)\}$. These two quantities are incomparable in general, and this can be seen by specializing the next example to the two policies case. For a fixed $r \in [0,1]$, define $q_r(\theta_1 \,\|\, \theta_2) = Q_\tau(\Theta)$ with $\tau(\theta_1) = r$; hence, $\mathcal{C}(\Theta) = \sup_{r\in[0,1]} q_r(\theta_1 \,\|\, \theta_2) =: \mathcal{C}(\theta_1, \theta_2)$. An explicit form for $q_r(\theta_1 \,\|\, \theta_2)$ is given by:

$$q_r(\theta_1 \,\|\, \theta_2) = r(1-r)\sum_x \frac{(\theta_1(x) - \theta_2(x))^2}{r\theta_1(x) + (1-r)\theta_2(x)} .$$

As a function of $r \in [0,1]$, $q_r(\theta_1 \,\|\, \theta_2)$ is concave with $q_0(\theta_1 \,\|\, \theta_2) = q_1(\theta_1 \,\|\, \theta_2) = 0$. This quantity is known in the literature as the Vincze–Le Cam divergence of order r (Makur and Zheng, 2020; Raginsky, 2016), which is an f-divergence with $f(x) = \frac{r(1-r)(x-1)^2}{r(x-1)+1}$. Corresponding to $r = 1/2$ is (half) the triangular discrimination Δ, immediately implying a lower bound for $\mathcal{C}$:

$$\mathcal{C}(\theta_1,\theta_2) \geq q_{1/2}(\theta_1 \,\|\, \theta_2) = \frac{1}{2}\sum_x \frac{(\theta_1(x)-\theta_2(x))^2}{\theta_1(x)+\theta_2(x)} = \frac{1}{2}\Delta(\theta_1,\theta_2)\,.$$

On the other hand, since $\frac{r(1-r)(x-1)^2}{r(x-1)+1} \leq (\sqrt{x}-1)^2$ for any $r \in (0,1)$ and $x \in [0,\infty)$, we can bound $\mathcal{C}$ in terms of the squared Hellinger distance H^2:

$$\mathcal{C}(\theta_1,\theta_2) \leq \sum_x \theta_2(x)\left(\sqrt{\theta_1(x)/\theta_2(x)}-1\right)^2 = 2H^2(\theta_1,\theta_2)\,. \tag{2.6}$$

Combining these observations with known inequalities (Topsoe, 2000), we obtain that

$$\delta^2 \leq \frac{\Delta}{2} \leq \mathcal{C} \leq 2H^2 \leq \Delta \leq 2\delta\,,$$

which shows that the capacity of two policies is of the same order as the squared Hellinger distance and the triangular discrimination. For the two policies case, we prove in Theorem 2.5 a lower bound of $\Omega(\sqrt{\mathcal{C}(\theta_1,\theta_2)\,T})$, which order-wise matches the bound of Theorem 2.1.

ε-Greedy Policies

Consider now a case in which $N = K$ and each policy θ is associated (one-to-one) with an outcome x_θ such that for any outcome x, $\theta(x) = (1-\varepsilon)/N + \varepsilon\mathbb{I}\{x = x_\theta\}$, where $\varepsilon \in [0,1]$. At $\varepsilon = 0$, all policies collapse to the uniform distribution, and we get that $\mathcal{C}(\Theta) = 0$. On the other hand, when $\varepsilon = 1$, the problem essentially reduces to a standard (unstructured) bandit problem with $\mathcal{C}(\Theta) = N-1$. Generally, for $\tau \in \mathcal{P}_\Theta$, $Q_\tau(\Theta)$ takes the following form:

$$Q_\tau(\Theta) = \varepsilon^2 \sum_\theta \frac{\tau(\theta)(1-\tau(\theta))}{\frac{1-\varepsilon}{N}+\varepsilon\tau(\theta)}\,.$$

For intermediate values of $\varepsilon \in (0,1)$, $Q_\tau(\Theta)$ is a strictly concave function in τ attaining its maximum value at the uniform distribution, entailing that $\mathcal{C}(\Theta) = \varepsilon^2(N-1)$. In comparison,

we have that

$$\mathcal{V}(\Theta) = \varepsilon(N-1) \qquad \text{and} \qquad d_{\chi^2}(\Theta) = \frac{\varepsilon(N-2)+2}{\varepsilon(N-1)+1} \cdot \frac{\varepsilon^2}{1-\varepsilon} N .$$

Notice that even though d_{χ^2} grows unbounded as ε approaches 1, it can be smaller than $\mathcal{V}$ for small enough ε. In Theorem 2.6, we prove a lower bound of order $\varepsilon\sqrt{NT}$ for this policy set structure, which matches the upper bound of Theorem 2.1 up to a logarithmic factor.

M-Supported Uniform Policies

Consider another policy set structure where all policies are uniform distributions over a support of $M \leq K$ outcomes. That is, if we denote by $\mathrm{Supp}(\theta)$ the support for policy θ, then we have that $\mathrm{Supp}(\theta) = M$ and for any outcome x, $\theta(x) = (1/M)\mathbb{I}\{x \in \mathrm{Supp}(\theta)\}$. Assume further that each outcome belongs to the support of at least one policy. For this structure, one can verify that $\mathcal{C}(\Theta) = \mathcal{V}(\Theta) = K/M - 1$. In fact, we have that $Q_\tau(\Theta) = \mathcal{C}(\Theta)$ for any $\tau \in \mathcal{P}_\Theta$ with full support. On the other hand, $d_{\chi^2} = \infty$ outside of the trivial case when $M = K$. In Theorem 2.7, we show that for a special family of M-supported uniform policies (where $N \geq K/M$), the regret of any algorithm is $\Omega\big(\sqrt{(K/M-1)\,T\log(N)/\log(K/M)}\big)$. This lower bound particularly shows that the logarithmic factor in the regret bound of Theorem 2.1 is at least partly unavoidable.

2.3.4 A Generalization for Time-Varying Policy Distributions

The next theorem extends the result of Theorem 2.1 by allowing the distributions of the policies to vary between rounds, modelling in this manner the problem of bandits with expert advice. We rely on an adaptive learning rate schedule, whose form is common in the online learning and bandits literature (Auer et al., 2002b; McMahan and Streeter, 2010; Neu, 2015a). The resulting bound replaces the dependence on the (per-round) capacity with the history-conditioned mutual chi-squared-information between the chosen policy and the drawn outcome, recalling that the former is an upper bound for the latter by definition. Additionally, the adopted learning rate in a given round only requires an upper bound on the capacity of the policies' distributions, thus affording one the flexibility of providing a quantity of simpler form like $\mathcal{V}$, or even just $\min\{N, K\}$. In terms of regret, this flexibility is paid for through a solitary additive term depending primarily on the largest of these provided bounds.

Algorithm 2.2 EXP4 (time-varying policy distributions)

input: sequence of learning rates $(\eta_t)_{t=1}^T$
initialize: $\forall \theta \in \Theta$, $\widehat{\ell}_0(\theta) = 0$
for $t = 1, \dots, T$ **do**
  observe distributions $\theta(\cdot; t) \quad \forall \theta \in \Theta$
  draw $\vartheta_t \sim p_t$, where $p_t(\theta) = \frac{\exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta))}{\sum_{\theta'} \exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta'))}$
  draw $X_t \sim \vartheta_t$, and observe loss $\ell_t(X_t)$
  $\forall \theta \in \Theta$, set $\widehat{\ell}_t(\theta) = \frac{\theta(X_t;t)}{\sum_{\theta'} p_t(\theta')\theta'(X_t;t)} \ell_t(X_t)$
end for

Only in the current scope, a member θ of the policy set Θ is not synonymous with a distribution over the outcomes; it serves solely as an identifier for a policy. Along with the sequence of losses, the environment draws for each $\theta \in \Theta$ a sequence of distributions $(\theta(\cdot; t))_{t=1}^T$ at the beginning of the game, where $\theta(x; t)$ is the probability assigned to outcome x by policy θ at round t. The distributions $(\theta(\cdot; t))_{\theta \in \Theta}$ associated with a given round t are revealed to the learner at the beginning of the round. Accordingly, we redefine the interaction history as $\mathcal{H}_t := \big((\vartheta_s, X_s, \ell_s(X_s))_{s \in [t]}, (\theta(\cdot; s))_{\theta \in \Theta, s \in [t+1]}\big)$, which includes all the information available to the learner before choosing a policy at round $t+1$.[4] The definition of the regret remains the same, only that the loss of policy θ at round t is now defined as $\ell_t(\theta) := \sum_x \theta(x; t)\ell_t(x)$. For a distribution $\tau \in \mathcal{P}_\Theta$ and round $t \in [T]$, we define

$$Q_{t,\tau} := \sum_\theta \tau(\theta) \chi^2\big(\theta(\cdot; t) \,\big\|\, \textstyle\sum_{\theta'} \tau(\theta')\theta'(\cdot; t)\big)$$

and $\mathcal{C}_t := \sup_{\tau \in \mathcal{P}_\Theta} Q_{t,\tau}$ as the time-varying analogues of Q_τ and $\mathcal{C}$. The following theorem still concerns the plain EXP4 algorithm, reformulated in Algorithm 2.2 for the time-varying case. Observe that for any round t, $I^t_{\chi^2}(\vartheta_t; X_t) = Q_{t,p_t}$.

Theorem 2.2. *Let $Z_t := \sum_{s=1}^t I^s_{\chi^2}(\vartheta_s; X_s)$ for all $t \in [T]$, and let $(J_t)_{t=1}^T$ be a non-decreasing sequence of non-negative real numbers such that J_t is $\mathcal{F}_{t-1}$-measurable and $J_t \geq \mathcal{C}_t$. Then, Algorithm 2.2 with $\eta_t = \sqrt{\frac{\log N}{\log N + e(Z_{t-1} + J_t)}}$ satisfies*

$$\begin{aligned} R_T &\leq \mathbb{E}\Big[2\sqrt{e \textstyle\sum_t I^t_{\chi^2}(\vartheta_t; X_t) \log N} + \log N + \sqrt{e J_T \log N}\Big] \\ &\leq \mathbb{E}\Big[2\sqrt{e \textstyle\sum_t \mathcal{C}_t \log N} + \log N + \sqrt{e J_T \log N}\Big]. \end{aligned}$$

[4] Consistently, the usage of the filtration $(\mathcal{F}_t)_t$ (and dependent quantities) in the context of the following result refers to this augmented definition of the history.

Proof. For any $\theta \in \Theta$ and $t \in [T]$, let, similarly to (2.2), $\widehat{\zeta}_t(\theta) := \widehat{\ell}_t(\theta) - \ell_t(X_t)$, with $\widehat{\ell}_t(\theta)$ as defined in Algorithm 2.2. Notice that the distributions $(\theta(\cdot;t))_{\theta\in\Theta}$ are measurable with respect to $\mathcal{F}_{t-1}$. Moreover, the sequence $(\eta_t)_t$ is non-increasing and $\eta_t \leq 1$ holds for all rounds. Hence, with the same arguments laid out in the proof of Theorem 2.1, one can show that

$$R_T \leq \mathbb{E}\Big[\frac{\log N}{\eta_T} + \frac{e}{2}\sum_t \eta_t \sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\Big].$$

Furthermore, similar to what was shown in the proof of Theorem 2.1, it holds for every t that $\mathbb{E}_t \sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2 \leq Q_{t,p_t} = I^t_{\chi^2}(\vartheta_t; X_t)$. Hence, since η_t is $\mathcal{F}_{t-1}$-measurable, we get that

$$R_T \leq \mathbb{E}\Big[\frac{\log N}{\eta_T} + \frac{e}{2}\sum_t \eta_t I^t_{\chi^2}(\vartheta_t; X_t)\Big].$$

We then conclude the proof by bounding the two terms inside the expectation. Starting with the second term, we have that

$$\begin{aligned}\frac{e}{2}\sum_t \eta_t I^t_{\chi^2}(\vartheta_t; X_t) &= \frac{\sqrt{\log N}}{2}\sum_t \frac{e I^t_{\chi^2}(\vartheta_t; X_t)}{\sqrt{\log N + e\,(Z_{t-1} + J_t)}} \\ &\leq \frac{\sqrt{\log N}}{2}\sum_t \frac{e I^t_{\chi^2}(\vartheta_t; X_t)}{\sqrt{e Z_t}} \leq \sqrt{e Z_T \log N},\end{aligned}$$

where the last inequality follows via Lemma 3.5 in (Auer et al., 2002b). Whereas

$$\frac{\log N}{\eta_T} = \sqrt{\log^2 N + e\,(Z_{T-1} + J_T)\log N} \leq \sqrt{e Z_T \log N} + \log N + \sqrt{e J_T \log N}.$$

□

Let $\mathcal{S}_t := \sum_x \max_\theta \theta(x;t)$ and $\mathcal{V}_t := \mathcal{S}_t - 1$. A reasonable choice is to set $J_t = \max_{s\leq t} \mathcal{V}_s$, which would only cause the bound to concede an added term of $\sqrt{e \max_{t\leq T} \mathcal{V}_t \log N}$ while lifting the more burdensome requirement of computing the capacity at each round. Notice that the first bound of Theorem 2.2 depends (in expectation) on the observed sequence of losses through its dependence on the algorithm's decision at each round (i.e., p_t). However, it is unclear whether this bound can take advantage of any particular benign property of the losses when compared with the second bound, which only depends on the policies' distributions. Nevertheless, for what concerns the bandits with expert advice problem, these bounds improve upon the state of the art bound of $\sqrt{\sum_t \mathcal{S}_t \log N}$ reported in (Lattimore and Szepesvári, 2020, Theorem 18.3).

2.4 Best-of-Both-Worlds Bounds

Besides the adversarial regime considered thus far, we study in this section a more benign setting where the dependence of the regret on the time horizon can be improved. Specifically, we will consider what we will refer to as the adversarially corrupted stochastic regime, where it is assumed that there exists a policy $\xi \in \Theta$ such that for every round t and policy $\theta \neq \xi$,

$$\mathbb{E}\big[\ell_t(\theta) - \ell_t(\xi)\big] \geq \Delta - B_t\,,$$

for some $\Delta \in (0,1]$ and $B_t \geq 0$. Additionally, we define $B \coloneqq \sum_{t=1}^{T} B_t$. This includes, as a special case, the canonical stochastic regime where the loss functions $(\ell_t)_t$ are independently and identically distributed across rounds. Notice that besides the addition of corruption, the more general stochastic regime we consider does not require the losses to be distributed either stationarily or independently. For similar setups, see, for example, (Dann et al., 2023; Ito et al., 2022; Wei and Luo, 2018; Zimmert and Seldin, 2021).

The main result of this section concerns, once again, the EXP4 algorithm, which we show to enjoy BOBW bounds when coupled with a certain learning rate schedule. For the stochastic regime, the algorithm achieves a bound linear in the capacity and only polylogarithmic in the time horizon. Simultaneously, it retains roughly the same worst-case guarantee as that of Theorem 2.1. This result, provided in the next theorem, is obtained by combining elements from the proof of Theorem 2.1 with the learning rate schedule and analysis technique used by Ito et al. (2022) in the setting of online learning with strongly observable feedback graphs. Before stating the theorem, we define

$$P(\theta) \coloneqq \sum_{t=1}^{T} (1 - p_t(\theta)) \qquad \text{and} \qquad \overline{P}(\theta) \coloneqq \mathbb{E} \sum_{t=1}^{T} (1 - p_t(\theta))$$

for every policy $\theta \in \Theta$, where $p_t(\theta) = \mathbb{P}^t(\vartheta_t = \theta)$ as specified in Algorithm 2.1. Moreover, for any distribution p over the policies, we let $H(p) \coloneqq \sum_\theta p(\theta) \log(1/p(\theta))$ denote its Shannon entropy as before.

Theorem 2.3. *Let* $\gamma = \sqrt{\frac{e\mathcal{C}(\Theta)\log(eT)}{2\log(N)}}$*, and suppose Algorithm 2.1 is run with* $\eta_t = \min\{1, \frac{1}{\beta_t}\}$ *for* $t \in [T+1]$*, where* $\beta_1 = \gamma$ *and for* $t \in [T]$*,* $\beta_{t+1} = \beta_t + \frac{\gamma}{\sqrt{1 + \frac{1}{\log N}\sum_{s=1}^{t} H(p_t)}}$*. Then, it holds in general that*

$$R_T \leq 3\sqrt{2e\mathcal{C}(\Theta)\, T \log(eT)\log(eN)} + \log N\,,$$

whereas in the adversarially corrupted stochastic regime, the algorithm additionally satisfies

$$R_T \leq 36e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N + 4\Delta\,.$$

Proof. Let $\theta^* \in \arg\min_{\theta\in\Theta}\mathbb{E}\sum_t \ell_t(\theta)$, which need not coincide with ξ. For policy θ and round t, let $\widehat{\zeta}_t(\theta)$ be defined as in (2.2). As shown in the proof of Theorem 2.1,[5] we have that $R_T = \mathbb{E}\sum_t\langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle$, where $\mathbf{e}_{\theta^*}\in\mathbb{R}^N$ is the indicator vector for θ^*. Also, similar to what was argued in that proof, Algorithm 2.1 produces its predictions according to the following FTRL rule:

$$p_t = \operatorname*{arg\,min}_{p\in\Delta_N}\left\langle\sum_{s=1}^{t-1}\widehat{\zeta}_s, p\right\rangle + \phi_t(p)\,,$$

where for $p\in\Delta_N$ and $t\in[T+1]$, $\phi_t(p) = -\frac{1}{\eta_t}H(p)$. Note that the sequences β_t and η_t are increasing and non-increasing, respectively. Hence, we have that $\phi_t(p) - \phi_{t+1}(p)\geq 0$. With this in mind, one can extract the following bound from the proof of Theorem 2.1 and the proof of Lemma 7.14 in (Orabona, 2023):

$$\begin{aligned}\sum_t\langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle &\leq \phi_{T+1}(\mathbf{e}_{\theta^*}) - \phi_1(p_1) + \sum_t\big(\phi_t(p_{t+1}) - \phi_{t+1}(p_{t+1})\big)\\ &\qquad + \frac{e}{2}\sum_t\eta_t\sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\\ &= \frac{1}{\eta_1}\log N + \sum_t\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)H(p_{t+1}) + \frac{e}{2}\sum_t\eta_t\sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\,.\end{aligned}$$

Since, η_t is measurable with respect to $\mathcal{F}_{t-1}$, we have via (2.5) that $\mathbb{E}_t\eta_t\sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2 \leq \eta_t Q_{p_t}(\Theta)\leq\eta_t\mathcal{C}(\Theta)$. Consequently, it holds that

$$R_T \leq \frac{1}{\eta_1}\log N + \mathbb{E}\sum_t\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)H(p_{t+1}) + \frac{e}{2}\mathcal{C}(\Theta)\,\mathbb{E}\sum_t\eta_t\,.$$

For every round t, we clearly have that $\eta_t\leq\frac{1}{\beta_t}$, and that $\frac{1}{\eta_t} = \max\{1,\beta_t\}$. Hence, since $\beta_{t+1}\geq\beta_t$, it holds that $\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} = \big(\max\{1,\beta_{t+1}\} - \max\{1,\beta_t\}\big)\leq\beta_{t+1}-\beta_t$. Consequently,

$$R_T \leq \max\{1,\gamma\}\log N + \mathbb{E}\sum_t\big(\beta_{t+1}-\beta_t\big)H(p_{t+1}) + \frac{e}{2}\mathcal{C}(\Theta)\,\mathbb{E}\sum_t\frac{1}{\beta_t}\,. \tag{2.7}$$

[5] We at times treat p_t and $\widehat{\zeta}_t$ as vectors in $\mathbb{R}^N$ in the manner described before in the proof of Theorem 2.1.

The following two facts can be extracted from the proof of Proposition 1 in (Ito et al., 2022):

$$\sum_t (\beta_{t+1} - \beta_t) H(p_{t+1}) \le 2\gamma\sqrt{\log N}\sqrt{\sum_t H(p_t)}$$
$$\sum_t \frac{1}{\beta_t} \le \frac{\log(eT)}{\gamma\sqrt{\log N}}\sqrt{\log N + \sum_t H(p_t)}\,.$$

Moreover, Lemma 4 in the same paper entails that $\sum_t H(p_t) \le P(\xi)\log\frac{eNT}{P(\xi)}$. Plugging these inequalities back into (2.7) yields that

$$R_T \le \max\{1,\gamma\}\log N + 2\gamma\sqrt{\log N}\,\mathbb{E}\sqrt{P(\xi)\log\frac{eNT}{P(\xi)}} + \frac{e\mathcal{C}(\Theta)\log(eT)}{2\gamma\sqrt{\log N}}\,\mathbb{E}\sqrt{\log N + P(\xi)\log\frac{eNT}{P(\xi)}}\,. \tag{2.8}$$

To obtain the worst-case bound, simply observe that $u\log\frac{eNT}{u}$ is an increasing function in u for $0 < u \le NT$. Hence, $P(\xi)\log\frac{eNT}{P(\xi)} \le T\log(eN)$, which, together with (2.8), implies that

$$R_T \le \max\{1,\gamma\}\log N + \Big(2\gamma\sqrt{\log N} + \frac{e\mathcal{C}(\Theta)\log(eT)}{\gamma\sqrt{\log N}}\Big)\sqrt{T\log(eN)}\,,$$

from which the desired bound can be seen to hold after plugging in the value of γ.

Towards proving the second bound, we take an alternative route following the proof of Theorem 4 in (Ito et al., 2022). Namely, we argue that if $P(\xi) \le e$, then $P(\xi)\log\frac{eNT}{P(\xi)} \le e\log(NT)$, otherwise, $P(\xi)\log\frac{eNT}{P(\xi)} \le P(\xi)\log(NT)$. Resuming again from (2.8), we get that

$$\begin{aligned}
R_T &\le \max\{1,\gamma\}\log N + \Big(2\gamma\sqrt{\log N} + \frac{e\mathcal{C}(\Theta)\log(eT)}{\gamma\sqrt{\log N}}\Big)\sqrt{\log(NT)}\,\mathbb{E}\sqrt{\max\{P(\xi), e\}} \\
&\le \max\Big\{\log N, \sqrt{\frac{e}{2}\mathcal{C}(\Theta)\log(eT)\log(N)}\Big\} + 2\sqrt{2e\mathcal{C}(\Theta)\log(eT)\log(NT)}\Big(\sqrt{\overline{P}(\xi)} + \sqrt{e}\Big) \\
&\le \log N + 3e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)}\Big(1 + \sqrt{\overline{P}(\xi)}\Big)\,,
\end{aligned}$$

where the second inequality follows after plugging in the value of γ and using Jensen's inequality. For what follows, we define $G_1 := \log N$ and $G_2 := 3e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)}$.

Now, in the adversarially corrupted stochastic regime, we observe that

$$R_T \ge \sum_t \mathbb{E}\big[\ell_t(\theta) - \ell_t(\xi)\big] \ge \Delta\,\mathbb{E}\sum_t \mathbb{I}\{\vartheta_t \ne \xi\} - B = \Delta\overline{P}(\xi) - B\,.$$

Combining this with the last bound, we obtain that for any $\lambda > 0$,

$$R_T = (1+\lambda)R_T - \lambda R_T \le (1+\lambda)\Big(G_1 + G_2 + G_2\sqrt{\overline{P}(\xi)}\Big) - \lambda\Delta\overline{P}(\xi) + \lambda B\,.$$

Using the fact that $2a\sqrt{u} - bu \le a^2/b$ for all $u, a, b \ge 0$, we get that

$$\begin{aligned} R_T &\le (1+\lambda)(G_1+G_2) + \lambda B + \frac{(1+\lambda)^2 G_2^2}{4\lambda\Delta} \\ &= (1+\lambda)(G_1+G_2) + \frac{G_2^2}{2\Delta} + \lambda\Big(B + \frac{G_2^2}{4\Delta}\Big) + \frac{G_2^2}{4\lambda\Delta}. \end{aligned}$$

If we choose $\lambda := \sqrt{\frac{G_2^2}{4\Delta}\Big/\Big(B + \frac{G_2^2}{4\Delta}\Big)}$, we obtain the following bound:

$$R_T \le 2G_1 + 2G_2 + \frac{G_2^2}{2\Delta} + 2\sqrt{\frac{G_2^2 B}{4\Delta} + \frac{G_2^4}{16\Delta^2}} \le 2G_1 + 2G_2 + \frac{G_2^2}{\Delta} + \sqrt{\frac{G_2^2 B}{\Delta}},$$

where we have also used the fact that $\lambda \le 1$. Using the definitions of G_1 and G_2, we can conclude that

$$\begin{aligned} R_T &\le 18e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N \\ &\qquad + 6e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)} \\ &\le 36e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N + 4\Delta\,, \end{aligned}$$

where we used that $6e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)} \le \max\Big\{18e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta}, 4\Delta\Big\}$. □

The appeal of this theorem is that it shows that the simple and fundamental EXP4 algorithm can obtain logarithmic regret in stochastic environments, always scaling with the policy set capacity. Notice that we require the uniqueness of the optimal policy to achieve this, though similar assumptions are common in BOBW works adopting the so-called self-bounding technique applied in the last proof (Dann et al., 2023; Ito et al., 2022; Wei and Luo, 2018; Zimmert and Seldin, 2021). There is, however, small room for improvement in terms of the dependence of both bounds on the time horizon. Compared to Theorem 2.1, the adversarial bound shown above is worse off by an extra $\sqrt{\log T}$ factor. At the same time, the stochastic regime bound scales as $\log^2 T$ instead of the typical $\log T$ rate. Arguably, these shortcomings are mild for BOBW bounds, especially considering the simplicity of the algorithm.

Nevertheless, the recent work of Dann et al. (2023) offers one way for further honing these bounds. There, a general reduction scheme is proposed, allowing the automatic synthesis of BOBW algorithms starting from traditional algorithms satisfying a certain importance-weighting (iw) stability condition. More precisely, this condition requires that if the algorithm receives feedback in round t only with probability q_t (communicated at the start of the round), it achieves a gracefully degrading bound of $\mathbb{E}\left[\sqrt{c_1 \sum_{t\le t'} 1/q_t} + c_2 \max_{t\le t'} 1/q_t\right]$ on the expected regret at any stopping time t' with some constants c_1 and c_2. Let upd_t be an indicator for whether the feedback is received at round t. For bandits with expert advice (or contextual bandits), Lemma 10 in their paper shows that EXP4 is iw-stable with $c_1 = \mathcal{O}(K \log N)$ and $c_2 = 0$ by scaling the loss estimators with upd_t/q_t and using a simple adaptive learning rate. This leads to BOBW bounds depending on the number of actions K. For our setting, one can combine their analysis with that of Theorem 2.1 (similarly scaling the estimated shifted losses $\widehat{\zeta}_t$ with upd_t/q_t) yielding that EXP4 is iw-stable with $c_1 = \mathcal{O}(\mathcal{C}(\Theta)\log N)$ and $c_2 = \mathcal{O}(\log N)$ using the learning rate

$$\eta_t = \min\left\{\min_{s\le t} q_s, \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)\sum_{s\le t} 1/q_s}}\right\}.$$

Hence, Theorems 6 and 11 in (Dann et al., 2023) imply the existence of an algorithm enjoying a bound of $\mathcal{O}\left(\sqrt{\mathcal{C}(\Theta)T\log(N)} + \log(N)\log^2(T)\right)$ for the adversarial regime, and

$$\mathcal{O}\left(\frac{\mathcal{C}(\Theta)\log(T)\log(N)}{\Delta} + \sqrt{\frac{\mathcal{C}(\Theta)\log(T)\log(N)B}{\Delta}} + \log(N)\log(T)\log\left(\frac{B}{\Delta}\right)\right)$$

for the adversarially corrupted stochastic regime. These bounds deliver the sought improvements at the minor cost of scaling the trailing terms in both bounds with $\log T$ factors. On the downside, achieving these bounds requires an arguably laborious and contrived combination of EXP4 with two meta-algorithms.

In the stochastic regime, competing results in the literature mainly include a bound of order $(1 + d_{\chi^2}(\Theta))^2 \log T \log N/\Delta$ in (Sen et al., 2018, Corollary 2),[6] and a bound of order $\sqrt{(1 + d_{\chi^2}(\Theta))\, T\log(NT)}$ in (Papini et al., 2019, Theorem 2). These bounds fall short of this section's results, primarily considering their dependence on the structure of Θ. Another notable result, though incomparable, is the constant (time-independent) regret bound in (Metelli et al., 2021, Theorem 5.2), which nonetheless requires $d_{\chi^2}(\Theta)$ to be finite.

[6] To be precise, Sen et al. (2018) consider a stochastic version of the bandits with expert advice problem where an expert's recommendation is a function of an i.i.d. context. There, the diameter d_{χ^2} is defined with respect to the conditional chi-squared divergence integrated over the context distribution.

2.5 Lower Bounds

In this section, we complement the regret bounds provided thus far by proving lower bounds for the families of policy sets described in Section 2.3.3. More precisely, for a given policy set, we aim to prove a lower bound for the minimax regret $\inf_\pi \sup_{(\ell_t)_t} R_T$, where π is the player's strategy. To this end, we will consider a class of environments, each identified by a vector $\mu \in [0,1]^K$ such that, for an outcome x, the loss $\ell_t(x)$ at every round t is drawn from a Bernoulli distribution with mean $\mu(x)$ in an i.i.d. manner. For $t \le T$, recall that $\mathcal{H}_t := (\vartheta_s, X_s, \ell_s(X_s))_{s=1}^t$ denotes the interaction history till the end of round t. The player's strategy π can be represented as a sequence of probability kernels $\{\pi_t\}_{t=1}^T$, each mapping the history so far to a distribution over the policies such that ϑ_t is sampled from $\pi_t(\cdot \mid \mathcal{H}_{t-1})$. Hence, under environment μ, it holds that $\mathbb{P}(\vartheta_t = \cdot \mid \mathcal{H}_{t-1}) = \pi_t(\cdot \mid \mathcal{H}_{t-1})$, $\mathbb{P}(X_t = \cdot \mid \mathcal{H}_{t-1}, \vartheta_t) = \vartheta_t(\cdot)$, and $\mathbb{P}(\ell_t(X_t) = \cdot \mid \mathcal{H}_{t-1}, \vartheta_t, X_t) = p_{\mu,X_t}(\cdot)$, where $p_{\mu,x}$ is the loss distribution of outcome x under μ. Consequently, each environment μ (coupled with the player's strategy) induces a probability distribution P_μ on $\mathcal{H}_T$ such that

$$P_\mu\big((\xi_1, x_1, l_1, \ldots, \xi_T, x_T, l_T)\big) = \prod_{t=1}^T \pi_t(\xi_t \mid \xi_1, x_1, l_1, \ldots, \xi_{t-1}, x_{t-1}, l_{t-1}) \xi_t(x_t) p_{\mu,x_t}(l_t)\,,$$

for any $(\xi_1, x_1, l_1, \ldots, \xi_T, x_T, l_T) \in (\Theta \times \mathcal{X} \times \{0,1\})^T$. For an environment μ, we define the stochastic regret as:

$$\overline{R}_T(\mu) := \max_{\theta^* \in \Theta} \mathbb{E}_\mu \sum_{t=1}^T \sum_{x \in \mathcal{X}} (\vartheta_t(x) - \theta^*(x)) \mu(x)\,, \tag{2.9}$$

where the subscript in $\mathbb{E}_\mu$ emphasizes the dependence on P_μ. For any μ, $\overline{R}_T(\mu)$ is a lower bound for $\sup_{(\ell_t)_t} R_T$. Thus, to prove a lower bound on the minimax regret, it is sufficient to prove a lower bound on $\sup_\mu \overline{R}_T(\mu)$ that holds for any strategy of the player. In the sequel, we will make use of the following lemma, which provides an expression for the KL-divergence between the probability distributions induced by two environments.

Lemma 2.4. *For a fixed player's strategy, policy set, and time horizon, any two environments μ and μ' satisfy $D(P_\mu \,\|\, P_{\mu'}) = \sum_\theta N_\mu(\theta; T) \sum_x \theta(x) d\big(\mu(x) \,\big\|\, \mu'(x)\big)$, where $N_\mu(\theta; T) := \mathbb{E}_\mu \sum_{t=1}^T \mathbb{I}\{\vartheta_t = \theta\}$, and $d(a \,\|\, b)$ is the KL-divergence between two Bernoulli distributions with means a and b.*

Proof. Using the chain rule of the KL-divergence, one can obtain that

$$D(P_\mu \,\|\, P_{\mu'}) = \sum\nolimits_t \mathbb{E}_\mu D(p_{\mu,x_t} \,\|\, p_{\mu',x_t})\,.$$

We then use the tower rule and the linearity of expectation to conclude the proof:

$$\begin{aligned}\sum_t \mathbb{E}_\mu D(p_{\mu,x_t} \,\|\, p_{\mu',x_t}) &= \sum_t \mathbb{E}_\mu\big[\mathbb{E}_\mu\big[D(p_{\mu,x_t} \,\|\, p_{\mu',x_t}) \mid \vartheta_t\big]\big] \\ &= \sum_t \mathbb{E}_\mu \sum_x \vartheta_t(x) D(p_{\mu,x} \,\|\, p_{\mu',x}) \\ &= \sum_\theta N_\mu(\theta; T) \sum_x \theta(x) D(p_{\mu,x} \,\|\, p_{\mu',x})\,.\end{aligned}$$

□

2.5.1 The Two Policies Case

The following theorem provides a lower bound for the two policies case examined in Section 2.3.3. The proof mostly follows the needle-in-a-haystack technique of Auer et al. (2002a). The key to obtaining this result is a careful choice of the mean loss for each outcome. This choice leads to a lower bound in terms of the Hellinger squared distance between the two policies, which is then related to the capacity via (2.6). This shows that the bound of Theorem 2.1 is order-wise unimprovable for this case.

Theorem 2.5. *Assume that* $\Theta = \{\theta_1, \theta_2\}$. *Then, for any algorithm and* $T \geq \frac{1}{8\log(4/3) H^2(\theta_1,\theta_2)}$, *there exists a sequence of losses such that* $R_T \geq \frac{1}{13\sqrt{2}} \sqrt{C(\theta_1,\theta_2)\, T}$.

Proof. We will consider two environments μ_1 and μ_2 such that for an outcome x, we choose

$$\mu_1(x) := \frac{1}{2} - \Delta \frac{\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}} \qquad \text{and} \qquad \mu_2(x) := \frac{1}{2} - \Delta \frac{\sqrt{\theta_2(x)} - \sqrt{\theta_1(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}}\,,$$

where $0 \leq \Delta \leq \frac{1}{4}$ is to be tuned later. We posit that $\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}$ is always positive by assuming, without loss of generality, that each outcome is in the support of at least one policy. Additionally, let μ_0 be an environment such that $\mu_0(x) := 1/2$ for any outcome x. Note that θ_1 (θ_2) is the optimal policy in μ_1 (μ_2). Indeed,

$$\sum_x (\theta_2(x) - \theta_1(x)) \mu_1(x) = \Delta \sum_x \big(\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}\big)^2 = 2\Delta H^2(\theta_1,\theta_2) > 0\,.$$

Symmetrically, we have that $\sum_x(\theta_1(x)-\theta_2(x))\mu_2(x) = 2\Delta H^2(\theta_1,\theta_2)$. Hence, it holds that

$$\begin{aligned}\overline{R}_T(\mu_1) &= \mathbb{E}_{\mu_1}\sum\nolimits_t \mathbb{I}\{\vartheta_t=\theta_2\}\sum\nolimits_x(\theta_2(x)-\theta_1(x))\mu_1(x)\\ &= 2\Delta H^2(\theta_1,\theta_2)(T-N_{\mu_1}(\theta_1;T))\\ &\geq 2\Delta H^2(\theta_1,\theta_2)\Big(T-N_{\mu_0}(\theta_1;T)-T\sqrt{\frac{1}{2}D\big(P_{\mu_0}\,\|\,P_{\mu_1}\big)}\Big),\end{aligned} \tag{2.10}$$

where the inequality follows by using that $N_{\mu_1}(\theta_1;T)-N_{\mu_0}(\theta_1;T)\leq T\delta(P_{\mu_0},P_{\mu_1})$ followed by an application of Pinsker's inequality. Note that for $\Delta\leq 1/4$ and $c\coloneqq 8\log(4/3)$,

$$d(\mu_0(x)\,\|\,\mu_1(x)) = d\Big(\frac{1}{2}\,\Big\|\,\frac{1}{2}-\Delta\frac{\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)}+\sqrt{\theta_2(x)}}\Big)\leq c\Delta^2\left(\frac{\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)}+\sqrt{\theta_2(x)}}\right)^2.$$

We also have that

$$\sum\nolimits_x\theta_1(x)\left(\frac{\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)}+\sqrt{\theta_2(x)}}\right)^2\leq\sum\nolimits_x\big(\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}\big)^2 = 2H^2(\theta_1,\theta_2),$$

with an analogous inequality holding for θ_2. Combining Lemma 2.4 with these observations gets us that

$$\begin{aligned}&D\big(P_{\mu_0}\,\|\,P_{\mu_1}\big)\\ &\quad= N_{\mu_0}(\theta_1;T)\sum\nolimits_x\theta_1(x)d\big(\mu_0(x)\,\|\,\mu_1(x)\big)+N_{\mu_0}(\theta_2;T)\sum\nolimits_x\theta_2(x)d\big(\mu_0(x)\,\|\,\mu_1(x)\big)\\ &\quad\leq 2c\Delta^2H^2(\theta_1,\theta_2)(N_{\mu_0}(\theta_1;T)+N_{\mu_0}(\theta_2;T)) = 2c\Delta^2H^2(\theta_1,\theta_2)T.\end{aligned}$$

Plugging back into (2.10) yields that

$$\overline{R}_T(\mu_1)\geq 2\Delta H^2(\theta_1,\theta_2)\Big(T-N_{\mu_0}(\theta_1;T)-T\Delta\sqrt{cH^2(\theta_1,\theta_2)T}\Big).$$

An analogous bound can be similarly shown to hold for environment μ_2. Hence, we can proceed by arguing that

$$\sup_\mu\overline{R}_T(\mu)\geq\frac{1}{2}(\overline{R}_T(\mu_1)+\overline{R}_T(\mu_2))\geq\Delta H^2(\theta_1,\theta_2)T\left(1-2\Delta\sqrt{cH^2(\theta_1,\theta_2)T}\right).$$

The theorem then follows by setting $\Delta\coloneqq\frac{1}{4\sqrt{cH^2(\theta_1,\theta_2)T}}$ and using that (see Section 2.3.3) $2H^2(\theta_1,\theta_2)\geq\mathcal{C}(\theta_1,\theta_2)$. Note that the stated condition on T ensures that $\Delta\leq 1/4$. □

2.5.2 ε-Greedy Policies

Next, we prove a lower bound for the ε-greedy case discussed in Section 2.3.3. Recall that for this case, $\mathcal{C}(\Theta) = \varepsilon^2(N-1)$; hence, the following lower bound matches the bound of Theorem 2.1 up to a logarithmic factor. Further, we can conclude from this result that for any $g \in [0, N-1]$, there exists a policy set Θ with $|\Theta| = N$ and $\mathcal{C}(\Theta) = g$ for which one has to incur regret of order at least $\sqrt{\mathcal{C}(\Theta)T}$.

Theorem 2.6. *Assume that the policy set conforms to the ε-greedy structure. Then, for $T \geq \frac{N}{4\log(4/3)}$ and any algorithm, there exists a sequence of losses such that $R_T \geq \frac{1}{18}\varepsilon\sqrt{NT}$.*

Proof. We will consider N environments $\{\mu_\theta\}_{\theta\in\Theta}$ such that for environment μ_θ and outcome x, $\mu_\theta(x) \coloneqq 1/2 - \Delta\mathbb{I}\{x = x_\theta\}$, where $0 \leq \Delta \leq 1/4$ is to be tuned later. Additionally, let μ_0 be an environment such that $\mu_0(x) \coloneqq 1/2$ for any outcome x. Notice that θ is the optimal policy in environment μ_θ. In particular, for $\theta' \in \Theta \setminus \{\theta\}$, we have that

$$\textstyle\sum_x (\theta'(x) - \theta(x))\mu_\theta(x) = \Delta(\theta(x_\theta) - \theta'(x_\theta)) = \Delta\varepsilon\,.$$

Thus,

$$\overline{R}_T(\mu_\theta) = \Delta\varepsilon(T - N_{\mu_\theta}(\theta;T)) \geq \Delta\varepsilon\Big(T - N_{\mu_0}(\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_0}\,\big\|\,P_{\mu_\theta}\big)}\Big), \tag{2.11}$$

where the inequality follows by using that $N_{\mu_\theta}(\theta;T) - N_{\mu_0}(\theta;T) \leq T\delta(P_{\mu_0}, P_{\mu_\theta})$ followed by an application of Pinsker's inequality. Starting from Lemma 2.4, we have that

$$\begin{aligned} D\big(P_{\mu_0}\,\big\|\,P_{\mu_\theta}\big) &= \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T) \sum_x \theta'(x)\, d\big(\mu_0(x)\,\big\|\,\mu_\theta(x)\big) \\ &= \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T)\theta'(x_\theta)\, d\Big(\frac{1}{2}\,\Big\|\,\frac{1}{2} - \Delta\Big) \\ &\leq c\Delta^2 \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T)\theta'(x_\theta) = c\Delta^2\left(\frac{1-\varepsilon}{N}T + \varepsilon N_{\mu_0}(\theta;T)\right), \end{aligned}$$

where the inequality holds for $\Delta \leq 1/4$ with $c \coloneqq 8\log(4/3)$. Plugging this result back into (2.11) allows us to conclude that

$$\begin{aligned} \sup_\mu \overline{R}_T(\mu) &\geq \frac{1}{N}\textstyle\sum_\theta \overline{R}_T(\mu_\theta) \\ &\geq \Delta\varepsilon\left(T - \frac{T}{N} - T\sqrt{\frac{c}{2}\Delta^2\left(\frac{1-\varepsilon}{N}T + \varepsilon\frac{T}{N}\right)}\right) \geq \Delta\varepsilon T\left(\frac{1}{2} - \Delta\sqrt{\frac{c}{2}\frac{T}{N}}\right), \end{aligned}$$

where the second inequality uses the concavity of the square root, and the third holds since $N \geq 2$. The theorem then follows by setting $\Delta := \frac{1}{4}\sqrt{\frac{2N}{cT}}$ and verifying that the stated condition on T ensures that $\Delta \leq 1/4$. □

2.5.3 The Multitask Bandits Structure

In this section, we prove a lower bound for a certain structure belonging to the family of M-supported uniform policies described in Section 2.3.3. Let M be a positive integer such that $q := K/M$ is an integer greater than or equal to 2. In this structure, the outcomes are divided into M sections, and each policy is a uniform distribution supported over M outcomes such that its support contains an outcome from each section. Assuming that the policy set contains all such policies, we have that $N = (K/M)^M$. For this particular structure, we will index the outcomes according to the section they belong to and their order therein: $\mathcal{X} = \{x_{i,j} : i \in [M], j \in [q]\}$. With this notation, we can describe the policy set as

$$\Theta = \left\{\theta \in \mathcal{U}_{K,M} : \forall i \in [M], \sum_{j=1}^{q} \theta(x_{i,j}) = \frac{1}{M}\right\},$$

where $\mathcal{U}_{K,M}$ is the set of all M-supported uniform distributions over K outcomes. Seeing the outcomes in one section as arms in a bandit game, this problem is, in a sense, equivalent to playing M bandit games simultaneously, with the choice of policy at each round dictating an arm choice for each game. The distinction is that only the loss incurred in a single randomly sampled game is observed, while the player nonetheless aims at minimizing their regret averaged over the M games. This type of structure (albeit with a different type of feedback) is commonly used to prove lower bounds for combinatorial bandits (see, e.g., Audibert et al., 2014). The following theorem, proved in Appendix A.1, provides a lower bound for our setting.

Theorem 2.7. *Suppose the policy set conforms to the multi-task structure. Then, for any algorithm and $T \geq \frac{K}{4\log(4/3)}$, there exists a sequence of losses such that $R_T \geq \frac{1}{18}\sqrt{KT}$.*

Recall from Section 2.3.3 that for M-supported uniform policies, $C(\Theta) = K/M - 1$. Also, note that $M = \log(N)/\log(K/M)$. Thus, the bound given by the theorem is of order $\sqrt{C(\Theta)T\log(N)/\log(C(\Theta)+1)}$. The distinguishing value of this lower bound is that it shows that the logarithmic dependence on the number of policies in the bound of Theorem 2.1 is not entirely spurious and that it becomes increasingly tight as $C(\Theta)$ decreases. This result can be seen as an analogue for our setting of the $\sqrt{KT\log(N)/\log(K)}$ lower bound proved by Seldin and Lugosi (2016) for the problem of bandits with expert advice, noting

that the construction of their bound relies on a (time-varying) sequence of deterministic expert recommendations.

2.6 An Impossibility Result Under Linear Bandit Feedback

In this section, we establish a separation between the mediator feedback model and the linear bandit model in terms of achievable regret. With $\Theta \subset \Delta_K$ as the action set, we consider a linear bandit problem where upon choosing a policy ϑ_t in round t, the learner directly observes $\ell_t(\vartheta_t) = \sum_x \vartheta_t(x)\ell_t(x)$. This is in contrast to the setting considered thus far, where the learner observes $(X_t, \ell_t(X_t))$ with X_t sampled from the distribution of ϑ_t. The notion of regret we aim to minimize in the linear bandit variant remains the same as before. The main message conveyed by this section's results is that obtaining information regarding individual outcomes is crucial for achieving regret guarantees that reflect the affinity of the policies' distributions.

Concretely, we will consider once again the ε-greedy structure described in Section 2.3.3. Similar to the previous section, we will rely on a class of environments, each identified by a vector $\mu \in [0,1]^K$. However, we will adopt a different scheme for generating the losses for the outcomes. For every round t, let Z_t be a random variable drawn in an i.i.d. manner from a normal distribution $\mathcal{N}(0,\sigma^2)$, with some $\sigma > 0$. Accordingly, for each outcome x, we set $\ell_t(x) \coloneqq \mu(x) + Z_t$. Hence, the losses assigned to the outcomes in a given round are correlated, see (Cohen et al., 2017) for a similar approach in the combinatorial bandit problem. It follows then that $\ell_t(\vartheta_t) = \langle \vartheta_t, \mu \rangle + Z_t$. In this section, we let P_μ denote the distribution induced by μ (and the player's strategy) over the interaction history in this variant, namely $(\vartheta_1, \ell_1(\vartheta_1), \dots, \vartheta_T, \ell_T(\vartheta_T))$. We will again study the stochastic regret $\overline{R}_T(\mu)$, still defined as in (2.9).

For the ε-greedy decision set, Theorem 22.1 in (Lattimore and Szepesvári, 2020) implies the existence of an algorithm enjoying a bound of order $\sqrt{NT\log(NT)}$ on the stochastic regret (when $\sigma = 1$), recalling that the members of Θ in this case are N-dimensional vectors. While for the adversarial regret, an upper bound of order $\sqrt{NT\log(N)}$ is achievable, see (Bubeck et al., 2012). The following two results show that, up to factors logarithmic in N and T, the cited bounds are unimprovable, no matter the value of ε. Hence, the growing similarity between the actions—or the shrinking diameter of Θ—as ε approaches 0 cannot be exploited. This is in sharp contrast to the mediator feedback setting, where Theorems 2.1 and 2.6 establish the minimax regret to be essentially of order $\varepsilon\sqrt{NT}$ for this family.

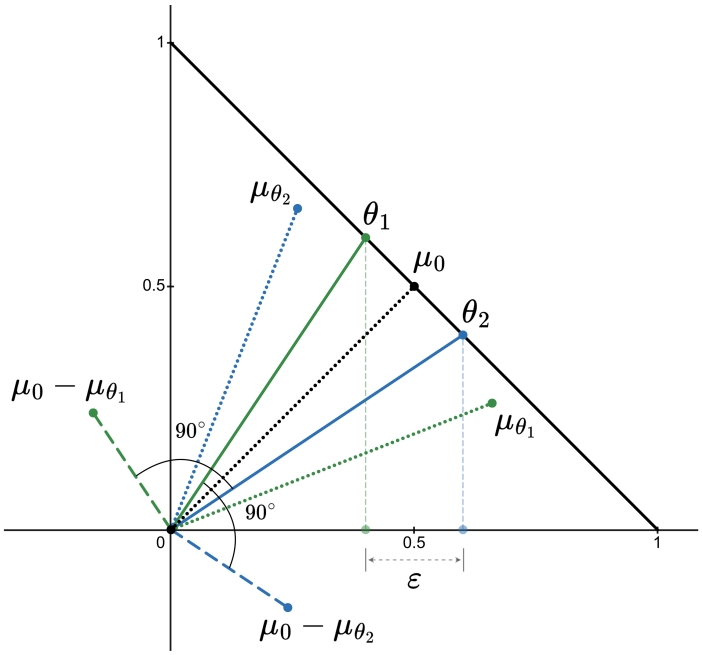

Fig. 2.1 This figure provides an illustration of the construction we use to prove the lower bound of Proposition 2.8 in the case when $N = 2$ (and thus, $K = 2$, as per the setup of the ε-greedy structure). Notice that μ_{θ_1} is designed such that $\langle \mu_0 - \mu_{\theta_1}, \theta_2 \rangle = 0$, meaning that one cannot distinguish between the environments μ_{θ_1} and μ_0 by playing θ_2. The same holds when the roles of θ_1 and θ_2 are exchanged. We mirror in this way the classic lower bound construction of Auer et al. (2002a) in the standard bandit problem, and indeed essentially recover their bound, no matter how similar the policies are.

Proposition 2.8. *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, for the class of linear bandit environments described above (with any given $\sigma > 0$), it holds for any algorithm and $T \geq \frac{\sigma^2 N}{\varepsilon^2}$ that* $\sup_\mu \overline{R}_T(\mu) \geq \frac{\sigma}{8}\sqrt{NT}$.

Proof. We will consider N environments $\{\mu_\theta\}_{\theta \in \Theta}$ such that for environment μ_θ and outcome x, we set

$$\mu_\theta(x) := \frac{1}{2} + \Delta\left(\frac{1-\varepsilon}{N} - \mathbb{I}\{x = x_\theta\}\right),$$

where $0 \leq \Delta \leq 1/2$ is to be tuned later. Thus, under environment μ_θ, we have that

$$\ell_t(\vartheta_t) = \sum_x \vartheta_t(x)\ell_t(x) = \frac{1}{2} + \Delta\left(\frac{1-\varepsilon}{N} - \vartheta_t(x_\theta)\right) + Z_t = \frac{1}{2} - \Delta\varepsilon\mathbb{I}\{\vartheta_t = \theta\} + Z_t\,. \tag{2.12}$$

Additionally, let μ_0 be an environment such that $\mu_0(x) = 1/2$ for any outcome x, implying that $\ell_t(\vartheta_t) = 1/2 + Z_t$. Notice that θ is the optimal policy in environment μ_θ. In particular,

for $\theta' \in \Theta \setminus \{\theta\}$, we have that

$$\sum_x (\theta'(x) - \theta(x))\mu_\theta(x) = \Delta(\theta(x_\theta) - \theta'(x_\theta)) = \Delta\varepsilon\,.$$

(See Figure 2.1 for a graphical illustration of this construction.) Hence, it holds that

$$\overline{R}_T(\mu_\theta) = \Delta\varepsilon(T - N_{\mu_\theta}(\theta;T)) \geq \Delta\varepsilon\Big(T - N_{\mu_0}(\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_0} \,\|\, P_{\mu_\theta}\big)}\Big)\,,$$

where the inequality follows by using that $N_{\mu_\theta}(\theta;T) - N_{\mu_0}(\theta;T) \leq T\delta(P_{\mu_0}, P_{\mu_\theta})$ followed by an application of Pinsker's inequality. Combining the observation in (2.12) with standard results, see, for example, Exercise 15.8 (b) and Exercise 14.7 in (Lattimore and Szepesvári, 2020); we can express the KL-divergence term as follows:

$$\begin{aligned} D\big(P_{\mu_0} \,\|\, P_{\mu_\theta}\big) &= \sum_{\theta'} N_{\mu_0}(\theta';T) D\big(\mathcal{N}(1/2,\sigma^2) \,\|\, \mathcal{N}(1/2 - \Delta\varepsilon\mathbb{I}\{\theta' = \theta\}, \sigma^2)\big) \\ &= N_{\mu_0}(\theta;T) D\big(\mathcal{N}(1/2,\sigma^2) \,\|\, \mathcal{N}(1/2 - \Delta\varepsilon, \sigma^2)\big) = \frac{\Delta^2\varepsilon^2}{2\sigma^2} N_{\mu_0}(\theta;T)\,. \end{aligned}$$

Consequently, we get that

$$\sup_\mu \overline{R}_T(\mu) \geq \frac{1}{N}\sum_\theta \overline{R}_T(\mu_\theta) \geq \Delta\varepsilon\left(T - \frac{T}{N} - \frac{\Delta\varepsilon}{2\sigma}T\sqrt{\frac{T}{N}}\right) \geq \Delta\varepsilon T\left(\frac{1}{2} - \frac{\Delta\varepsilon}{2\sigma}\sqrt{\frac{T}{N}}\right),$$

where the second inequality holds by the concavity of the square root, and the third since $N \geq 2$. The proposition then follows by choosing $\Delta := \frac{\sigma}{2\varepsilon}\sqrt{\frac{N}{T}}$. Note that the stated condition on T ensures that $\Delta \leq 1/2$. □

As the construction of this lower bound relied on normally distributed (hence unbounded) losses, a lower bound in the adversarial setting is not immediately implied. Instead, the following theorem (proved in Appendix A.2) provides the sought bound at the cost of an extra $1/\sqrt{\log(T)}$ factor by combining Proposition 2.8 with a simple truncation argument due to Cohen et al. (2017). Notice that the resulting bound can be made arbitrarily larger than the mediator feedback guarantee of Theorem 2.1 by picking a small enough ε and a suitably long horizon.

Theorem 2.9. *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, under linear bandit feedback, we have that for any algorithm and $T \geq \frac{N}{8\varepsilon^2}$, there exists a sequence of losses (bounded in $[0,1]$) such that $R_T \geq \frac{1}{64\sqrt{2\log(16T)}}\sqrt{NT}$.*

Algorithm 2.3 OMD on the convex hull of the policies under full-information

input: learning rate η, initial distribution $\alpha^* \in \mathrm{co}(\Theta) : \alpha^*(x) > 0 \; \forall x \in \mathcal{X}$
initialize: $u_1 = \alpha^*$
for $t = 1, \dots, T$ **do**
    pick distribution $p_t \in \mathcal{P}_\Theta$ such that $\sum_\theta p_t(\theta)\theta = u_t$
    draw $\vartheta_t \sim p_t$
    observe the loss vector ℓ_t
    set $u_{t+1} = \arg\min_{u \in \mathrm{co}(\Theta)} \eta \langle u, \ell_t \rangle + D(u \,\|\, u_t)$
end for

2.7 The Full-Information Case

In this last section, we briefly examine a full-information variant of the problem, where the entire loss map $(\ell_t(x))_{x \in \mathcal{X}}$ is observed at every round. One can see this as a variant of the classical prediction with expert advice problem (Cesa-Bianchi et al., 1997), with the outcomes representing the actions (or experts). The distinction is that the learner can only sample from a mixture of the distributions of the policies in Θ. Moreover, the learner competes with the best policy, aiming to minimize the same notion of regret as before. We show in the following theorem that a simple strategy enjoys a regret guarantee depending on the more standard notion of capacity based on the KL-divergence. Precisely, if ϑ and X are two random variables taking values respectively over Θ and $\mathcal{X}$ such that $\mathbb{P}_{X|\vartheta=\theta}(x) = \theta(x)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. Then, we define the KL-capacity of the policy set as:

$$\mathcal{C}_{\mathrm{KL}}(\Theta) := \mathcal{C}_{\mathrm{KL}}(\mathbb{P}_{X|\vartheta}) = \max_{\tau \in \mathcal{P}_\Theta} \sum_\theta \tau(\theta) D\Big(\theta \,\Big\|\, \textstyle\sum_{\theta'} \tau(\theta')\theta'\Big).$$

Via (Polyanskiy and Wu, 2023, Corollary 5.6), $\mathcal{C}_{\mathrm{KL}}(\Theta)$ can alternatively be interpreted as the "radius" of Θ in terms of the KL-divergence:

$$\mathcal{C}_{\mathrm{KL}}(\Theta) = \min_{\alpha \in \Delta_K} \max_{\theta \in \Theta} D(\theta \,\|\, \alpha). \tag{2.13}$$

The idea of Algorithm 2.3 is to run an Online Mirror Descent (OMD) algorithm directly on the outcome space restricted to the convex hull of the policy set, henceforth denoted as $\mathrm{co}(\Theta)$. The key to obtaining the following result is a tailored choice of the initial distribution that utilizes the interpretation in (2.13).

Theorem 2.10. *Algorithm 2.3 run with* $\alpha^* \in \arg\min_{\alpha \in \Delta_K} \max_{\theta \in \Theta} D(\theta \,\|\, \alpha)$ *and* $\eta = \sqrt{\frac{2\mathcal{C}_{\mathrm{KL}}(\Theta)}{T}}$ *satisfies* $R_T \le \sqrt{2\mathcal{C}_{\mathrm{KL}}(\Theta)\,T}$.

Proof. Firstly, we note that setting $u_1 = \alpha^*$ is a valid choice since the minimum value of $\max_{\theta\in\Theta} D(\theta \,\|\, \alpha)$ in $\alpha \in \Delta_K$ can only be attained in $\mathrm{co}(\Theta)$; see Theorem 11.6.1 in Cover and Thomas (2006). Let $\theta^* \in \arg\min_{\theta\in\Theta} \mathbb{E}\sum_t \ell_t(\theta)$. As ℓ_t and ϑ_t are independent given the events up to the end round $t-1$, we have that $R_T = \mathbb{E}\sum_t \langle \vartheta_t - \theta^*, \ell_t\rangle = \mathbb{E}\sum_t \langle u_t - \theta^*, \ell_t\rangle$. A standard regret bound for OMD with the negative entropy regularizer (see, e.g., Lemma 6.16 and the proof of Theorem 10.2 in Orabona, 2023) allows us to conclude that

$$\sum_t \langle u_t - \theta^*, \ell_t\rangle \le \frac{D(\theta^* \,\|\, \alpha^*)}{\eta} + \frac{\eta}{2}\sum_t \sum_x u_t(x)\ell_t^2(x) \le \frac{D(\theta^* \,\|\, \alpha^*)}{\eta} + \frac{\eta}{2}T \le \frac{\mathcal{C}_{\mathrm{KL}}(\Theta)}{\eta} + \frac{\eta}{2}T,$$

where the last step follows from (2.13) and the definition of α^*. The theorem then follows after plugging in the value of η. □

We remark that $\mathcal{C}_{\mathrm{KL}}(\Theta) \le \min\{\log N, \log K\}$, see Section 2.2.1. In particular, the first bound is attained when the policies have non-overlapping supports, while the second is attained when each outcome is matched with a policy entirely concentrated on that outcome. Notice that these two cases essentially reduce the problem to a standard prediction with expert advice problem on the policy and outcome spaces, respectively, where the bound of Theorem 2.10 matches the minimax regret up to constants (Cesa-Bianchi et al., 1997). Beyond these extreme cases, the bound improves as the capacity of the policy set—or its information radius—shrinks.

2.8 Conclusions and Future Directions

In this chapter, we have focused on the mediator feedback framework and studied to what extent the structure of the problem can be exploited by the learner. In particular, we have introduced the policy set capacity as a measure of the effective size (or complexity) of the policy set. For various setups, we have derived new and improved regret bounds for EXP4 featuring the capacity. Further, the lower bounds we provided establish the capacity as a fundamental indicator of the difficulty of the problem for a rich collection of policy sets. Ultimately, we leave open the study of the optimality of the capacity on a more fine-grained level; specifically, whether a regret lower bound in terms of the capacity can be shown to hold for *any* given policy set.

Another direction for improvement is providing bounds that hold with high probability rather than in expectation, noting that prior works on bandits with expert advice (or contextual bandits) such as (Beygelzimer et al., 2011) and (Neu, 2015b) obtained high prob-

ability bounds only of order $\sqrt{KT\log N}$. Yet another direction is proving data-dependent (or small loss) bounds, following again previous works on bandits with expert advice (e.g., Allen-Zhu et al., 2018). Concerning the stochastic regime, improving the dependence of the bounds on the sub-optimality gaps (beyond the crude scaling with the smallest gap) is another interesting problem. Additionally, for cases when the policy set is very large, achieving similar regret guarantees via more computationally efficient strategies is a worthy direction. Finally, save for the discussion regarding the lower bound of Theorem 2.7, we have not touched upon the question of whether the $\sqrt{\log N}$ factor featured in the upper bounds can be improved. This will be the topic of Chapter 4, which will build upon tools and results from this chapter and the next.

Chapter 3

On the Minimax Regret for Online Learning with Feedback Graphs*

3.1 Introduction

The problem of bandits with graph feedback, first proposed by Mannor and Shamir (2011), provides an elegant interpolation between the classical problems of multi-armed bandits and prediction with expert advice. Given an undirected feedback graph G over the K actions of a bandit problem, the learner observes not only the loss of the action chosen in each round, but also the loss of the actions that are adjacent to it in the graph. The standard bandit problem corresponds to the case when G has no edges, whereas prediction with expert advice corresponds to the other extreme where G is a clique. When losses are generated adversarially, the regret in this setting has been shown to scale with the independence number α of G. Intuitively, denser graphs, which correspond to smaller independence numbers, provide more feedback to the learner, thus enabling a better control on the regret.

More specifically, the best known upper and lower bounds on the regret after T rounds are $\mathcal{O}\big(\sqrt{\alpha T \log K}\big)$ and $\Omega\big(\sqrt{\alpha T}\big)$ (Alon et al., 2013, 2017). It has been known for three decades that this upper bound is tight when $\alpha = 1$ as this corresponds to the experts case, see (Cesa-Bianchi et al., 1993, 1997). When $\alpha = K$, we are in the standard bandit case, where the lower bound $\Omega\big(\sqrt{KT}\big)$—which has also been known for nearly three decades (Auer et al., 1995, 2002a)—was matched by a corresponding upper bound $\mathcal{O}\big(\sqrt{KT}\big)$ only in 2009 by Audibert and Bubeck. These results show that in feedback graphs, the logarithmic

*This chapter is based on (Eldowa, Esposito, Cesari, and Cesa-Bianchi, 2023b).

factor $\sqrt{\log K}$ is necessary (at least) for the $\alpha = 1$ case, while it must vanish from the minimax regret as α grows from 1 to K, but the current bounds fail to capture this fact. In this chapter, we prove new upper and lower regret bounds that for the first time account for this vanishing logarithmic factor.

To prove our new upper bound, we use the standard FTRL algorithm run with the q-Tsallis entropy regularizer (q-FTRL for short). It is well-known (Abernethy et al., 2015) that for $q = 1/2$ this algorithm (run with appropriate loss estimates) achieves regret $\mathcal{O}\big(\sqrt{KT}\big)$ when $\alpha = K$ (bandits case), while for $q \to 1^-$ the same algorithm (without loss estimates) recovers the bound $\mathcal{O}\big(\sqrt{T\log K}\big)$ when $\alpha = 1$ (experts case). We show in Theorem 3.2 that, if q is chosen as a certain function $q(\alpha, K)$, then $q(\alpha, K)$-FTRL, run with standard importance-weighted loss estimates, achieves regret $\mathcal{O}\big(\sqrt{\alpha T(1+\log(K/\alpha))}\big)$. This is a strict improvement over the previous bound, and matches the lower bounds for bandits and experts while interpolating between them in the intermediate cases. This interpolation is reflected by our choice of q, which goes from $1/2$ to 1 as α ranges from 1 to K. The main technical hurdle in proving this result is an extension to arbitrary values of $q \in [1/2, 1)$ of a standard result (see, e.g., Mannor and Shamir, 2011, Lemma 3) that bounds in terms of α the variance term in the regret of q-FTRL.

In Theorem 3.3, using a modified loss estimate, this result is extended to the strongly observable (undirected) feedback graphs setting (Alon et al., 2015), a slightly more general class of online learning problems with graph-structured feedback where some of the actions may not reveal their loss when chosen by the learner. In Theorem 3.4, we show via a doubling trick that our new upper bound can also be obtained (up to constant factors) without the need of knowing (or computing) α. As the resulting algorithm is oblivious to α, our analysis also applies to arbitrary sequences of graphs G_t, where K is constant but the independence number α_t of G_t can change over time, and the algorithm observes G_t only after choosing an action (the so-called uninformed case). In this setting, the analysis of the doubling trick is complicated by the non-trivial dependence of the regret on the α_t sequence.

We also improve on the $\Omega\big(\sqrt{\alpha T}\big)$ lower bound by proving a new $\Omega\big(\sqrt{\alpha T\log_\alpha K}\big)$ lower bound for all $\alpha > 1$. This is the first result showing the necessity—outside of the experts case—of a logarithmic factor in the minimax regret for all $\alpha < K$. Our proof uses a stochastic adversary generating both losses and feedback graphs via i.i.d. draws from a joint distribution. This sequence of losses and feedback graphs can be used to define a hard instance of the multi-task bandit problem, a variant of the combinatorial bandits framework (Cesa-Bianchi and Lugosi, 2012). We then prove our result by adapting known

lower bounding techniques for multi-task bandits (Audibert et al., 2014). In fact, the construction used to prove this lower bound is essentially the same as that used to prove the lower bound of Theorem 2.7 in the mediator feedback chapter, where we did rely on the same multi-task reduction idea. Note that for values of α bounded away from 2 and K, the logarithmic factor $\log_\alpha K$ in the lower bound is smaller than the corresponding factor $1+\log(K/\alpha)$ in the upper bound. Following the publication of these results, this gap has been closed (from the lower bound's side) for a certain class of graphs in a subsequent work by Chen et al. (2024), we refer to Section 3.7 at the end of this chapter for more details.

3.1.1 Additional related work

Several previous works have used the q-Tsallis regularizer with q tuned to specific values other than $1/2$ and 1. For example, in (Zimmert and Lattimore, 2019, Section 4), q is chosen as a function of K to prove a regret bound of $\mathcal{O}\big(\sqrt{\alpha T(\log K)^3}\big)$ for any strongly observable directed feedback graph, which shaves off a $\log T$ factor compared to previous works. This bound is worse than the corresponding bounds for undirected graphs because the directed setting is harder. Specific choices of q have been considered to improve the regret in settings of online learning with standard bandit feedback. For example, the choice $q = 2/3$ was used in (Rouyer and Seldin, 2020) to improve the analysis of regret in bandits with decoupled exploration and exploitation. Regret bounds for arbitrary choices of q are derived in (Zimmert and Seldin, 2021; Jin et al., 2023) for a best-of-both-worlds analysis of bandits, though $q = 1/2$ remains the optimal choice. The $1/2$-Tsallis entropy and the Shannon entropy ($q = 1$) regularizers have been combined before in different ways to obtain best-of-both-worlds guarantees for the graph feedback problem (Erez and Koren, 2021; Ito et al., 2022).

The idea of using values of $q \in (1/2, 1)$ for feedback graphs is quite natural and has been brought up before, e.g., by Rouyer et al. (2022), but achieving an improved dependence on the graph structure by picking a suitable value of q has not been, to the best of our knowledge, successfully pursued before. On the other hand, an approach based on a similar use of the q-Tsallis regularizer (though presented through the Poly-INF framework of Audibert and Bubeck, 2009, 2010) has been employed by Kale (2014) for a generalization of the bandits with expert advice problem, where a bound of order $\sqrt{\min\{K,N\}T(1+\log(N/\min\{K,N\}))}$ (with N being the number of experts) is achieved for the standard version of the problem. Additionally, Kwon and Perchet (2016) use a similar regularization scheme for the problem of bandits with sparse losses to achieve a

regret bound of order $\sqrt{sT\log(K/s)}$, where s is the maximum number of nonzero losses at any round.

In addition to its similarity (in derivation and form) to the lower bound of Theorem 2.7 in the mediator feedback setting, our lower bound is also reminiscent of the $\Omega\big(\sqrt{KT\log_K N}\big)$ lower bound proved by Seldin and Lugosi (2016) for the related bandits with expert advice problem (with $N \geq K$). We also refer to (Vural et al., 2019) for a similar lower bound in a related setting.

3.2 Problem Setting

We consider the following game played over T rounds between a learner with action set $V := [K]$ and the environment. At the beginning of the game, the environment secretly selects a sequence of losses $(\ell_t)_{t\in[T]}$, where $\ell_t \colon V \to [0,1]$,[1] and a sequence of undirected graphs $(G_t)_{t\in[T]}$ over the set of actions V, that is, $G_t := (V, E_t)$. At any time t, the learner selects an arm I_t (possibly at random), then pays loss $\ell_t(I_t)$ and observes the feedback graph G_t and all losses $\ell_t(i)$ of neighbouring actions $i \in N_{G_t}(I_t)$, where $N_{G_t}(i) := \{j \in V : (i,j) \in E_t\}$ (see the interaction protocol below). We only focus on strongly observable graphs (Alon et al., 2015); an undirected graph G is said to be strongly observable if for every $i \in V$, at least one of the following holds: $i \in N_G(i)$ or $i \in N_G(j)$ for all $j \neq i$.

The performance of the learner is measured by the regret

$$R_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t(i)\,.$$

where the expectation is over the learner's internal randomization.

Interaction Protocol Online learning with feedback graphs

environment: (hidden) losses $\ell_t \colon V \to [0,1]$ and graphs $G_t = (V, E_t)$, for all $t = 1, \ldots, T$
for $t = 1, \ldots, T$ **do**
  The learner picks an action $I_t \in V$ (possibly at random)
  The learner incurs loss $\ell_t(I_t)$
  The learner observes losses $\big\{\big(i, \ell_t(i)\big) : i \in N_{G_t}(I_t)\big\}$ and graph G_t
end for

[1] As in the previous chapter, we will sometimes treat the loss functions $\ell_t \colon V \to [0,1]$ as vectors with components $\ell_t(1), \ldots, \ell_t(K)$.

As usual, we denote by Δ_K the probability simplex in $\mathbb{R}^K$. For any graph G, we define its independence number as the cardinality of the largest set of nodes such that no two nodes are neighbors, and denote it by $\alpha(G)$. For simplicity, we use N_t to denote the neighbourhood N_{G_t} in the graph G_t and we use α_t to denote the independence number $\alpha(G_t)$ of G_t at time t.

3.3 FTRL with Tsallis Entropy Regularization for Feedback Graphs

As a building block, in this section, we focus on the case when all the feedback graphs $G_1, \dots, G_T$ have the same independence number $\alpha_1 = \cdots = \alpha_T = \alpha$, whereas the general case is treated in the next section. For simplicity, we start with the assumption that all nodes have self-loops: $(i,i) \in E_t$ for all $i \in V$ and all t. We later lift this requirement and show that the regret guarantees that we provide can be extended to general strongly observable undirected feedback graphs, only at the cost of a constant multiplicative factor.

The algorithm we analyze is q-FTRL (described in Algorithm 3.1), which is an instance of the FTRL framework with the (negative) q-Tsallis entropy

$$\psi_q(x) \coloneqq \frac{1}{1-q}\left(1 - \sum_{i \in V} x(i)^q\right) \qquad \forall x \in \Delta_K,$$

as the regularizer, whose parameter $q \in (0,1)$ can be tuned according to our needs. Since we do not observe all the losses in a given round, the algorithm makes use of unbiased estimates for the losses. When all self-loops are present, we define the estimated losses in the following standard manner. Let I_t be the action picked at round t, which is drawn from the distribution $p_t \in \Delta_K$ maintained by the algorithm, the loss estimate for an action $i \in V$ at round t is given by

$$\widehat{\ell}_t(i) \coloneqq \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\{I_t \in N_t(i)\}, \tag{3.1}$$

where $P_t(i) \coloneqq \mathbb{P}\big(I_t \in N_t(i)\big) = \sum_{j \in N_t(i)} p_t(j)$. This estimate is unbiased in the sense that $\mathbb{E}_t\big[\widehat{\ell}_t(i)\big] = \ell_t(i)$ for all $t \in [T]$ and all $i \in V$, where we denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$.

A key part of the standard regret analysis of q-FTRL (see, e.g., the proof of Lemma B.2 in Appendix B.1) is handling the variance term, which, with the choice of estimator given

Algorithm 3.1 q-FTRL for undirected feedback graphs

input: $q \in (0,1)$, $\eta > 0$
initialization: $p_1(i) \leftarrow 1/K$ for all $i = 1, \dots, K$
for $t = 1, \dots, T$ **do**
  Select action $I_t \sim p_t$ and incur loss $\ell_t(I_t)$
  Observe losses $\{(i, \ell_t(i)) : i \in N_t(I_t)\}$ and graph G_t
  Construct a loss estimate $\widehat{\ell}_t$ for ℓ_t ▷ e.g., (3.1) or (3.6)
  Let $p_{t+1} \leftarrow \operatorname{argmin}_{p \in \Delta_K} \eta \langle \sum_{s=1}^{t} \widehat{\ell}_s, p \rangle + \psi_q(p)$
end for

in (3.1), takes the following form

$$B_t(q) \coloneqq \sum_{i \in V} \frac{p_t(i)^{2-q}}{P_t(i)}. \tag{3.2}$$

By Hölder's inequality, this term can be immediately upper bounded by

$$B_t(q) \le \sum_{i \in V} p_t(i)^{1-q} \le \left(\sum_{i \in V} p_t(i) \right)^{1-q} \left(\sum_{i \in V} 1^{1/q} \right)^{q} = K^q,$$

while previous results on the regret analysis of bandits with graph feedback (Mannor and Shamir, 2011; Alon et al., 2017) would give that

$$B_t(q) \le \sum_{i \in V} \frac{p_t(i)}{P_t(i)} \le \alpha.$$

However, the former result would only recover a $\mathcal{O}(\sqrt{KT})$ regret bound (regardless of α) with the best choice of $q = 1/2$, which could be trivially achieved by ignoring side-observations of the losses, whereas the latter bound would only manage to achieve a $\mathcal{O}(\sqrt{\alpha T \log K})$ regret bound, incurring the extra $\sqrt{\log K}$ factor for all values of α. Other results in the literature (e.g., see Alon et al., 2013, 2015; Kocák et al., 2014; Zimmert and Lattimore, 2019; Esposito et al., 2022; Ito et al., 2022; Rouyer et al., 2022; Dann et al., 2023) do not bring an improvement in this setting when bounding the $B_t(q)$ term and, hence, do not suffice for achieving the desired regret bound. The following lemma provides a novel and improved bound on quantities of the same form as $B_t(q)$ in terms of the independence number $\alpha_t = \alpha$ of the undirected graph G_t.

Lemma 3.1. *Let $G = (V, E)$ be any undirected graph with $|V| = K$ vertices and independence number $\alpha(G) = \alpha$. Let $b \in [0,1]$, $p \in \Delta_K$ and consider any nonempty subset $U \subseteq \{v \in V : v \in$*

$N_G(v)\}$. *Then,*

$$\sum_{v\in U} \frac{p(v)^{1+b}}{\sum_{u\in N_G(v)} p(u)} \le \alpha^{1-b}.$$

Proof. First of all, observe that we can restrict ourselves to the subgraph $G[U]$ induced by U, i.e., the graph $G[U] = (U, E \cap (U \times U))$. This is because the neighbourhoods in this graph are such that $N_{G[U]}(v) \subseteq N_G(v)$ for all $v \in U$, and its independence number is $\alpha(G[U]) \le \alpha(G)$. Hence, it suffices to prove the claimed inequality for any undirected graph $G = (V, E)$ with all self-loops, any $p \in [0,1]^K$ such that $\|p\|_1 \le 1$, and the choice $U = V$. We assume this in what follows without loss of generality.

For any subgraph $H \subseteq G$ with vertices $V(H) \subseteq V$, denote the quantity we want to upper bound by

$$Q(H) \coloneqq \sum_{v\in V(H)} \frac{p(v)^{1+b}}{\sum_{u\in N_G(v)} p(u)}.$$

Our aim is thus to provide an upper bound to $Q(G)$.

Consider a greedy algorithm that incrementally constructs a subset of vertices in the following way: at each step, it selects a vertex v that maximizes $p(v)^b / \big(\sum_{u\in N_G(v)} p(u)\big)$, it adds v to the solution, and it removes v from G together with its neighbourhood $N_G(v)$. This step is iterated on the remaining graph until no vertex is left.

Let $S \coloneqq \{v_1, \dots, v_s\} \subseteq V$ be the solution returned by the above greedy algorithm on G. Also let $G_1, \dots, G_{s+1}$ be the sequence of graphs induced by the operations of the algorithm, where $G_1 = G$ and G_{s+1} is the empty graph, and let $N_r(v) \coloneqq N_{G_r}(v)$ for $v \in V(G_r)$. At every step $r \in [s]$ of the greedy algorithm, the contribution to $Q(G)$ of the removed vertices $N_r(v_r)$ amounts to

$$\begin{aligned} Q(G_r) - Q(G_{r+1}) = \sum_{v\in N_r(v_r)} \frac{p(v)^{1+b}}{\sum_{u\in N_1(v)} p(u)} &\le \sum_{v\in N_r(v_r)} p(v) \frac{p(v_r)^b}{\sum_{u\in N_1(v_r)} p(u)} \\ &\le \frac{\sum_{v\in N_1(v_r)} p(v)}{\sum_{u\in N_1(v_r)} p(u)} p(v_r)^b = p(v_r)^b, \end{aligned}$$

where the last inequality is due to the fact that $N_i(v) \subseteq N_j(v)$ for all $i \ge j$ and $v \in V_i$. Therefore, we can observe that

$$Q(G) = \sum_{r=1}^{s} \big(Q(G_r) - Q(G_{r+1})\big) \le \sum_{v\in S} p(v)^b.$$

The solution S is an independent set of G by construction. Consider now any independent set $A \subseteq V$ of G. We have that

$$\sum_{v \in A} p(v)^b \leq \max_{x \in \Delta_K} \sum_{v \in A} x(v)^b = |A| \max_{x \in \Delta_K} \sum_{v \in A} \frac{x(v)^b}{|A|}$$
$$\leq |A| \max_{x \in \Delta_K} \Big(\frac{1}{|A|} \sum_{v \in A} x(v) \Big)^b \leq |A|^{1-b} \leq \alpha^{1-b}, \tag{3.3}$$

where the second inequality follows by Jensen's inequality and the fact that $b \in [0,1]$. □

Observe that this upper bound is tight for general probability distributions $p \in \Delta_K$ over the vertices V of any strongly observable undirected graph G (containing at least one self-loop), as it is exactly achieved by the distribution $p^\star \in \Delta_K$ defined as $p^\star(i) \coloneqq \frac{1}{|S|} \mathbb{I}\{i \in S\}$ for some maximal independent set $S \subseteq V$ of G. Using this lemma, the following theorem provides our improved upper bound under the simplifying assumptions we made thus far.

Theorem 3.2. *Let $G_1, \dots, G_T$ be a sequence of undirected feedback graphs, where each G_t contains all self-loops and has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 3.1 is run with input*

$$q = \frac{1}{2}\Big(1 + \frac{\log(K/\alpha)}{\sqrt{\log(K/\alpha)^2 + 4} + 2}\Big) \in [1/2, 1) \qquad \text{and} \qquad \eta = \sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}},$$

and loss estimates (3.1), then its regret satisfies $R_T \leq 2\sqrt{e\alpha T\big(2 + \log(K/\alpha)\big)}$.

Proof. One can verify that for any $i \in V$, the loss estimate $\widehat{\ell}_t(i)$ defined in (3.1) satisfies $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \leq 1/P_t(i)$. Hence, using also that $\mathbb{E}_t\big[\widehat{\ell}_t(i)\big] = \ell_t(i)$, Lemma B.1 in Appendix B.1 implies that

$$R_T \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\Big[\sum_{i \in V} \frac{p_t(i)^{2-q}}{P_t(i)}\Big] \tag{3.4}$$
$$\leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q} \alpha^q T, \tag{3.5}$$

where the second inequality follows by Lemma 3.1 with $b = 1-q$ since all actions $i \in V$ are such that $i \in N_G(i)$. Our choices for q and η allow us to further upper bound the

right-hand side of (3.5) by

$$\begin{aligned}\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}}T &= \sqrt{2T\exp\Big(1+\frac{1}{2}\log(\alpha K)-\frac{1}{2}\sqrt{\log(K/\alpha)^2+4}\Big)\Big(2+\sqrt{\log(K/\alpha)^2+4}\Big)} \\ &\le \sqrt{2e\alpha T\Big(2+\sqrt{\log(K/\alpha)^2+4}\Big)} \le 2\sqrt{e\alpha T\sqrt{\log(K/\alpha)^2+4}} \\ &\le 2\sqrt{e\alpha T\left(2+\log(K/\alpha)\right)}\,. \end{aligned}$$

□

The bound stated in the above theorem coincides (up to constants) with the optimal regret bound for the experts setting (i.e., $\alpha = 1$) and the bandits setting (i.e., $\alpha = K$) simultaneously. Moreover, it interpolates between them in the intermediate cases where α ranges between 1 and K, introducing the multiplicative logarithmic factor only for graphs with independence number strictly smaller than K. We remark that the chosen values of q and η do in fact minimize the right-hand side of (3.5). Note that we relied on the knowledge of α to tune the parameter q. This is undesirable in general. We will show how to lift this requirement in Section 3.4. The same comment applies to Theorem 3.3, below.

We now show how to achieve the improved regret bound of Theorem 3.2 in the case of strongly observable undirected feedback graphs where some self-loops may be missing; i.e., there may be actions $i \in V$ such that $i \notin N_G(i)$. Using the loss estimator defined in (3.1) may lead to a large variance term due to the presence of actions without self-loops. One approach to deal with this (see, e.g., Zimmert and Lattimore, 2019 or Luo et al., 2023) is to suitably alter the loss estimates of these actions.

Define $S_t \coloneqq \{i \in V : i \notin N_t(i)\}$ as the subset of actions without self-loops in the feedback graph G_t at each time step $t \in [T]$. The idea is that we need to carefully handle some action $i \in S_t$ only in the case when the probability $p_t(i)$ of choosing i at round t is sufficiently large, say, larger than $1/2$. Define the set of such actions as $J_t \coloneqq \{i \in S_t : p_t(i) > 1/2\}$ and observe that $|J_t| \le 1$. Similarly to (Zimmert and Lattimore, 2019), define new loss estimates

$$\widehat{\ell}_t(i) \coloneqq \begin{cases} \frac{\ell_t(i)}{P_t(i)}\mathbb{I}\{I_t \in N_t(i)\} & \text{if } i \in V \setminus J_t \\ \frac{\ell_t(i)-1}{P_t(i)}\mathbb{I}\{I_t \in N_t(i)\}+1 & \text{if } i \in J_t \end{cases} \tag{3.6}$$

for which it still holds that $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$ and that $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$ for all $i \notin J_t$. This change, along with the use of Lemma 3.1 for the actions in $V \setminus S_t$, suffices in order to prove the following regret bound (see Appendix B.2 for the proof) when the feedback graphs do not necessarily contain self-loops for all actions.

Theorem 3.3. *Let $G_1, \dots, G_T$ be a sequence of strongly observable undirected feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 3.1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\log(K/\alpha)}{\sqrt{\log(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \qquad \text{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}},$$

and loss estimates (3.6), then its regret satisfies $R_T \le 6\sqrt{e\alpha T\left(2 + \log(K/\alpha)\right)}$.

3.4 Adapting to Arbitrary Sequences of Graphs

In the previous section, we assumed for simplicity that all the graphs have the same independence number. This independence number was then used to tune q, the parameter of the Tsallis entropy regularizer used by the algorithm. In this section, we show how to extend our approach to the case when the independence numbers of the graphs are neither the same nor known a-priori by the learner. Had these independence numbers been known a-priori, one approach is to set q as in Theorem 3.3 but using the average independence number

$$\overline{\alpha}_T := \frac{1}{T}\sum_{t=1}^{T} \alpha_t .$$

Doing so would allow us to achieve a $\mathcal{O}\left(\sqrt{\sum_{t=1}^{T} \alpha_t (1 + \log(K/\overline{\alpha}_T))}\right)$ regret bound. We now show that we can still recover a bound of the same order without prior knowledge of $\overline{\alpha}_T$. For round t and any fixed $q \in [0,1]$, define

$$H_t(q) := \sum_{i \in V \setminus S_t} \frac{p_t(i)^{2-q}}{P_t(i)} .$$

We know from Lemma 3.1 that $H_t(q) \le \alpha_t^q$. Thus, we can leverage these observations and use a doubling trick (similar in principle to Alon et al., 2017) to guess the value of $\overline{\alpha}_T$. This approach is outlined in Algorithm 3.2. Starting with $r = 0$ and $T_r = 1$, the idea is to instantiate Algorithm 3.1 at time-step T_r with q and η set as in Theorem 3.3 but with 2^r replacing the independence number. Then, at $t \ge T_r$, we increment r and restart Algorithm 3.1 only if

$$\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1},$$

since (again thanks to Lemma 3.1) the left-hand side of the above inequality is always upper bounded by $\overline{\alpha}_T$. The following theorem shows that this approach essentially enjoys the same regret bound of Theorem 3.3 up to an additive $\log_2 \overline{\alpha}_T$ term.

Algorithm 3.2 q-FTRL for an arbitrary sequence of strongly observable undirected graphs

input: Time horizon T
define: For each $r \in \{0, \dots, \lfloor \log_2 K \rfloor\}$,

$$q_r := \frac{1}{2}\left(1 + \frac{\log(K/2^r)}{\sqrt{\log(K/2^r)^2 + 4} + 2}\right) \qquad \text{and} \qquad \eta_r := \sqrt{\frac{2q_r K^{1-q_r}}{11T(1-q_r)\,(2^r)^{q_r}}}$$

initialization: $T_0 \leftarrow 1$, $r \leftarrow 0$, instantiate Algorithm 3.1 with $q = q_0$, $\eta = \eta_0$, and loss estimates (3.6)
for $t = 1, \dots, T$ **do**
  Perform one step of the current instance of Algorithm 3.1
  if $\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1}$ **then**
    $r \leftarrow r + 1$
    $T_r \leftarrow t + 1$
    Restart Algorithm 3.1 with $q = q_r$, $\eta = \eta_r$, and loss estimates (3.6)
  end if
end for

Theorem 3.4. *Let* $C := 4\sqrt{6e}\frac{\sqrt{\pi}+\sqrt{4-2\log 2}}{\log 2}$. *Then, the regret of Algorithm 3.2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T} \alpha_t \left(2 + \log\left(\frac{K}{\overline{\alpha}_T}\right)\right)} + \log_2 \overline{\alpha}_T\,.$$

Proof sketch. For simplicity, we sketch here the proof for the case when in every round t, all the nodes have self-loops; hence, $H_t(q) = B_t(q)$. See the full proof in Appendix B.3, which treats the general case in a similar manner. Let $n := \lceil \log_2 \overline{\alpha}_T \rceil$ and assume without loss of generality that $\overline{\alpha}_T > 1$. Since Lemma 3.1 implies that for any r and t, $B_t(q_r) \le \alpha_t^{q_r}$, we have as a consequence that for any $t \ge T_r$,

$$\frac{1}{T}\sum_{s=T_r}^{t} B_s(q_r)^{1/q_r} \le \frac{1}{T}\sum_{s=T_r}^{t} \alpha_s \le \overline{\alpha}_T \le 2^n\,.$$

Hence, the maximum value of r that the algorithm can reach is $n-1$. In doing so, we will execute n instances of Algorithm 3.1, each corresponding to a value of $r \in \{0, \dots, n-1\}$. For every such r, we upper bound the instantaneous regret at step $T_{r+1} - 1$ (the step when the restarting condition is satisfied) by 1, hence the added $\log_2 \overline{\alpha}_T$ term in the regret bound. For the rest of the interval; namely, for $t \in [T_r, T_{r+1} - 2]$, we have via (3.4) that the regret of

Algorithm 3.1 is bounded by

$$\frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r)\,. \tag{3.7}$$

Define $T_{r:r+1} = T_{r+1} - T_r - 1$, and notice that

$$\begin{aligned}\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r) &\le T_{r:r+1}\Big(\frac{1}{T_{r:r+1}}\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r)^{1/q_r}\Big)^{q_r} \\ &\le T_{r:r+1}\Big(\frac{T}{T_{r:r+1}}2^{r+1}\Big)^{q_r} \le 2T\big(2^r\big)^{q_r}\,,\end{aligned}$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0,1)$, and the second follows from the restarting condition of Algorithm 3.2. After, plugging this back into (3.7), we can simply use the definitions of η_r and q_r and bound the resulting expression in a similar manner to the proof of Theorem 3.2. Overall, we get that

$$R_T \le 4\sqrt{3eT}\sum_{r=0}^{n-1}\sqrt{2^r\log\big(e^2K2^{-r}\big)} + \log_2\overline{\alpha}_T\,,$$

from which the theorem follows by using Lemma B.3 in Appendix B.1, which shows, roughly speaking, that the sum on the right-hand side is of the same order as its last term. □

Although Algorithm 3.2 requires knowledge of the time horizon, this can be dealt with by applying a standard doubling trick on T at the cost of a larger constant factor. It is also noteworthy that the bound we obtained is of the form $\sqrt{T\overline{\alpha}_T(1+\log(K/\overline{\alpha}_T))}$ and not $\sqrt{\sum_t \alpha_t(1+\log(K/\alpha_t))}$. Although both coincide with the bound of Theorem 3.3 when α_t is the same for all time steps, the latter is smaller via the concavity of $x(1+\log(K/x))$ in x. It is not clear, however, whether there is a tuning of $q \in (0,1)$ that can achieve the second bound (even with prior knowledge of the entire sequence $\alpha_1,\dots,\alpha_T$ of independence numbers).

3.5 An Improved Lower Bound via Multitask Learning

In this section we provide a new lower bound on the minimax regret showing that, apart from the bandit case, a logarithmic factor is indeed necessary in general. When the graph is fixed over time, it is known that a lower bound of order $\sqrt{\alpha T}$ holds for any value of α (Mannor and Shamir, 2011; Alon et al., 2017). Whereas for the experts case

($\alpha = 1$), the minimax regret is of order[2] $\sqrt{T \log K}$ (Cesa-Bianchi et al., 1997). The following theorem provides, for the first time, a lower bound that interpolates between the two aforementioned bounds for the intermediate values of α.

Theorem 3.5. *Pick any $K \geq 2$ and any α such that $2 \leq \alpha \leq K$. Then, for any algorithm and for all $T \geq \frac{\alpha \log_\alpha K}{4 \log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1, \ldots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t = 1, \ldots, T$ and*

$$R_T \geq \frac{1}{18\sqrt{2}} \sqrt{\alpha T \log_\alpha K}.$$

In essence, the proof of this theorem (see Appendix B.4) constructs a sequence of feedback graphs and losses that is equivalent to a hard instance of the multitask bandit problem (MTB) (Cesa-Bianchi and Lugosi, 2012), an important special case of combinatorial bandits with a convenient structure for proving lower bounds (Audibert et al., 2014; Cohen et al., 2017; Ito et al., 2019). We consider a variant of MTB which we have already relied on for proving the lower bound of Theorem 2.7 in the mediator feedback setting. In this variant, at the beginning of each round, the decision-maker selects an arm to play in each one of M stochastic bandit games. Subsequently, the decision-maker only observes (and suffers) the loss of the arm played in a single randomly selected game. For proving the lower bound, we use a class of stationary stochastic adversaries (i.e., environments), each generating graphs and losses in a manner that simulates an MTB instance.

Fix $2 \leq \alpha \leq K = |V|$ and assume for simplicity that $M = \log_\alpha K$ is an integer. We now construct an instance of online learning with time-varying feedback graphs $G_t = (V, E_t)$ with $\alpha(G_t) = \alpha$ that is equivalent to an MTB instance with M bandit games each containing α "base actions". Since $K = \alpha^M$, we can uniquely identify each action in V with a vector $a = \big(a(1), \ldots, a(M)\big)$ in $[\alpha]^M$. The action $a_t \in V$ chosen by the learner at round t is equivalent to a choice of base actions $a_t(1), \ldots, a_t(M)$ in the M games. The feedback graph at every round is sampled uniformly at random from a set of M undirected graphs $\{G^i\}_{i=1}^M$, where $G^i = (V, E^i)$ is such that $(a, a') \in E^i$ if and only if $a(i) = a'(i)$. This means (see Figure 3.1) that each graph G^i consists of α isolated cliques $\{C_{i,j}\}_{j=1}^\alpha$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. Clearly, the independence number of any such graph is α. Drawing feedback graph $G_t = G^i$ corresponds to the activation of game i in the MTB instance. Hence, choosing $a_t \in V$ with feedback graph $G_t = G^i$ is equivalent to playing base action $a_t(i)$ in game i in the MTB.

[2] As a lower bound, this is known to hold asymptotically as K and T grow. However, it can also be shown to hold non-asymptotically (though with worse leading constants); see (Haussler et al., 1998, Theorem 3.22) or (Cesa-Bianchi and Lugosi, 2006, Theorem 3.6).

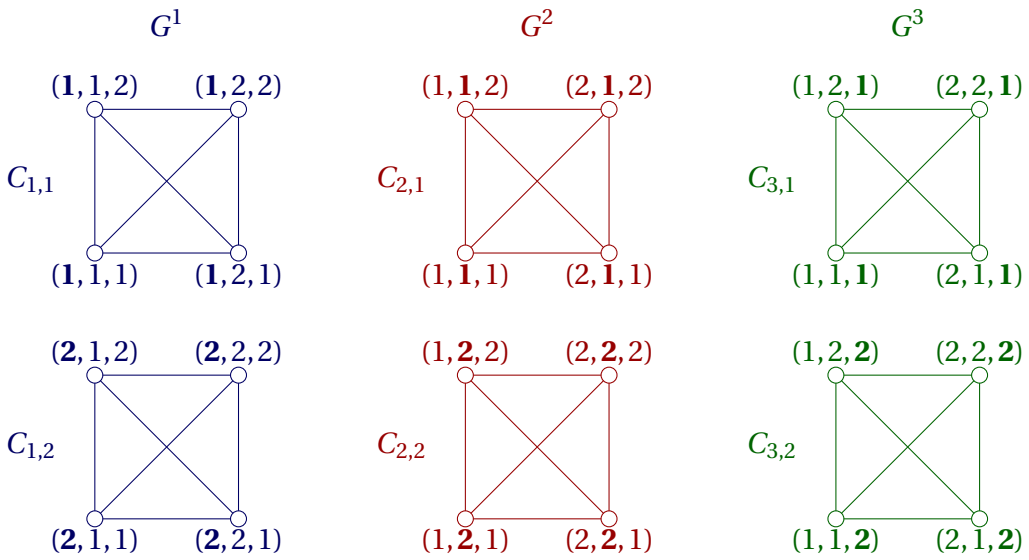

Fig. 3.1 This figure shows an example of the multi-task bandit construction used to prove the lower bound. Here, $K = 8$ and $\alpha = 2$; thus, the number of games is $M = 3$. Each action is identified by a tuple of three numbers, each corresponding to a choice of one out of a pair of "base actions" in each game. Each of the three graphs in the figure corresponds to a game, such that two actions share an edge if and only if they choose the same base action in the corresponding game. At every round, a graph is randomly drawn, and all actions belonging to the same clique suffer the same loss.

As for the losses, we enforce that, given a feedback graph G_t, all actions that belong to the same clique of the feedback graph are assigned the same loss. Namely, if $G_t = G^i$ and $a(i) = a'(i) = j$, then $\ell_t(a) = \ell_t(a')$, which can be seen as the loss $\ell_t(j)$ assigned to base action j in game G^i. To choose the distribution of the losses for the base actions, we apply the classic needle-in-a-haystack approach of Auer et al. (2002a) over the M games. More precisely, we construct a different environment for each action $a \in V$ in such a way that the distribution of the losses in each MTB game slightly favors (with a difference of a small $\varepsilon > 0$) the base action corresponding to a in that game. The proof then proceeds similarly to that of Theorem 2.7.

While both our upper and lower bounds achieve the desired goal of interpolating between the minimax rates of experts and bandits, the logarithmic factors in the two bounds are not exactly matching. In particular, if we compare $1+\log_2(K/\alpha)$ and $\log_\alpha K$, we can see that although they coincide at $\alpha = 2$ and $\alpha = K$, the former is larger for intermediate values. It is reasonable to believe that the upper bound is of the correct order, seeing as it came about naturally as a result of choosing the best parameter for the Tsallis entropy regularizer, whereas achieving the extra logarithmic factor in the lower bound required a somewhat contrived construction that appealed to the flexibility of the time-varying

graphs formulation. This has been corroborated by the later work of Chen et al. (2024) as we discuss shortly.

3.6 Conclusions and Future Work

We have shown in this chapter that a proper tuning of the q-FTRL algorithm allows one to achieve a $\mathcal{O}\big(\sqrt{\alpha T(1+\log(K/\alpha))}\big)$ regret for the problem of online learning with undirected strongly observable feedback graphs. Our bound interpolates between the minimax regret rates of the bandits and the experts problems, the two extremes of the strongly observable graph feedback spectrum. Furthermore, we have shown that an analogous bound can be achieved when the graphs vary over time, and without requiring any prior knowledge on the graphs. These results are complemented by our new lower bound of $\Omega\big(\sqrt{\alpha T(\log K)/(\log \alpha)}\big)$, which holds for $\alpha \geq 2$ and shows the necessity of a logarithmic factor in the minimax regret except for the bandits case.

At the time of their publication, these findings had provided the tightest characterization of the minimax rate for this setting. Later, Chen et al. (2024) showed a lower bound for fixed feedback graphs composed of disjoint cliques that implies worst-case optimality (up to constant factors) of our proposed algorithm for each pair of K and α—see the next section for a more detailed comparison with results therein. Extending our results to the case of directed strongly observable feedback graphs is a harder task, better understanding this more general setting is an interesting future direction.

3.7 Comparison with Chen et al. (2024)

Chen et al. (2024) consider a special case of the undirected feedback graph problem where the graph (fixed and known) is composed of α disjoint cliques with self-loops. For $j \in [\alpha]$, let m_j denote the number of actions in the j-th clique, implying that $\sum_{j=1}^{\alpha} m_j = K$ (the number of arms). For this problem, Theorem 4 of Chen et al. (2024) provides a lower bound of order $\sqrt{T\sum_{j=1}^{\alpha}\log(m_j+1)}$. In particular, if the cliques are balanced (i.e., $m_1 = \cdots = m_\alpha = K/\alpha$), the lower bound becomes of order $\sqrt{\alpha T\log(1+K/\alpha)}$, thus matching the regret bound of Algorithm 3.1. This means that, for any value of $1 \leq \alpha \leq K$, there are feedback graphs on K nodes with independence number α such that no other algorithm can achieve a better minimax regret guarantee than that of our proposed algorithm.

We emphasize that this does not imply *graph-specific* minimax optimality. Indeed, as shown by Chen et al. (2024), when the cliques are unbalanced, the regret guarantee of our algorithm can be inferior to that of the algorithm they proposed, which matches the $\sqrt{T\sum_{j=1}^{\alpha}\log(m_j+1)}$ bound. However, beyond the disjoint cliques case, their algorithm requires computing a minimum clique cover for the given feedback graph G, which is known to be NP-hard (Karp, 1972). More importantly, their reliance on a clique cover leads to a dependence of the regret on the clique cover number $\theta(G)$ instead of the independence number $\alpha(G)$. One can argue that the ratio between $\theta(G)$ and $\alpha(G)$ can be $\Omega(K/(\log K)^2)$ for most graphs on a sufficiently large number K of vertices (e.g., see Mannor and Shamir, 2011, Section 6). Finally, it is not clear how to generalize their approach to time-varying feedback graphs (informed or uninformed). Hence, despite the contributions of our work and those of Chen et al. (2024), the problem of characterizing the minimax regret rate at a graph-based granularity still calls for further investigation.

Chapter 4

Improved Regret Bounds for Bandits with Expert Advice*

4.1 Introduction

In this short chapter, we revisit the bandits with expert advice problem, for which we have already provided a novel regret bound in Section 2.3.4, having seen this problem as a variant of the mediator feedback framework. Here, we combine tools from the last two chapters to improve still upon this result, as well as to prove a new lower bound. Firstly, we recall in the following the description of the problem along with relevant results from the literature, expanding upon our discussion in Section 1.2.3.

As introduced by Auer et al. (1995, 2002a), this problem is an extension of the non-stochastic multi-armed bandit problem where the learner receives in every round a recommendation, in the form of a probability distribution over the actions, from each expert in a given set. The goal of the learner is to minimize their expected regret with respect to the best expert in hindsight; that is, the difference between their expected cumulative loss and that of the best expert. This provides a simple and general framework for incorporating contextual information into the bandit problem, seeing the set of experts as a set of benchmark strategies each mapping an (unobserved) context to a (randomized) action choice. Auer et al. (2002a) proposed the EXP4 algorithm as a solution strategy that has since become an important baseline or building block for addressing many bandit settings; for example, sleeping bandits (Kleinberg et al., 2010), online multi-class classification (Daniely and Helbertal, 2013), online non-parametric learning (Cesa-Bianchi

*This chapter is based on (Cesa-Bianchi, Eldowa, Esposito, and Olkhovskaya, 2024).

et al., 2017), and non-stationary bandits (Luo et al., 2018). They proved a bound of order $\sqrt{KT\log N}$ on the expected regret incurred by the EXP4 strategy, where T denotes the number of rounds, K the number of actions, and N the number of experts. This result is of a worst-case nature, in that it holds for any sequence of losses assigned to the actions and any sequence of expert recommendations.

The appealing feature of the bound of Auer et al. (2002a) is that it exhibits only a logarithmic dependence on the number of experts, in addition to the $\sqrt{K}$ dependence on the number of actions known to be unavoidable in the classical bandit problem, where the learner competes with the best fixed action. While the minimax regret in the latter problem is known to be of order $\sqrt{KT}$ (Audibert and Bubeck, 2009), a similar exact characterization remains missing for the expert advice problem. Kale (2014) studied a generalized version of the bandits with expert advice problem—originally proposed by Seldin et al. (2013)—where the learner is only allowed to query the advice of $M \le N$ experts. When $M = N$, the results of Kale (2014) imply an upper bound of order $\sqrt{\min\{K,N\}T(1+\log(N/\min\{K,N\}))}$ on the minimax regret, improving upon the bound of Auer et al. (2002a). Unlike the latter, the logarithmic factor in Kale's (2014) bound diminishes as K increases with respect to N, leading to a bound of order $\sqrt{NT}$ when $N \le K$, which is tight in general as the experts in that case can be made to emulate an N-armed bandit problem. This improved bound was achieved via the PolyINF algorithm (Audibert and Bubeck, 2009, 2010) played on the expert set utilizing the importance-weighted loss estimators of EXP4. Later, Seldin and Lugosi (2016) proved a lower bound of order $\sqrt{KT(\log N)/(\log K)}$ for $N \ge K$.

As these upper and lower bounds still do not match, the correct minimax rate remains unclear. Here, we take a step towards resolving this issue by showing that the upper bound is not improvable in general under a restricted feedback model in which the importance weighted loss estimators used by EXP4 or PolyINF remain implementable. In this restricted model, without observing the experts' recommendations, the learner picks an expert (possibly at random) at the beginning of each round, and the environment subsequently samples the action to be executed from the chosen expert's distribution. Afterwards, the learner only observes the distributions of the experts that had assigned positive probability to the chosen action. Via a reduction from the problem of bandits with graph feedback (see Chapter 3), we use the recent results of Chen et al. (2024) to obtain a lower bound of order $\sqrt{KT\log(N/K)}$ for $N > K$.

Departing from the worst-case results discussed thus far, a few works have obtained instance-dependent bounds for this problem. The dependence on the instance can be in terms of the assigned sequence of losses through small loss bounds (see Allen-Zhu et al.,

2018), or in terms of the sequence of expert recommendations through bounds that reflect the similarity between the recommended expert distributions like that of Theorem 2.2, see also (McMahan and Streeter 2009; Lattimore and Szepesvári 2020, Theorem 18.3). Our focus here is on the latter case, where the best bound to our knowledge is the bound of order $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t \log N}$ that we derived in Theorem 2.2 for the EXP4 algorithm, where, to recall, $\mathcal{C}_t$ is the (chi-squared) capacity of the recommended distributions at round t, which measures the dissimilarity between the experts' recommendations and satisfies $0 \le \mathcal{C}_t \le \min\{K, N\} - 1$. Improving upon this result, we illustrate that it is possible to achieve a bound of order $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t \big(1 + \log(N / \max\{\bar{\mathcal{C}}_T, 1\})\big)}$, where $\bar{\mathcal{C}}_T = \sum_{t=1}^{T} \mathcal{C}_t / T$ is the average capacity. This bound combines the best of the bound of Theorem 2.2 (its dependence on the agreement between the experts) and that of Kale (2014) (its improved log factor), simultaneously outperforming both. Notice that the form of this bound is the same as that of Theorem 3.4 in the feedback graphs setting, with the capacity replacing the independence number. Indeed, we arrive at this result by building upon the doubling trick and algorithmic scheme used to achieve the guarantee in Theorem 3.4.

Road map. We formalize the problem setting and introduce some notation in the next section. In Section 4.3, as a preliminary building block, we present Algorithm 4.1, an adaptation of the q-FTRL algorithm from Chapter 3 to the bandits with expert advice problem. This algorithm is essentially equivalent to an instance of the PolyINF algorithm (see Abernethy et al., 2015; Audibert et al., 2011) which was used by Kale (2014) to achieve the best known worst-case upper bound for this problem. We then show in Section 4.4 that combining this algorithm with a doubling trick (in the spirit of Algorithm 3.2) allows us to achieve the improved instance-based bound mentioned above. The lower bound for the restricted feedback setting is presented in Section 4.5. Finally, we provide some concluding remarks in Section 4.6.

4.2 Preliminaries

For $x, y \in \mathbb{R}$, let $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. Moreover, we define $x_+ := x \vee 0$. Let $V = [N]$ be a set of N experts and $\mathcal{A} = [K]$ be a set of K actions. We consider a sequential decision-making problem where a learner interacts with an unknown environment for T rounds. The environment is characterized by a fixed and unknown sequence of loss vectors $(\ell_t)_{t \in [T]}$, where $\ell_t \in [0,1]^K$ is the assignment of losses for the actions at round t, and a fixed and unknown sequence of expert advice $(\theta_t^i)_{i \in V, t \in [T]}$, where $\theta_t^i \in \Delta_K$ is the

distribution over actions recommended by expert i at round t.[1] At the beginning of each round $t \in [T]$, the expert recommendations $(\theta_t^i)_{i \in V}$ are revealed to the learner, who then selects (possibly at random) an action $A_t \in \mathcal{A}$ and subsequently suffers and observes the loss $\ell_t(A_t)$. For an expert $i \in V$, we define $y_t(i) := \sum_{a \in \mathcal{A}} \theta_t^i(a) \ell_t(a)$ as its loss in round t. The goal is to minimize the expected regret with respect to the best expert in hindsight:

$$R_T := \mathbb{E}\Big[\sum_{t=1}^T \ell_t(A_t)\Big] - \min_{i \in V} \sum_{t=1}^T y_t(i)\,,$$

where the expectation is taken with respect to the randomization of the learner.

4.3 q-FTRL for Bandits with Expert Advice

As mentioned in Chapter 2, the EXP4 algorithm can be seen as an instance of the FTRL framework where a distribution p_t over the experts is maintained at each round t and updated as follows

$$p_{t+1} \leftarrow \operatorname*{arg\,min}_{p \in \Delta_N} \eta \Big\langle \sum_{s=1}^t \widehat{y}_s, p \Big\rangle + \sum_{i \in V} p(i) \log p(i)\,,$$

where $\eta > 0$ is the learning rate, the second term is the negative Shannon entropy of p, and $\widehat{y}_s(i)$ is an importance-weighted estimate of $y_s(i)$. The action A_t is drawn from the mixture distribution $\sum_{i \in V} p_t(i) \theta_t^i(\cdot)$. In the spirit of the q-FTRL algorithm of Chapter 3, we consider here a generalized strategy (outlined in Algorithm 4.1) where the negative Shannon entropy is replaced with the negative q-Tsallis entropy, which, to recall, for $q \in (0,1)$ is given by

$$\psi_q(x) := \frac{1}{1-q}\left(1 - \sum_{i \in V} x(i)^q\right) \qquad \forall x \in \Delta_N\,,$$

where the negative Shannon entropy is recovered in the limit when $q \to 1$. The following theorem provides a regret bound for the algorithm, which takes the same form as that we proved in Theorem 3.2 for q-FTRL in the feedback graphs setting. This result is not novel, a similar bound is implied by Theorem 2 in Kale (2014) for a closely related algorithm in a more general setting. We provide a concise proof of the result for completeness. As mentioned before, when $N \le K$, this bound is trivially tight in general. While when $N > K$,

[1] To reduce clutter, we use here a slightly different notation from that used in Chapter 2 in the context of the mediator feedback framework.

Algorithm 4.1 q-FTRL for bandits with expert advice

input: $q \in (0,1)$, $\eta > 0$
initialization: $p_1(i) \leftarrow 1/N$ for all $i \in V$
for $t = 1, \dots, T$ **do**
  receive expert advice $(\theta_t^i)_{i \in V}$
  draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
  construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) := \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
  let $p_{t+1} \leftarrow \operatorname{argmin}_{p \in \Delta_N} \eta \langle \sum_{s=1}^t \widehat{y}_s, p \rangle + \psi_q(p)$
end for

we prove an order-wise matching minimax lower bound in Section 4.5 under additional restrictions on the received feedback.

Theorem 4.1. *Algorithm 4.1 run with*

$$q = \frac{1}{2}\left(1 + \frac{\log\big(N/(K \wedge N)\big)}{\sqrt{\log\big(N/(K \wedge N)\big)^2 + 4} + 2}\right) \in [1/2, 1) \quad \textit{and} \quad \eta = \sqrt{\frac{2qN^{1-q}}{T(1-q)(K \wedge N)^q}},$$

satisfies

$$R_T \le 2\sqrt{e(K \wedge N)T\big(2 + \log\big(N/(K \wedge N)\big)\big)}.$$

Proof. Let $i^* \in \operatorname{argmin}_{i \in V} \sum_{t=1}^T y_t(i)$, and note that $R_T = \mathbb{E}\sum_{t=1}^T \big(y_t(I_t) - y_t(i^*)\big)$ because $\mathbb{E}\,\ell_t(A_t) = \mathbb{E}\,y_t(I_t)$. For round $t \in [T]$, let $\mathcal{F}_t := \sigma(I_1, A_1, \dots, I_t, A_t)$ denote the σ-algebra generated by the random events up to the end of round t, and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_0$ being the trivial σ-algebra. For action $a \in \mathcal{A}$, let $\phi_t(a) := \sum_{i \in V} p_t(i)\theta_t^i(a)$ and note that conditioned on $\mathcal{F}_{t-1}$, A_t is distributed according to ϕ_t. As p_t is $\mathcal{F}_{t-1}$-measurable, it is then easy to verify that $\mathbb{E}_t \widehat{y}_t = y_t$. Hence, Lemma B.1 implies that

$$R_T \le \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q}\sum_{t=1}^T \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{y}_t(i)^2\right]. \tag{4.1}$$

For fixed $t \in [T]$ and $i \in V$, we have that

$$\mathbb{E}_t\left[\widehat{y}_t(i)^2\right] = \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\ell_t(A_t)^2\right] \le \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\right] = \mathbb{E}_t\left[\sum_{a \in \mathcal{A}} \frac{\theta_t^i(a)^2}{\phi_t(a)^2}\mathbb{I}\{a = A_t\}\right] = \sum_{a \in \mathcal{A}} \frac{\theta_t^i(a)^2}{\phi_t(a)} \tag{4.2}$$

where the inequality holds because $\ell_t(A_t) \in [0,1]$ and the final equality holds because $\mathbb{E}_t \mathbb{I}\{a = A_t\} = \mathbb{P}(a = A_t \mid \mathcal{F}_{t-1}) = \phi_t(a)$. Hence, it holds that

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q}\,\widehat{y}_t(i)^2\right] &= \sum_{a\in\mathcal{A}} \frac{\sum_{i\in V} p_t(i)^{2-q}\theta_t^i(a)^2}{\phi_t(a)} \\
&\le \sum_{a\in\mathcal{A}} \frac{\sum_{i\in V} p_t(i)^{2-q}\theta_t^i(a)^{2-q}}{\phi_t(a)} \max_{i\in V}\theta_t^i(a)^q \\
&\le \sum_{a\in\mathcal{A}} \frac{\left(\sum_{i\in V} p_t(i)\theta_t^i(a)\right)^{2-q}}{\phi_t(a)} \max_{i\in V}\theta_t^i(a)^q \\
&= \sum_{a\in\mathcal{A}} \phi_t(a)\Big(\frac{\max_{i\in V}\theta_t^i(a)}{\phi_t(a)}\Big)^q \le \Big(\sum_{a\in\mathcal{A}} \max_{i\in V}\theta_t^i(a)\Big)^q \le (K\wedge N)^q\,,
\end{aligned}$$

where the second inequality follows from the superadditivity of x^{2-q} for $x \ge 0$ and $q \in (0,1)$, the third inequality follows from the concavity of x^q for $q \in (0,1)$ because of Jensen's inequality, and the last inequality holds since $\max_{i\in V}\theta_t^i(a) \le \min\{1, \sum_{i\in V}\theta_t^i(a)\}$. Substituting back into (4.1) yields that

$$R_T \le \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q}(K\wedge N)^q T\,.$$

For brevity, let $\xi := (K \wedge N)$. In a similar manner to the proof of Theorem 3.2, substituting the specified values of η and q allows us to conclude the proof:

$$\begin{aligned}
R_T &\le \sqrt{\frac{2N^{1-q}\xi^q}{q(1-q)}T} \\
&= \sqrt{2T\exp\Big(1+\frac{1}{2}\log(\xi N) - \frac{1}{2}\sqrt{\log(N/\xi)^2+4}\Big)\Big(2+\sqrt{\log(N/\xi)^2+4}\Big)} \\
&\le \sqrt{2T\exp\Big(1+\frac{1}{2}\log(\xi N) - \frac{1}{2}\log(N/\xi)\Big)\Big(2+\sqrt{\log(N/\xi)^2+4}\Big)} \\
&= \sqrt{2e\xi T\Big(2+\sqrt{\log(N/\xi)^2+4}\Big)} \le 2\sqrt{e\xi T\sqrt{\log(N/\xi)^2+4}} \\
&\le 2\sqrt{e\xi T\left(2+\log(N/\xi)\right)}\,.
\end{aligned}$$

□

4.4 An Improved Instance-Based Regret Bound

We now obtain a more refined regret bound whose form is analogous to the bound of Theorem 4.1 except that it depends on the similarity between the experts' recommendations

at each round, replacing $K \wedge N$ with the average capacity of the experts' recommendations, which can be seen as an effective number of experts. Before discussing the algorithm, we recall some relevant quantities from Chapter 2.[2] For any round $t \in [T]$ and $\tau \in \Delta_N$,

$$Q_t(\tau) \coloneqq \sum_{i \in V} \tau(i) \chi^2\big(\theta_t^i \,\|\, \textstyle\sum_{j \in V} \tau(j) \theta_t^j\big) = \sum_{a \in \mathcal{A}} \frac{\sum_{i \in V} \tau(i) \theta_t^i(a)^2}{\sum_{j \in V} \tau(j) \theta_t^j(a)} - 1,$$

where $\chi^2(p \,\|\, q) \coloneqq \sum_{a \in \mathcal{A}} q(a) \big(p(a)/q(a) - 1\big)^2 = \sum_{a \in \mathcal{A}} p(a)^2/q(a) - 1$ is the chi-squared divergence between distributions $p, q \in \Delta_K$. Whereas

$$\mathcal{C}_t \coloneqq \sup_{\tau \in \Delta_N} Q_t(\tau) \qquad \text{and} \qquad \bar{\mathcal{C}}_T \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathcal{C}_t$$

denote the chi-squared capacity of the recommended distributions at round t and its average over the T rounds. As was shown in Chapter 2, $\mathcal{C}_t$ is never larger than $(K \wedge N) - 1$ and can be arbitrarily smaller depending on the agreement between the experts at round t. In particular, it vanishes when all recommendations are identical.

The idea of Algorithm 4.2 is to tune Algorithm 4.1 as done in Theorem 4.1 but with $\bar{\mathcal{C}}_T$ replacing $K \wedge N$. However, to avoid requiring prior knowledge of $\bar{\mathcal{C}}_T$, we rely on a doubling trick to adapt to its value, similar to what we did in Section 3.4 to adapt to the value of the average independence number in the feedback graphs problem. In a given round t, we maintain a running instance of Algorithm 4.1 tuned with an estimate for $\bar{\mathcal{C}}_T$. Let m_t be the round when the present execution of Algorithm 4.1 had started. If the current estimate is found to be smaller than $\frac{1}{2T} \sum_{s=m_t}^{t} Q_s(p_s)$, the algorithm is restarted and the estimate is (at least) doubled. This quantity we test against is a simple lower bound for $\bar{\mathcal{C}}_T/2$ that can be constructed without computing the capacity at any round. As the value of $\bar{\mathcal{C}}_T$ can be arbitrarily close to zero, the initial guess (which ideally should be a lower bound for $\bar{\mathcal{C}}_T$) is left as a user-specified parameter (denoted by J) for the algorithm, and appears in the first (and more general) bound of Theorem 4.2. The second statement of the theorem shows that choosing $\log(e^2 N)/T$ as the initial guess suffices to obtain a bound of order $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t \big(1 + \log(N/\max\{\bar{\mathcal{C}}_T, 1\})\big)}$, up to an additive $\log N$ term. This simultaneously outperforms the $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t \log N}$ bound of Theorem 2.2 and the $\sqrt{(K \wedge N) T \big(1 + \log(N/(K \wedge N))\big)}$ bound of Kale (2014).

The proof combines elements from the proof of Theorem 2.1 (the main result from the mediator feedback setting) and the proof of Theorem 3.4 (the regret bound we provided

[2] Again, with a slightly varied notation that is more apt for the present setting.

Algorithm 4.2 q-FTRL with the doubling trick for bandits with expert advice

input: $J \in (0, N]$
initialization: $r_1 \leftarrow \lceil \log_2 J \rceil - 1$, $m_1 \leftarrow 1$, $p_1(i) \leftarrow 1/N$ for all $i \in V$
define: For each integer $r \in (-\infty, \log_2 N]$,

$$q_r := \frac{1}{2}\Big(1 + \frac{\log(N/2^r)}{\sqrt{\log(N/2^r)^2 + 4} + 2}\Big)$$

$$\eta_r := \min\left\{ \sqrt{\frac{q_r(N^{1-q_r} - 1)}{eT(1-q_r)\,(2^r)^{q_r}}}\,,\ \frac{q_r}{1-q_r}\Big(1 - e^{\frac{q_r - 1}{2 - q_r}}\Big) \right\}$$

for $t = 1, \dots, T$ **do**
  receive expert advice $(\theta_t^i)_{i \in V}$
  draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
  construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) := \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
  if $\frac{1}{T}\sum_{s=m_t}^t Q_s(p_s) > 2^{r_t+1}$ **then**
    $p_{t+1}(i) \leftarrow 1/N$ for all $i \in V$
    $r_{t+1} \leftarrow \lceil \log_2\big(\frac{1}{T}\sum_{s=m_t}^t Q_s(p_s)\big) \rceil - 1$, $m_{t+1} \leftarrow t+1$
  else
    $p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta_{r_t} \big\langle \sum_{s=m_t}^t \widehat{y}_s, p \big\rangle + \psi_{q_{r_t}}(p)$
    $r_{t+1} \leftarrow r_t$, $m_{t+1} \leftarrow m_t$
  end if
end for

in Chapter 3 for time-varying graphs). Compared to the latter, we require a more refined analysis to account for the case when $\bar{C}_T < 1$. This refinement is achieved in part via the use of Lemma C.1 (found in Appendix C.1), which also allows adapting the analysis of Theorem 2.1 to account for the fact that we use the q-Tsallis entropy as a regularizer in place of the Shannon entropy.

Theorem 4.2. *Assuming that $T \geq \log(e^2 N)$, Algorithm 4.2 run with input $J \in (0, N]$ satisfies*

$$R_T \leq 38e\sqrt{(\bar{C}_T \vee J)\, T \log\left(\frac{e^2 N}{\bar{C}_T \vee J \vee 1}\right)} + \log_2\left(\frac{\bar{C}_T}{J}\right)_+ + \frac{18e}{5}\log_2\left(\frac{4\Big(\big(JT \vee \bar{C}_T T\big) \wedge \log(e^2 N)\Big)}{JT}\right)_+ \log\big(e^2 N\big) + 1\,.$$

In particular, setting $J = \log(e^2 N)/T$ yields that

$$R_T \le 38e\sqrt{\bar{C}_T T \log\left(\frac{e^2 N}{\bar{C}_T \vee 1}\right)} + \log_2\left(\frac{\bar{C}_T T}{\log(e^2 N)}\right)_+ + 46e\log\big(e^2 N\big) + 1\,.$$

Proof. For brevity, we define $U := \bar{C}_T \vee J$. Let $s := \lceil \log_2 J \rceil - 1$ and $n := \lceil \log_2 U \rceil - 1$, the latter of which is the largest value that r_t can take, since for any round t,

$$\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le \frac{1}{T}\sum_{s=1}^{T} Q_s(p_s) \le \frac{1}{T}\sum_{s=1}^{T} C_s \le 2^{n+1}\,.$$

Without loss of generality, we assume that for any (integer) $r \in \{s,\dots,n\}$, there are at least two rounds in which $r_t = r$, and we use T_r to refer to the index of the first such round. Additionally, we define $T_{n+1} := T+2$. Note that for any r in this range, $q_r \in [1/2, 1)$. Let $i^* \in \operatorname{argmin}_{i\in V} \sum_{t=1}^{T} y_t(i)$. We start by decomposing the regret over the intervals corresponding to fixed values of $r_t \in \{s,\dots,n\}$ and bounding the instantaneous regret at the last step of each but the last interval by 1:

$$\begin{aligned} R_T &= \mathbb{E}\Big[\sum_{t=1}^{T}\big(y_t(I_t) - y_t(i^*)\big)\Big] \\ &\le \mathbb{E}\Big[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\Big] + n - s \\ &\le \mathbb{E}\Big[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\Big] + \log_2\big(U/J\big) + 1\,. \end{aligned} \tag{4.3}$$

Let $\mathbf{e}_{i^*} \in \mathbb{R}^N$ be the indicator vector for i^* and define (similarly to the proof of Theorem 2.1) $\widetilde{y}_t \in \mathbb{R}^N$ where $\widetilde{y}_t(i) := \widehat{y}_t(i) - \ell_t(A_t)$ for every $i \in V$. Similarly to the proof of Theorem 3.4,

we note that for each $r \in \{s, \dots, n\}$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \big(y_t(I_t) - y_t(i^*)\big)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\Big\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\Big\}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\Big\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\Big\}\langle p_t - \mathbf{e}_{i^*}, \widehat{y}_t\rangle\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\Big\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\Big\}\langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t\rangle\right]
\end{aligned}$$

where (a) follows since $\mathbb{E}_t\big[y_t(I_t)\big] = \sum_{i\in V} p_t(i)y_t(i)$, $\mathbb{E}_t\big[\widehat{y}_t\big] = y_t$, and the indicator at round t is measurable with respect to $\mathcal{F}_{t-1}$ (where $\mathcal{F}_{t-1}$ and $\mathbb{E}_t$ are defined in the same way as in the proof of Theorem 4.1); and (b) follows since $p_t, \mathbf{e}_{i^*} \in \Delta_N$ and $\widehat{y}_t(i) - \widetilde{y}_t(i) = \ell_t(A_t)$ is identical for all $i \in V$. Similarly to the last argument, the fact that $\big\langle \widetilde{y}_s - \widehat{y}_s, p - q\big\rangle = 0$ holds for any $p, q \in \Delta_N$ at any round s implies that p_{t+1} can be equivalently defined as $\operatorname{argmin}_{p\in\Delta_N} \eta_{r_t}\big\langle \sum_{s=m_t}^{t} \widetilde{y}_s, p\big\rangle + \psi_{q_{r_t}}(p)$. Hence, using that $\widetilde{y}_t(i) \ge -1$, we can invoke Lemma C.1 (with $b = 1$ and $c = e$) to obtain that

$$\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t\rangle \le \frac{N^{1-q_r} - 1}{(1-q_r)\eta_r} + \frac{e\eta_r}{2q_r} \sum_{t=T_r}^{T_{r+1}-2} \sum_{i\in V} p_t(i)^{2-q_r}\, \widetilde{y}_t(i)^2 .$$

For any round $t \in [T]$ and action $a \in \mathcal{A}$, recall the definition $\phi_t(a) \coloneqq \sum_{i\in V} p_t(i)\theta_t^i(a)$. Similarly to the proof of Theorem 2.1, we have that

$$\begin{aligned}
\mathbb{E}_t\left[\widetilde{y}_t(i)^2\right] &= \mathbb{E}_t\left[\ell_t(A_t)^2 \frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&\le \mathbb{E}_t\left[\frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&= \sum_{a\in\mathcal{A}} \frac{\big(\theta_t^i(a) - \phi_t(a)\big)^2}{\phi_t(a)} = \sum_{a\in\mathcal{A}} \phi_t(a)\Big(\frac{\theta_t^i(a)}{\phi_t(a)} - 1\Big)^2 = \chi^2(\theta_t^i \,\|\, \phi_t) .
\end{aligned}$$

Hence, for any round t and any $r \in \{s, \dots, n\}$, it holds that

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{y}_t(i)^2\right] &\le \sum_{i\in V} p_t(i)^{2-q_r}\chi^2(\theta_t^i \,\|\, \phi_t) \\
&= Q_t(p_t)\sum_{i\in V}\frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)}p_t(i)^{1-q_r} \\
&\le Q_t(p_t)\left(\sum_{i\in V}\frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)}p_t(i)\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r}\left(\sum_{i\in V} p_t(i)^2\chi^2(\theta_t^i \,\|\, \phi_t)\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r}\left(\sum_{i\in V} p_t(i)^2\sum_{a\in\mathcal{A}}\frac{\theta_t^i(a)^2}{\phi_t(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r}\left(\sum_{a\in\mathcal{A}}\frac{\sum_{i\in V} p_t(i)^2\theta_t^i(a)^2}{\sum_{j\in V} p_t(j)\theta_t^j(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&\le Q_t(p_t)^{q_r}\left(\sum_{a\in\mathcal{A}}\sum_{i\in V} p_t(i)\theta_t^i(a) - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r}\left(1 - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \le Q_t(p_t)^{q_r},
\end{aligned}$$

where the second inequality follows from the definition of $Q_t(p_t)$ and the fact that x^{1-q_r} is concave in $x \ge 0$, and the third inequality uses the superadditivity of x^2 for non-negative real numbers and the non-negativity of the quantity in brackets. Let $T_{r:r+1} \coloneqq T_{r+1} - T_r - 1$, it then holds that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{y}_t(i)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\Big\}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{y}_t(i)^2\right] \\
&\le \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)^{q_r}\right] \\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{1}{T_{r:r+1}}\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)\right)^{q_r}\right] \\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{T}{T_{r:r+1}}2^{r+1}\right)^{q_r}\right] \le 2T\left(2^r\right)^{q_r},
\end{aligned}$$

where the second inequality uses the concavity of x^{q_r} in $x \ge 0$ and the third inequality uses that $(1/T)\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t) \le 2^{r+1}$ since the algorithm is not reset in the interval $[T_r, T_{r+1}-2]$.

Overall, we have shown that

$$\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\left(y_t(I_t)-y_t(i^*)\right)\right] \le \frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T\,.$$

If $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} \le \frac{q_r}{1-q_r}\left(1-e^{\frac{q_r-1}{2-q_r}}\right)$, then substituting the values of η_r and q_r gives that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T &= 2\sqrt{\frac{e(N^{1-q_r}-1)\,(2^r)^{q_r}\,T}{q_r(1-q_r)}} \\
&= 2\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{\frac{eN^{1-q_r}\,(2^r)^{q_r}\,T}{q_r(1-q_r)}} \\
&\le 2e\sqrt{2}\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{2^r\left(2+\log(N2^{-r})\right)T} \\
&\le 2e\sqrt{2}\left(\sqrt{\frac{\log N}{\log(N2^{-r})}}\wedge 1\right)\sqrt{2^r\left(2+\log(N2^{-r})\right)T} \\
&= 2e\sqrt{2}\sqrt{2^r\log\left(e^2N(2^{-r}\wedge 1)\right)T}\,,
\end{aligned}$$

where the first inequality holds via the same arguments laid in the last passage of the proof of Theorem 4.1, and the second inequality holds since

$$\begin{aligned}
\frac{N^{1-q_r}-1}{N^{1-q_r}} &= 1-\exp\left(-\log\left(N^{1-q_r}\right)\right) \\
&\le (1-q_r)\log N \\
&= \frac{1}{2}\left(1-\frac{\log(N/2^r)}{\sqrt{\log(N/2^r)^2+4}+2}\right)\log N \\
&= \frac{\log N}{2\log(N/2^r)}\left(2+\log(N/2^r)-\sqrt{\log(N/2^r)^2+4}\right) \le \frac{\log N}{\log(N/2^r)}\,,
\end{aligned}$$

where the inequality follows from the fact that $1-e^{-x} \le x$. Otherwise, if $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} > \frac{q_r}{1-q_r}\left(1-e^{\frac{q_r-1}{2-q_r}}\right)$, then η_r takes the latter value and we obtain that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T &\le \frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \eta_r \frac{N^{1-q_r}-1}{(1-q_r)} \left(\frac{1-q_r}{q_r\left(1-e^{\frac{q_r-1}{2-q_r}}\right)}\right)^2 \\
&= 2\frac{N^{1-q_r}-1}{q_r\left(1-e^{\frac{q_r-1}{2-q_r}}\right)} \\
&\le \frac{18\left(N^{1-q_r}-1\right)}{5q_r(1-q_r)} \\
&= \frac{18\,(2^r)^{-q_r}\left(N^{1-q_r}-1\right)(2^r)^{q_r}}{5q_r(1-q_r)} \\
&\le \frac{18e}{5}\left(2^r\right)^{1-q_r}\log\left(e^2N(2^{-r}\wedge 1)\right) \\
&\le \frac{18e}{5}\left(1\vee\sqrt{2^r}\right)\log\left(e^2N(2^{-r}\wedge 1)\right),
\end{aligned}$$

where the last inequality holds since $q_r \ge 1/2$, and the second inequality holds since

$$1-e^{\frac{q_r-1}{2-q_r}} \ge \frac{1-q_r}{2-q_r} - \frac{1}{2}\left(\frac{1-q_r}{2-q_r}\right)^2 = \frac{3-q_r}{2(2-q_r)^2}(1-q_r) \ge \frac{5}{9}(1-q_r)\log\left(e^2N(2^{-r}\wedge 1)\right),$$

where the first step uses that $e^{-x} \le 1-x+x^2/2$ for $x \ge 0$, and the last step uses again that $q_r \ge 1/2$. Hence, the results above yield that

$$\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\left(y_t(I_t)-y_t(i^*)\right)\right] \le \max\Big\{2e\sqrt{2}\sqrt{2^rT\log\left(e^2N(2^{-r}\wedge 1)\right)}, \frac{18e}{5}\left(1\vee\sqrt{2^r}\right)\log\left(e^2N(2^{-r}\wedge 1)\right)\Big\}. \quad (4.4)$$

Let $M := \log(e^2 N)/T$ and $m := \log_2 M$, and note that $m \le 0$ (and $M \le 1$) by the assumption that $T \ge \log(e^2 N)$. In the case when $n \le 0$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\qquad\le \frac{18e}{5}\big((n \wedge \lfloor m \rfloor) - s + 1\big)_+ \log\big(e^2 N\big) + 2e\sqrt{2}\sum_{r=n\wedge\lceil m\rceil}^{n}\sqrt{2^r T \log\big(e^2 N\big)} \\
&\qquad\le \frac{18e}{5}\log_2\big(4(U \wedge M)/J\big)_+ \log\big(e^2 N\big) + 8e\sqrt{2UT\log\big(e^2 N\big)},
\end{aligned}$$

where the second inequality uses that

$$\sum_{r=\alpha}^{n}\big(\sqrt{2}\big)^r = \big(\sqrt{2}\big)^\alpha \sum_{r=0}^{n-\alpha}\big(\sqrt{2}\big)^r = \big(\sqrt{2}\big)^\alpha \frac{\big(\sqrt{2}\big)^{n-\alpha+1} - 1}{\sqrt{2}-1} \le \frac{\sqrt{2}}{\sqrt{2}-1}\big(\sqrt{2}\big)^n \le 4\sqrt{U},$$

with $\alpha := n \wedge \lceil m \rceil$. Otherwise, if $n > 0$, then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\quad\le \frac{18e}{5}\log_2(4M/J)_+ \log\big(e^2 N\big) + 8e\sqrt{2T\log\big(e^2 N\big)} + \mathbb{E}\left[\sum_{r=s_+}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\quad\le \frac{18e}{5}\log_2(4M/J)_+ \log\big(e^2 N\big) + 8e\sqrt{2T\log\big(e^2 N\big)} + \frac{18e}{5}\sum_{r=0}^{n}\sqrt{2^r \log\big(e^2 N 2^{-r}\big)T} \\
&\quad\le \frac{18e}{5}\log_2(4M/J)_+ \log\big(e^2 N\big) + 8e\sqrt{2T\log\big(e^2 N\big)} + 26e\sqrt{UT\log\big(e^2 N/U\big)} \\
&\quad\le \frac{18e}{5}\log_2(4M/J)_+ \log\big(e^2 N\big) + 38e\sqrt{UT\log\big(e^2 N/U\big)},
\end{aligned}$$

where the first inequality follows from the analysis of the first case with $n = 0$, the second inequality uses that $r \ge 0$ and the assumption that $T \ge \log(e^2 N)$, the third inequality uses Lemma B.3, and the fourth uses that $x\log(e^2 N/x)$ is increasing in $[0, eN]$ and that $U \ge 2$ in this case. The theorem then follows by combining the bounds provided for the two cases with (4.3). □

4.5 A Lower Bound for Restricted Expert Advice via Feedback Graphs

In this section, we provide a novel lower bound on the minimax regret for a slightly harder formulation of the multi-armed bandit problem with expert advice. We consider a setting where the learner picks an expert I_t (possibly at random) at the beginning of each round $t \in [T]$ without observing any of the experts' recommendations beforehand. The action A_t to be executed is subsequently drawn by the environment from the chosen expert's distribution, i.e., $A_t \sim \theta_t^{I_t}$. Afterwards, the learner observes A_t, the incurred loss $\ell_t(A_t)$, and the advice θ_t^i only of experts $i \in V$ that have the drawn action A_t in their support, i.e., $\theta_t^i(A_t) > 0$. For experts outside this set, the learner can only infer that, by definition, $\theta_t^i(A_t) = 0$. We will refer to this variation of the problem as the multi-armed bandit with *restricted* expert advice (note that this differs from the limited expert advice model studied by Kale 2014). Observe that Algorithm 4.1 is still implementable in this scenario and guarantees a regret upper bound of order $\sqrt{\xi T\left(1+\log(N/\xi)\right)}$ for $\xi := K \wedge N$, as previously analyzed. Here we show that the regret of Algorithm 4.1 is the best regret we can hope for, up to constant factors, for any number K of actions and any number N of experts. While a $\Omega(\sqrt{NT})$ regret lower bound in the case $N \le K$ is immediate (as mentioned before), the following theorem provides an $\Omega\big(\sqrt{KT\log(N/K)}\big)$ lower bound when $N > K$, improving upon the $\Omega\big(\sqrt{KT(\log N)/(\log K)}\big)$ lower bound of Seldin and Lugosi (2016).

In what follows, we fix $N > K \ge 2$. We derive the lower bound relying on a reduction from bandits with graph feedback. To recall, in that problem we assume there exists a graph $G = (V, E)$ over a finite set $V = [N]$ of actions from which the learner selects one action $J_t \in V$ at each round $t \in [T]$. Then, the learner observes the losses of the neighbours of J_t in G. For the construction of the lower bound, it suffices to assume that G is undirected and contains all self-loops, i.e., $(i, i) \in E$ for each $i \in V$. Consequently, the learner always observes the loss of the selected action and the graph G is strongly observable—see (Alon et al., 2015) for a classification of feedback graphs. We particularly focus on a specific family of graphs (also considered in the recent work of Chen et al. 2024) where the N vertices are partitioned into disjoint cliques with self-loops. Precisely, we let $M := \lfloor K/2 \rfloor \ge 1$ be the number of disjoint cliques in G. For any $k \in [M]$, let C_k be the set of vertices of the k-th clique in G. Since each C_k is a clique with all self-loops, we have that $(i, j) \in E$ if and only if $i, j \in C_k$ for some $k \in [M]$, and thus $E = \bigcup_{k\in[M]} (C_k \times C_k)$. Additionally, for our purposes, we only consider the partition into cliques $C_k = \{i \in [N] : i \equiv k \mod M\}$ of roughly the same size $|C_k| \ge \lfloor N/M \rfloor \ge \lfloor 2N/K \rfloor \ge N/K$.

Hence, we will focus on the class of instances, denoted by Ξ_{FG}, of the multi-armed bandit problem with feedback graphs where the graph assumes the particular structure described above. In particular, any instance $\mathcal{I} \in \Xi_{\mathrm{FG}}$ is defined as a tuple $\mathcal{I} := (T, G, \mathcal{L})$ containing the number T of rounds, the feedback graph $G = (V, E)$ over $V = [N]$ composed of the disjoint cliques $C_1, \dots, C_M$ as defined above, and the sequence $\mathcal{L} := (\ell_t)_{t\in[T]}$ of binary loss functions $\ell_t \colon V \to \{0,1\}$ over V. On the other hand, we let Ξ_{REA} be the class of instances for the multi-armed bandit problem with restricted expert advice, with N experts and K actions. An instance $\mathcal{I} \in \Xi_{\mathrm{REA}}$ is a tuple $\mathcal{I} := \big(T, V, \mathcal{A}, \Theta, \mathcal{L}\big)$ containing the number T of rounds, the set $V = [N]$ of experts, the set $\mathcal{A} = [K]$ of actions, the sequence $\Theta := (\theta_t^i)_{i\in V, t\in[T]}$ of expert advice where $\theta_t^i \in \Delta_K$, and the sequence $\mathcal{L} := (\ell_t)_{t\in[T]}$ of loss functions $\ell_t \colon \mathcal{A} \to \{0,1\}$ over $\mathcal{A}$. The sought result is established by showing that the worst-case regret of any algorithm against a particular subset of instances in Ξ_{REA} is order-wise at least as large as the minimax regret on Ξ_{FG}, combined with a lower bound on the latter quantity by Chen et al. (2024).

Theorem 4.3. *Let $\mathcal{B}$ be any possibly randomized algorithm for the multi-armed bandit problem with restricted expert advice for any number $K \geq 2$ of actions $\mathcal{A} = [K]$ and any number $N > K$ of experts $V = [N]$. Then, for a sufficiently large T, there exist a sequence $\ell_1, \dots, \ell_T \colon \mathcal{A} \to \{0,1\}$ of binary loss functions and a sequence $(\theta_t^i)_{i\in V, t\in[T]}$ of expert advice such that the expected regret of $\mathcal{B}$ is $\Omega\big(\sqrt{KT\log(N/K)}\big)$.*

Proof. We first describe a reduction from the multi-armed bandit problem with feedback graphs to the multi-armed bandit problem with restricted expert advice. We accomplish this by providing a mapping $\rho \colon \Xi_{\mathrm{FG}} \to \Xi_{\mathrm{REA}}$ from the considered instance class Ξ_{FG} of the former problem to the instance class Ξ_{REA} of the latter.

Consider any instance $\mathcal{I} := (T, G, \mathcal{L}) \in \Xi_{\mathrm{FG}}$ and recall that $G = (V, E)$ is a union of $M = \lfloor K/2 \rfloor$ disjoint cliques $C_1, \dots, C_M$ over $V = [N]$. The mapped instance $\rho(\mathcal{I}) := (T, V, \mathcal{A}, \Theta, \mathcal{L}')$, which belongs to Ξ_{REA}, is defined over the same number of rounds T and an experts set corresponding to the actions V in the original instance $\mathcal{I}$, whose sequence of recommendations is provided by $\Theta = (\theta_t^i)_{i\in V, t\in[T]}$. We first observe that the cardinality of the new action set $\mathcal{A} = [K]$ does relate to the number of cliques M. In particular, considering the partition of experts given by the cliques in G, we also partition the actions (in the expert advice instance $\rho(\mathcal{I})$) by associating 2 actions to each clique. Precisely, for any $k \in [M]$, we associate actions $\mathcal{A}_k := \{2k-1, 2k\}$ to C_k. If K is even, this partitions the entire set of actions $\mathcal{A}$, while it leaves out action K otherwise. We can ignore the latter case and assume K is even without loss of generality, since we can otherwise leave action K outside

of the support of any expert advice $\theta_t^i \in \Delta_K$ in the following construction (thus becoming a spurious action).

Second, we focus on the construction of the loss sequence $\mathcal{L}' := (\ell'_1, \dots, \ell'_T)$. For any $t \in [T]$, we define $\ell'_t \in \{0,1\}^{\mathcal{A}}$ as

$$\ell'_t(2k-1) := 0 \qquad \text{and} \qquad \ell'_t(2k) := 1 \qquad \forall k \in [M]\cdot$$

Finally, we define the sequence of expert advice $(\theta_t^i)_{i \in V, t \in [T]}$ depending on the sequence of losses $\mathcal{L}$ of the starting instance $\mathcal{I}$. For any $t \in [T]$, any $k \in [M]$, and any $i \in C_k$, we define $\theta_t^i \in \Delta_K$ as

$$\theta_t^i := \begin{cases} \delta_{2k-1} & \text{if } \ell_t(i) = 0 \\ \delta_{2k} & \text{if } \ell_t(i) = 1 \end{cases},$$

where $\delta_j \in \Delta_K$ is the Dirac delta at $j \in \mathcal{A}$. This ensures that the loss of expert i at round t, given by $y_t(i) = \sum_{a \in \mathcal{A}} \theta_t^i(a) \ell'_t(a)$ coincides with $\ell_t(i)$, the loss of action i in the original feedback graphs instance at the same round. Moreover, the knowledge of $\ell_t(i)$ suffices to infer θ_t^i.

At this point, given our instance mapping ρ and our algorithm $\mathcal{B}$, we design an algorithm $\mathcal{B}_\rho$ for the class Ξ_{FG}. Consider any instance $\mathcal{I} \in \Xi_{\text{FG}}$. Over the interaction period, the algorithm $\mathcal{B}_\rho$, without requiring prior knowledge of $\mathcal{I}$, maintains a running realization of $\mathcal{B}$ on instance $\rho(\mathcal{I})$. At any round $t \in [T]$, let I_t be the expert selected by algorithm $\mathcal{B}$ in $\rho(\mathcal{I})$, and let $k_t \in [M]$ be the index of the clique I_t belongs to, i.e., $I_t \in C_{k_t}$. Algorithm $\mathcal{B}_\rho$, interacting with the instance $\mathcal{I}$, executes action $J_t = I_t$ provided by $\mathcal{B}$ and observes the losses $(\ell_t(i))_{i \in C_{k_t}}$. Then, thanks to the design of the mapping ρ, $\mathcal{B}_\rho$ can construct and provide $\mathcal{B}$ the feedback it requires and which complies with instance $\rho(\mathcal{I})$. Namely, it determines that $A_t = 2k_t - 1$ if $\ell_t(J_t) = 0$ or else that $A_t = 2k_t$, then passes A_t, its loss $\ell'_t(A_t)$ (trivially determined), and the restricted advice $(\theta_t^i)_{i \in C_{k_t}}$ to $\mathcal{B}$. The last of which is a super-set of the recommended distributions having positive support on A_t since A_t is never picked by experts outside C_{k_t} by construction.

Now, let

$$R^{\mathcal{B}}(\mathcal{I}') := \mathbb{E}\left[\sum_{t=1}^T \ell'_t(A_t)\right] - \min_{i \in V} \sum_{t=1}^T \sum_{a \in \mathcal{A}} \theta_t^i(a) \ell'_t(a) = \mathbb{E}\left[\sum_{t=1}^T y_t(I_t)\right] - \min_{i \in V} \sum_{t=1}^T \sum_{a \in \mathcal{A}} \theta_t^i(a) \ell'_t(a)$$

be the expected regret of algorithm $\mathcal{B}$ on some instance $\mathcal{I}' = \left(T, V, \mathcal{A}, (\theta_t^i)_{i\in V, t\in[T]}, (\ell'_t)_{t\in[T]}\right)$ belonging to Ξ_{REA}. Similarly, let

$$R^{\mathcal{B}_\rho}(\mathcal{I}) := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(J_t)\right] - \min_{i\in V} \sum_{t=1}^{T} \ell_t(i)$$

be the expected regret of algorithm $\mathcal{B}_\rho$ on some instance $\mathcal{I} = \left(T, G, (\ell_t)_{t\in[T]}\right) \in \Xi_{\mathrm{FG}}$. Since $J_t = I_t$, we have that $y_t(I_t) = \ell_t(J_t)$ via the properties of ρ laid out before. Hence, we can conclude that $R^{\mathcal{B}}(\rho(\mathcal{I})) = R^{\mathcal{B}_\rho}(\mathcal{I})$ for any instance $\mathcal{I} \in \Xi_{\mathrm{FG}}$. Define $\rho(\Xi_{\mathrm{FG}}) := \{\rho(\mathcal{I}) : \mathcal{I} \in \Xi_{\mathrm{FG}}\} \subseteq \Xi_{\mathrm{REA}}$ as the subclass of instances in Ξ_{REA} obtained from Ξ_{FG} via ρ. Then, it holds that

$$\sup_{\mathcal{I}\in\Xi_{\mathrm{REA}}} R^{\mathcal{B}}(\mathcal{I}) \geq \sup_{\mathcal{I}\in\rho(\Xi_{\mathrm{FG}})} R^{\mathcal{B}}(\mathcal{I}) = \sup_{\mathcal{I}\in\Xi_{\mathrm{FG}}} R^{\mathcal{B}}(\rho(\mathcal{I})) = \sup_{\mathcal{I}\in\Xi_{\mathrm{FG}}} R^{\mathcal{B}_\rho}(\mathcal{I})\,.$$

On the other hand, Lemma E.1 in (Chen et al., 2024) implies that

$$\sup_{\mathcal{I}\in\Xi_{\mathrm{FG}}} R_T^{\mathcal{B}_\rho}(\mathcal{I}) = \Omega\left(\sqrt{T \sum_{k\in[M]} \log(1+|C_k|)}\right) = \Omega\left(\sqrt{KT\log(N/K)}\right)$$

for sufficiently large T since $\sum_{k\in[M]} \log(1+|C_k|) \geq M\log(N/M) \geq K\log(2N/K)/4$, thus concluding the proof. □

4.6 Conclusion

As the lower bound of Theorem 4.3 was proved for a harder formulation of the problem, it remains to be shown whether the same impossibility result holds for the more standard setup. We conjecture it should be possible to prove such a lower bound. If it indeed holds, this would imply that the minimax regret in the two variants is of the same order; that is, as far as we are only concerned with the worst-case regret, the standard feedback setup would be shown to be essentially as hard as the restricted one. In fact, in a concurrent work, Ito (2024) proved a lower bound of the same order for a less restricted setting by directly adapting the techniques of Chen et al. (2024). In Ito's (2024) setting, and unlike ours, the recommendations of *all* the experts are revealed to the learner at the end of each round; nonetheless, the learner must still choose an expert to follow before receiving the recommendations.

Chapter 5

Online Convex Markov Decision Processes with Bandit Feedback*

5.1 Introduction

Notation We collect here some notation used throughout this chapter. For some integer $d > 0$, let $\mathbb{1}_d \in \mathbb{R}^d$ denote the vector with all entries equal to one, and let $\mathbf{I}_d$ be the $d \times d$ identity matrix. For a finite set E, let $\Delta_E := \Delta_{|E|}$. We denote by $\mathbb{B}^d$ and $\mathbb{S}^d$ the unit Euclidean ball and sphere respectively in $\mathbb{R}^d$.

Let $N, S, A \geq 2$ be three positive integers. Consider an episodic Markov decision process (MDP; Puterman, 2014) characterized by the number of steps N within an episode, a finite state space $\mathcal{X} = [S]$, a finite action space $\mathcal{A} = [A]$, a sequence of loss functions $(\ell_n)_{n\in[N]}$ where $\ell_n \colon \mathcal{X} \times \mathcal{A} \to [0,1]$ evaluates the desirability of executing a certain action in a certain state at step n, an initial state distribution $p_1 \in \Delta_{\mathcal{X}}$, and a sequence of probability transition kernels $(p_n)_{n\in\{2,\dots,N\}}$ where $p_n \colon \mathcal{X}\times\mathcal{X}\times\mathcal{A} \to [0,1]$ is such that $p_n(x'|x,a)$ denotes the probability of transitioning to state x' at step n having executed action a in state x at step $n-1$; hence, $p_n(\cdot|x,a) \in \Delta_{\mathcal{X}}$. Lumping together the initial state distribution and the transition kernels, we refer to $p := (p_n)_{n\in[N]}$ simply as the (transition) dynamics or kernel of the MDP. An agent interacts with the environment using a policy $\pi := (\pi_n)_n$, where $\pi_n \colon \mathcal{A} \times \mathcal{X}$ is such that $\pi_n(a|x) \to [0,1]$ denotes the probability of picking action a at state x when following policy π at the n-th step of the episode; hence, $\pi_n(\cdot|x) \in \Delta_{\mathcal{A}}$. Denote by Π the set of all such policies.

*This chapter is partly based on (Moreno, Eldowa, Gaillard, Brégère, and Oudjane, 2025).

Executing policy π in this environment induces a random trajectory $(x_1, a_1, \ldots, x_N, a_N)$, where $x_1 \sim p_1$, $a_n \sim \pi_n(\cdot|x_n)$ for $1 \le n \le N$, and $x_n \sim p_n(\cdot|x_n, a_n)$ for $2 \le n \le N$. The general goal is to find a policy that minimizes the accumulated losses in an episode;

$$\min_{\pi\in\Pi} \mathbb{E}_\pi \sum_{n=1}^{N} \ell_n(x_n, a_n)\,, \tag{5.1}$$

where the $\mathbb{E}_\pi$ denotes the expectation over the induced trajectory when executing policy π. If the MDP is known to the learner (specifically, the sequences of loss functions and transition kernels), this problem can be solved via dynamic programming (Puterman, 2014). In absence of such knowledge, solving this problem is the subject of reinforcement learning (RL; Bertsekas, 2019; Sutton and Barto, 2018), where one must learn about the environment via repeated interaction.

Another way to phrase this objective is through the notion of occupancy measures. For a fixed MDP with dynamics p, a policy π can be identified via a sequence of probability distributions $\mu^{\pi,p} := (\mu_n^{\pi,p})_{n\in[N]}$ where $\mu_n^{\pi,p} \in \Delta_{\mathcal{X}\times\mathcal{A}}$ is the probability distribution over state-action pairs induced by policy π at step n. This is referred to as the (state-action) occupancy measure induced by policy π. Similarly, its state-occupancy measure is denoted by $\rho^{\pi,p} := (\rho_n^{\pi,p})_{n\in[N]}$ where $\rho_n^{\pi,p} \in \Delta_{\mathcal{X}}$ is the marginal of $\mu_n^{\pi,p}$ over the states; that is,[1] $\rho_n^{\pi,p}(x) = \sum_{a\in\mathcal{A}} \mu_n^{\pi,p}(x,a)$. Note that for any π, $\rho_1^{\pi,p} = p_1$, the fixed initial state distribution of the MDP. In what follows, and whenever clear from the context, we drop p from the superscripts of the symbols defined above. Viewing μ^π collectively as a vector in $(\Delta_{\mathcal{X}\times\mathcal{A}})^N$ and defining $\ell := (\ell_n)_{n\in[N]}$ as a vector in $[0,1]^{NSA}$, we can rewrite the objective in (5.1) as

$$\min_{\pi\in\Pi} \langle \mu^\pi, \ell \rangle\,.$$

This casts the problem as that of minimizing a linear function in the occupancy measure induced by the policy. Starting from this viewpoint, an immediate generalization is to consider an objective that is convex in the occupancy measure. Precisely, let $F\colon (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ be a convex function; the objective becomes

$$\min_{\pi\in\Pi} F(\mu^\pi)\,.$$

This problem has received increasing attention in recent years (Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022; Barakat et al., 2023; Mutti et al., 2023; Moreno et al., 2024a) where it has been studied under the name 'convex Markov deci-

[1] Formally, we use $\mu_n^{\pi,p}(x,a)$ as shorthand for $\mu_n^{\pi,p}((x-1)A+a)$.

sion processes', 'convex reinforcement learning', or 'concave utility reinforcement learning' (CURL). The increased generality of this framework allows it encompass interesting objectives beyond simple reward maximization. One example is that of the entropy maximization objective studied, for example, by Hazan et al. (2019) and Mutti et al. (2022b). There, one adopts a pure exploration objective of maximizing the entropy of the state-occupancy measure: $\min_\pi \sum_{n\in[N]} \langle \rho_n^\pi, \log \rho_n^\pi \rangle$ or $\min_\pi \langle \bar{\rho}_{1:n}^\pi, \log \bar{\rho}_{1:n}^\pi \rangle$, where $\bar{\rho}_{1:n}^\pi := {}^1/_N \sum_{n\in[N]} \rho_n^\pi$. Another example is imitation learning (Abbeel and Ng, 2004; Ghasemipour et al., 2020; Lavington et al., 2022), where the goal is to imitate the behavior of an expert, evading in this way the often difficult task of designing a reward function. The objective in this case can be the minimization of some convex divergence function (e.g., an f-divergence) between the induced occupancy measure and a target one: $\min_\pi \sum_{n\in[N]} D(\mu_n^\pi || \mu_n^*)$. The non-linearity of the objectives in these problems presents new challenges as some classic tools in the reinforcement learning literature, such as Bellman equations, are no longer applicable.

5.1.1 Problem Setting

We study in this chapter an online version of the convex MDP problem. In this setting, a learner interacts in a series of T rounds with an environment characterized by an episodic MDP (as described above) with fixed (and known) transition dynamics $p := (p_n)_{n\in[N]}$ and a series of (apriori unknown) convex objective functions $(F^t)_{t\in[T]}$ acting on occupancy measures, where $F^t \colon (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to [0, N]$. In each round t, the learner chooses a policy π^t (possibly at random) and suffers loss $F^t(\mu^{\pi^t})$. The objective becomes minimizing one's regret with respect to the best fixed policy in hindsight:

$$R_T := \mathbb{E}\left[\sum_{t=1}^{T} F^t(\mu^{\pi^t})\right] - \min_{\pi\in\Pi} \sum_{t=1}^{T} F^t(\mu^\pi)\,. \tag{5.2}$$

We assume the learner only receives bandit feedback, meaning that the value $F^t(\mu^{\pi^t})$ is the only information provided regarding the loss function at round t.

5.1.2 Related Works

In the case when F^t is linear, i.e., the standard RL objective, regret minimization has been studied extensively over the past two decades (e.g., Even-Dar et al., 2009; Jaksch et al., 2010; Zimin and Neu, 2013; Azar et al., 2017; Rosenberg and Mansour, 2019; Jin

et al., 2020; Cai et al., 2020; Luo et al., 2021). This literature covers many settings, drawing distinctions between, for instance, stochastic and adversarial losses; full information and bandit feedback; known and unknown transition dynamics; and between approaches relying on policy updates and those relying on occupancy measure updates.

In the general convex case, less settings seem to have been studied; we are not aware of works that address online learning with bandit feedback. Even in the more common research line concerned with finding a near-optimal policy rather than regret minimization, full knowledge of the objective function is the predominant assumption. The closest settings are those of Chatterji et al. (2021) and Mutti et al. (2022a, 2023), where the loss function is assumed to be an unknown linear function (or specified by a generalized linear model) of a known feature map of the trajectory; hence, the learner has richer knowledge concerning the objective than what we assume here.

These aforementioned works consider regret minimization, but with a fixed stochastic objective. In contrast, only the two works of Moreno et al. (2024a,b) seem to address adversarially changing loss functions, as we do here. Both assume partial knowledge of the MDP's dynamics through a given model, and only deal with the 'full-information' feedback case where $\nabla F^t(\mu^{\pi^t})$ is observed at the end of round t. The first work assumes the transition kernels are fixed and deals with the static regret objective defined above, while the second allows the transition kernels to vary between rounds and considers a notion of dynamic regret (competing against any sequence of policies).

We note at this point that interpreting the requirement of bandit feedback in convex MDPs is less straightforward than in the standard linear objective case. In the latter, bandit feedback usually means that the learner observes the value of the losses $\ell_n(x_n, a_n)$ incurred over all the steps $n \in [N]$ of a sampled trajectory. This amounts to observing n (random) coordinates of the gradient of the linear objective $\langle \cdot, \ell \rangle$. A few works, for example those of Efroni et al. (2021) and Cohen et al. (2021), do consider the more difficult case in which the learner observes only $\sum_{n\in[N]} \ell_n(x_n, a_n)$, the sum of the incurred losses along the sampled trajectory. In either case, the observation depends on a sampled trajectory collected using a certain policy, and not directly on the occupancy measure of the policy; meaning that, in the second case, we do not observe $\langle \mu^\pi, \ell \rangle$ upon executing policy π, but a random variable whose expectation is this quantity, recalling that $\mathbb{E}_\pi \sum_{n\in[N]} \ell_n(x_n, a_n) = \langle \mu^\pi, \ell \rangle$.

On the other hand, $\mathbb{E}_\pi F(\widehat{\mu}^\pi)$, where $\widehat{\mu}^\pi$ is an empirical occupancy measure encoding one sampled trajectory (or the average of a finite number of trajectories) obtained via policy π, does not coincide with $F(\mu^\pi)$ for general convex functions. Our objective being

minimizing functions acting on occupancy measures, we assume direct access to $F(\mu^\pi)$ upon choosing to execute policy π. In terms of practicality, one can motivate this setting by scenarios where the same policy is executed independently by a very large number of agents and the performance of the system (the observed loss) is a function of the collective behavior of the agents (see, e.g., the energy management problem discussed in Moreno et al., 2024a and Coffman et al., 2023).

On a related note, Mutti et al. (2023) argue that in practical scenarios where the learner is evaluated based on their expected performance in one or a handful of episodes, one ought to adopt an objective of the form $\mathbb{E}_\pi F(\widehat{\mu}^\pi)$ in the first place, whose minimizer does not match that of $F(\mu^\pi)$ in general. They show that non-Markovian policies[2] can be necessary to optimize this finite trials objective, which entails an increased computational burden. While interesting, we do not pursue this direction here. Lastly, we mention that here, unlike most related works, we assume knowledge of the transition dynamics of the MDP. This way, we focus solely on addressing the challenges of bandit feedback in convex MDPs. We briefly discuss in Section 5.5 the setting where the transition dynamics must be learned from observed trajectories. Such a setting, however, is perhaps more interesting from a theoretical viewpoint than from a practical one, having argued that the type of feedback we assume here involves the assessment of the average behavior of a selected policy on a very large number of trajectories, which could render estimating the MDP a trivial task if all these trajectories are observed.

5.1.3 Contributions and Road map

The problem setting we consider here can be broadly categorized as a bandit convex optimization (BCO) problem, we refer to the recent notes by Lattimore (2024) for an overview of BCO algorithms and results. We will adopt here one of the simpler techniques in the BCO literature. Popularized by Flaxman et al. (2005), this technique essentially consists in using a standard online convex optimization algorithm (such as mirror descent) while employing at every round a stochastic estimate of the gradient of the loss function at a certain point. These are one-point estimates, constructed by querying the loss function at a single point, usually a random perturbation of the point where we want to estimate the gradient. This method yields $T^{3/4}$ regret under convex and Lipschitz conditions, and is incapable of doing better (Hu et al., 2016). Although more advanced algorithms and analyses achieve $\sqrt{T}$ regret (Bubeck et al., 2021; Fokkema et al., 2024; Hazan and

[2] These are policies whose choice of action at any step depends on the interaction history so far, not just on the current state.

Li, 2016), they are arguably less practical, more complicated, and have worse dimension dependence. Thus, we settle here for the more basic combination of mirror descent and gradient estimation, allowing us nonetheless to obtain the first sub-linear regret bounds for the present setting.

The main challenge in our endeavor here is posed by the peculiar nature of the decision set, effectively the set of valid occupancy measure. In Section 5.2, we mold this set into a more convenient form, readily compatible with the gradient estimation method, and present some related preliminary results. In Section 5.3, as a warm-up, we present a first algorithm and its regret analysis, combining the approach of Moreno et al. (2024a) with a basic gradient estimator. As the regret bound we obtain for this algorithm features a certain undesirable dependence on the structure of the MDP, we remedy this in Section 5.4 by adopting another algorithm based on self-concordant regularization and a more refined gradient estimator. We show that this second algorithm enjoys an improved bound unimpaired by the aforementioned issue. Finally, we mention in Section 5.5 some possible future directions.

5.2 On the Representation of Occupancy Measures

Notation Before proceeding further, we add here an extra notational remark for clarity. Let $\dot{A} \in \{A, A-1\}$. For a vector $\xi \in \mathbb{R}^{NS\dot{A}}$ and $(n,x,a) \in [N] \times [S] \times [\dot{A}]$, we use $\xi_n(x,a)$ as a shorthand for $\xi((n-1)S\dot{A}+(x-1)\dot{A}+a)$. Similarly, let $\ddot{A} \in \{A, A-1\}$. Then, for an $NS\dot{A} \times NS\ddot{A}$ matrix M, we use $M(n,x,a,n',x',a')$ to denote the item in row $(n-1)S\dot{A}+(x-1)\dot{A}+a$ and column $(n'-1)S\ddot{A}+(x'-1)\ddot{A}+a'$ of M.

Fix an MDP with dynamics $p = (p_n)_{n\in[N]}$. The occupancy measures $\mu^{\pi,p}$ and $\rho^{\pi,p}$ induced by a policy $\pi \in \Pi$ can be recursively computed as follows: firstly, we have that $\rho_1^{\pi,p}(x) = p_1(x)$; then, for $n \geq 1$,

$$\mu_n^{\pi,p}(x,a) = \pi_n(a|x)\rho_n^{\pi,p}(x) \qquad \text{and} \qquad \rho_{n+1}^{\pi,p}(x) = \sum_{x',a'} \mu_n^{\pi,p}(x',a')\, p_n(x|x',a')\,.$$

Notice that the state distribution $\rho_{n+1}^{\pi,p}(x)$ at step $n+1$ is determined by the state-action distribution $\mu_n^{\pi,p}(x)$ at the previous step and the transition kernel p_{n+1}. Consider then the

set

$$\mathcal{M}^p := \Big\{\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N : \sum_a \mu_n(x,a) = \sum_{x',a'} \mu_{n-1}(x',a')\, p_n(x|x',a')\ \forall x \in \mathcal{X}, 2 \le n \le N \\ \text{and } \sum_a \mu_1(x,a) = p_1(x)\ \forall x \in \mathcal{X}\Big\},$$

which can be easily verified to be convex. Clearly $\mu^{\pi,p} \in \mathcal{M}^p$ for any $\pi \in \Pi$, recalling that $\rho_n^{\pi,p}(x) = \sum_a \mu_n^{\pi,p}(x,a)$. Conversely, for any $\mu \in \mathcal{M}^p$, one can find some $\pi \in \Pi$ such that $\mu^{\pi,p} = \mu$ simply by setting $\pi_n(a|x) = \mu_n(x,a)/\sum_{a'} \mu_n(x,a')$ when $\sum_{a'} \mu_n(x,a') \neq 0$ and choosing some arbitrary $\pi_n(\cdot|x) \in \Delta_{\mathcal{A}}$ otherwise. Hence, $\mathcal{M}^p = \{\mu^{\pi,p} : \pi \in \Pi\}$. This equivalence implies that we can rewrite the regret as

$$R_T = \mathbb{E}\left[\sum_{t=1}^T F^t\big(\mu^{\pi^t,p}\big)\right] - \min_{\mu\in\mathcal{M}^p} \sum_{t=1}^T F^t(\mu)\,.$$

Viewing $\mathcal{M}^p$ as our decision set readily places the tools of online convex optimization at our disposal, since $F^t(\mu^{\pi,p})$ is convex as a function of $\mu^{\pi,p}$, but not necessarily as a function of π (Hazan et al., 2019).

One obstacle, however, is that the gradient estimation techniques that we will adopt from the bandit convex optimization literature require that the decision set has a non-empty interior. This is because, roughly speaking, the gradient at some point μ is estimated by querying the objective at a nearby perturbation $\mu+\delta\boldsymbol{u}$, where δ is a scale parameter and $\boldsymbol{u}$ is a randomly sampled vector from the unit sphere; hence, ensuring the perturbed point is within the decision set requires it to have a non-empty interior. In our case, $\mathcal{M}^p$, as a subset of $\mathbb{R}^{NSA}$, clearly has an empty interior much in the same way that the probability simplex Δ_d has an empty interior in $\mathbb{R}^d$. Note that we require the queried point to belong to $\mathcal{M}^p$ in order to adhere to the interaction protocol of the problem; nevertheless, even its superset $(\Delta_{\mathcal{X}\times\mathcal{A}})^N$, which we take as the domain of the objective functions $(F^t)_{t\in[T]}$, has an empty interior in $\mathbb{R}^{NSA}$.

The solution we adopt here is to map $\mathcal{M}^p$ to a lower-dimensional space where it can have a non-empty interior. This is analogous to how one can define a bijective map between the simplex Δ_d and the set $\{x \in \mathbb{R}^{d-1} : \mathbb{1}_{d-1}^\intercal x \le 1 \text{ and } x_i \ge 0\ \forall i \in [d-1]\}$, which is the intersection of the positive orthant of R^{d-1} with the L_1 unit ball, see (Jézéquel et al., 2022, Section 2). This can be done since any coordinate x_{i^*} of a vector $x \in \Delta_d$ can be recovered from the rest of the coordinates: $x_{i^*} = 1 - \sum_{i\neq i^*} x_i$. In our case, denoting by a^* the last action in $\mathcal{A}$ (i.e., $a^* = A$, recalling that $\mathcal{A} = [A]$), we will represent the

occupancy measures as vectors in $\mathbb{R}^{NS(A-1)}$ by omitting all coordinates that correspond to this action. We can afford to do so, since for any $\mu \in \mathcal{M}^p$, we have that $\mu_n(x,a^*) = \rho^\mu_n(x) - \sum_{a \neq a^*} \mu_n(x,a)$, where we define $\rho^\mu_n(\cdot) \in \Delta_{\mathcal{X}}$ as the state distribution corresponding to the occupancy measure μ at step n: $\rho^\mu_n(x) \coloneqq \sum_a \mu_n(x,a)$, which at any step $n \geq 2$ is recoverable from the previous step: $\rho^\mu_n(x) = \sum_{x',a'} \mu_{n-1}(x',a') p_n(x|x',a')$, and is given at the first step by the initial state distribution: $\rho^\mu_1(x) = p_1(x)$. In the following, we use this idea to define the sought mapping.

Define the $A \times (A-1)$ matrix

$$G \coloneqq \begin{bmatrix} \mathbf{I}_{A-1} \\ -\mathbb{1}^\intercal_{A-1} \end{bmatrix}$$

and let H be the $NSA \times NS(A-1)$ matrix obtained via taking the direct sum of NS copies of G: $H \coloneqq \bigoplus_{i=1}^{NS} G$.[3] Define $\boldsymbol{w}^{p,1} \in \mathbb{R}^{NSA}$ such that

$$\boldsymbol{w}^{p,1}_n(x,a) \coloneqq p_1(x) \mathbb{I}\{n=1, a=a^*\}.$$

Next, for every $2 \leq m \leq N$, we define $W^{p,m}$ as the $NSA \times NSA$ matrix where

$$W^{p,m}(n,x,a,n',x',a') \coloneqq \mathbb{I}\{n=m, n'=m-1, a=a^*\} p_m(x|x',a').$$

Then, we define the $NSA \times NS(A-1)$ matrix

$$B^p \coloneqq (\mathbf{I}_{NSA} + W^{p,N}) \dots (\mathbf{I}_{NSA} + W^{p,3})(\mathbf{I}_{NSA} + W^{p,2}) H$$

and the vector

$$\boldsymbol{\beta}^p \coloneqq (\mathbf{I}_{NSA} + W^{p,N}) \dots (\mathbf{I}_{NSA} + W^{p,3})(\mathbf{I}_{NSA} + W^{p,2}) \boldsymbol{w}^{p,1}.$$

Finally, define the function $\Xi_p \colon \mathbb{R}^{NS(A-1)} \to \mathbb{R}^{NSA}$ where

$$\Xi_p(\xi) \coloneqq B^p \xi + \boldsymbol{\beta}^p$$

for $\xi \in \mathbb{R}^{NS(A-1)}$.

To explain the semantics of Ξ_p, let $\mu \in \mathcal{M}^p$ and $\widetilde{\mu} \in \mathbb{R}^{NS(A-1)}$ be such that $\widetilde{\mu}_n(x,a) \coloneqq \mu_n(x,a)$ for all $(n,x,a) \in [N] \times \mathcal{X} \times \mathcal{A} \setminus a^*$. It then holds that $\Xi_p(\widetilde{\mu}) = \mu$. To see this, note

[3] For an $n \times m$ matrix M and an $n' \times m'$ matrix M', $M \oplus M'$ is the $(n+n') \times (m+m')$ block matrix $\begin{bmatrix} M & \mathbf{0} \\ \mathbf{0} & M' \end{bmatrix}$.

that $H\widetilde{\mu}$ expands $\widetilde{\mu}$ setting $(H\widetilde{\mu})_n(x,a^*) = -\sum_{a\neq a^*}\mu_n(x,a)$. To fully recover $\mu_n(x,a^*)$, what remains is to add $\rho_n^\mu(x)$. This is achieved at $n=1$ by adding $\boldsymbol{w}^{p,1}$ to $H\widetilde{\mu}$ since $\boldsymbol{w}_n^{p,1}(x,a^*) = p_1(x) = \rho_1^\mu(x)$ and $\boldsymbol{w}_n^{p,1}(x,a) = 0$ for $a \neq a^*$. Next, at $n=2$, the matrix $W^{p,2}$ extracts the values $\rho_2^\mu(x)$ when operated on $H\widetilde{\mu} + \boldsymbol{w}^{p,1}$ such that $\mu_2(x,a^*)$ is recovered at coordinate $(2,x,a^*)$ of $(\mathbf{I}_{NSA} + W^{p,2})(H\widetilde{\mu} + \boldsymbol{w}^{p,1})$. Iterating this procedure until step N allows us to fully recover μ from $\widetilde{\mu}$. For any generic $\xi \in \mathbb{R}^{NS(A-1)}$, $\Xi_p(\xi)$ is the unique vector in $\mathbb{R}^{NSA}$ satisfying $(\Xi_p(\xi))_n(x,a) = \xi_n(x,a)$ for all n,x, and $a \neq a^*$; $(\Xi_p(\xi))_1(x,a^*) = p_1(x) - \sum_{a\neq a^*}(\Xi_p(\xi))_1(x,a)$ for all x; and $(\Xi_p(\xi))_n(x,a^*) = \sum_{x',a'}(\Xi_p(\xi))_{n-1}(x',a')p_n(x|x',a') - \sum_{a\neq a^*}(\Xi_p(\xi))_n(x,a)$ for all x and $n \geq 2$.

Note that B^p has full column rank since for any $\xi \in \mathbb{R}^{NS(A-1)}$, $B^p\xi$ is only an expansion of ξ; hence, we can define its left pseudo-inverse $(B^p)^+ := ((B^p)^\intercal B^p)^{-1}(B^p)^\intercal$, which satisfies $(B^p)^+ B^p = \mathbf{I}_{NS(A-1)}$. On the other hand, the matrix $B^p(B^p)^+$ projects vectors in $\mathbb{R}^{NSA}$ onto the column space of B^p, which is given by

$$\Big\{\mu \in \mathbb{R}^{NSA} : \sum_a \mu_n(x,a) = \sum_{x',a'} \mu_{n-1}(x',a')p_n(x|x',a')\ \forall x \in \mathcal{X}, 2 \leq n \leq N \\ \text{and } \sum_a \mu_1(x,a) = 0\ \forall x \in \mathcal{X}\Big\}. \quad (5.3)$$

It is easy to verify that for any $\mu, \mu' \in \mathcal{M}^p$, $\mu - \mu'$ lies in the column space of B^p (recall that $\sum_a \mu_1(x,a) = \sum_a \mu_1'(x,a) = p_1(x)$). Moreover, $\boldsymbol{\beta}^p \in \mathcal{M}^p$ as it corresponds to a policy π where $\pi_n(a^*|x) = 1$ for all n and x. Therefore, for any $\mu \in \mathcal{M}^p$, $\mu - \boldsymbol{\beta}^p$ belongs to the column space of B^p, and we consequently have that

$$\Xi_p\big((B^p)^+(\mu - \boldsymbol{\beta}^p)\big) = B^p(B^p)^+(\mu - \boldsymbol{\beta}^p) + \boldsymbol{\beta}^p = \mu.$$

Hence, by the definition of Ξ_p, $(B^p)^+(\mu - \boldsymbol{\beta}^p)$ coincides with μ on all coordinates $(n,x,a) \in [N] \times \mathcal{X} \times \mathcal{A} \setminus \{a^*\}$ (since the map Ξ_p only expands the input vector adding the coordinates corresponding to action a^*), and is then the only point in $\mathbb{R}^{NS(A-1)}$ that Ξ_p maps to μ. In light of this, we define

$$(\mathcal{M}^p)^- := \{\xi \in \mathbb{R}^{NS(A-1)} \mid \Xi_p(\xi) \in \mathcal{M}^p\},$$

the pre-image of $\mathcal{M}^p$ under Ξ_p. Accordingly, we define $\Lambda_p \colon (\mathcal{M}^p)^- \to \mathcal{M}^p$ as the restriction of Ξ_p to $(\mathcal{M}^p)^-$; that is,

$$\Lambda_p := \Xi_p|_{(\mathcal{M}^p)^-}.$$

This then is a bijective function, with $\Lambda_p^{-1}(\mu) = (B^p)^+(\mu - \boldsymbol{\beta}^p)$.

Still, $(\mathcal{M}^p)^-$ is not guaranteed to have a non-empty interior. Suppose that some state x^* is not reachable at a certain step n^*; that is, for every state x and action a, $p_{n^*}(x^*|x,a) = 0$ if $n^* \geq 2$, or just that $p_1(x^*) = 0$ if $n^* = 1$. Then, for any $\mu \in \mathcal{M}^p$, $\mu_{n^*}(x^*,a) = 0$ for every action a. This implies that for every $\xi \in (\mathcal{M}^p)^-$, $\xi_{n^*}(x^*,a) = 0$ for all $a \neq a^*$ (since these coordinates are preserved under Λ_p), and hence, $(\mathcal{M}^p)^-$ has an empty interior. To remedy this, we introduce the following condition.

Definition 5.1. *We say that the states of an MDP with dynamics p are uniformly reachable if for every state $x \in \mathcal{X}$ and step $n \in [N]$, there exists a policy $\pi \in \Pi$ such that $\rho_n^\pi(x) > 0$.*

This is equivalent to imposing that for every state x, $p_1(x) > 0$ and there exists for every step n a state-action pair (x',a') such that $p_{n+1}(x|x',a') > 0$. Note that this condition can be imposed without loss of generality. Defining $\mathcal{X}_n \subseteq \mathcal{X}$ as the subset of states reachable at step n, one can represent occupancy measures as sequences of distributions in $(\Delta_{\mathcal{X}_n \times \mathcal{A}})_{n\in[N]}$. For simplicity, however, we assume in the coming sections that the uniform reachability condition holds. We show next that this condition is sufficient for $(\mathcal{M}^p)^-$ to have a non-empty interior. We first present an alternative characterization of $(\mathcal{M}^p)^-$.

Lemma 5.2. *For any MDP with dynamics p, it holds that*

$$(\mathcal{M}^p)^- = \left\{\xi \in \mathbb{R}^{NS(A-1)} \colon B^p \xi \geq -\boldsymbol{\beta}^p\right\}.$$

Hence, $(\mathcal{M}^p)^-$ is a polyhedral set formed by NSA constraints; namely that for $n,x,a \in [N] \times \mathcal{X} \times \mathcal{A}$, $B^p(n,x,a,\cdot,\cdot,\cdot)^\intercal \xi + \boldsymbol{\beta}_n^p(x,a) \geq 0$.

Proof. Any $\xi \in (\mathcal{M}^p)^-$ clearly satisfies $B^p\xi \geq -\boldsymbol{\beta}^p$ since $B^p\xi + \boldsymbol{\beta}^p = \Lambda_p(\xi) \in \mathcal{M}^p$, whose coordinates are non-negative. Conversely, assume that $B^p\xi \geq -\boldsymbol{\beta}^p$ for some $\xi \in \mathbb{R}^{NS(A-1)}$, and let $\mu \coloneqq \Xi_p(\xi) = B^p\xi + \boldsymbol{\beta}^p$. Showing that $\xi \in (\mathcal{M}^p)^-$ is equivalent, by definition, to showing that $\Xi_p(\xi) \in \mathcal{M}^p$. Since $\boldsymbol{\beta}^p \in \mathcal{M}^p$ and $B^p\xi$ belongs to the column space of B^p specified in (5.3), it holds that

$$\sum_a \mu_n(x,a) = \sum_{x',a'} \mu_{n-1}(x',a')\, p_n(x|x',a')$$

for every $n \geq 2$, and that $\sum_a \mu_1(x,a) = \sum_a \boldsymbol{\beta}_1^p(x,a) = \rho_1^{\boldsymbol{\beta}^p}(x) = p_1(x)$. Then, to show that $\mu \in \mathcal{M}^p$, it remains to show that $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$. By assumption, μ only has non-negative coordinates; therefore, we only have to show that $\sum_{x,a} \mu_n(x,a) = 1$ at every n. This easily

done via induction: $\sum_{x,a}\mu_1(x,a)=\sum_x p_1(x)=1$, and for $n\geq 2$,

$$\sum_{x,a}\mu_n(x,a)=\sum_x\sum_{x',a'}\mu_{n-1}(x',a')p_n(x|x',a')=\sum_{x',a'}\mu_{n-1}(x',a')\,.$$

□

Lemma 5.3. *For an MDP with dynamics p, $(\mathcal{M}^p)^-$ has a non-empty interior if and only if the MDP's states are uniformly reachable.*

Proof. Necessity is immediate as argued before. We prove sufficiency utilizing an argument from the proof of Proposition 2.3 in (Wolsey and Nemhauser, 1999). For every step-state-action triple (n,x,a), it is easy to verify that uniform reachability implies the existence of some $\mu\in\mathcal{M}^p$ such that $\mu_n(x,a)>0$. Taking a convex combination with full support of one such occupancy measure for every (n,x,a) results, via the convexity of $\mathcal{M}^p$, in an occupancy measure $\mu^*\in\mathcal{M}^p$ whose entries are all strictly positive. Hence, $\xi^*:=\Lambda_p^{-1}(\mu^*)$ is an interior point of the polyhedral set $(\mathcal{M}^p)^-$ as it satisfies with strict inequality all the constraints defining it. □

To be used in the next section, we now explicitly construct an $NS(A-1)$-dimensional ellipsoid that fits inside $(\mathcal{M}^p)^-$ assuming uniform reachability. For every state x, define

$$\varepsilon_1^p(x):=p_1(x)\qquad\text{and}\qquad\varepsilon_n^p(x):=\sum_{x',a'}\frac{\varepsilon_{n-1}^p(x')}{2A-1+\sqrt{A-1}}p_n(x|x',a')\text{ for }2\leq n\leq N\,.\tag{5.4}$$

A simple induction argument shows that $\varepsilon_n^p(x)>0$ for all n and x, using that $p_1(x)>0$ and that there exists a state-action pair (x',a') such that $p_n(x|x',a')>0$. Let $\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}$ (the unit Euclidean ball in $\mathbb{R}^{NS(A-1)}$), and define $\xi^{\boldsymbol{\nu},p}\in\mathbb{R}^{NS(A-1)}$ such that

$$\xi_n^{\boldsymbol{\nu},p}(x,a):=\frac{\varepsilon_n^p(x)(1+\boldsymbol{\nu}_n(x,a)/2)}{A-1/2+1/2\sqrt{A-1}}\,.\tag{5.5}$$

Now let $\zeta^{\boldsymbol{\nu},p}:=\Xi_p(\xi^{\boldsymbol{\nu},p})=B^p\xi^{\boldsymbol{\nu},p}+\boldsymbol{\beta}^p$, we show that all the coordinates of $\lambda^{p,\boldsymbol{\nu}}$ are non-negative, implying via Lemma 5.2 that $\xi^{\boldsymbol{\nu},p}\in(\mathcal{M}^p)^-$. Specifically, we will show that $\zeta_n^{\boldsymbol{\nu},p}(x,a)\geq\frac{\varepsilon_n^p(x)}{2A-1+\sqrt{A-1}}$ for every (n,x,a), which is strictly positive. For $a\neq a^*$, since $\boldsymbol{\nu}_n(x,a)\geq-1$ and $\xi^{\boldsymbol{\nu},p}$ coincides with $\zeta^{\boldsymbol{\nu},p}$ on these actions, we directly have that

$$\zeta_n^{\boldsymbol{\nu},p}(x,a)=\xi_n^{\boldsymbol{\nu},p}(x,a)\geq\frac{\varepsilon_n^p(x)}{2A-1+\sqrt{A-1}}\,.$$

For a^*, we proceed via induction, tracing the transformation performed by Ξ_p. Firstly, note that at any step n and state x,

$$\sum_{a\neq a^*}\zeta_n^{\boldsymbol{\nu},p}(x,a)=\sum_{a\neq a^*}\xi_n^{\boldsymbol{\nu},p}(x,a)=\frac{\varepsilon_n^p(x)\left(A-1+1/2\sum_{a\neq a^*}\boldsymbol{\nu}_n(x,a)\right)}{A-1/2+1/2\sqrt{A-1}}$$
$$\leq\frac{\varepsilon_n^p(x)(A-1+1/2\sqrt{A-1})}{A-1/2+1/2\sqrt{A-1}}, \tag{5.6}$$

where the last step uses the Cauchy-Schwarz inequality:

$$\sum_{a\neq a^*}\boldsymbol{\nu}_n(x,a)\leq\sqrt{A-1}\sqrt{\sum_{a\neq a^*}\boldsymbol{\nu}_n^2(x,a)}\leq\sqrt{A-1}\|\boldsymbol{\nu}\|_2\leq\sqrt{A-1}.$$

Since $p_1(x)=\varepsilon_1^p(x)$, we have via the definition of Ξ_p and (5.6) that

$$\zeta_1^{\boldsymbol{\nu},p}(x,a^*)=p_1(x)-\sum_{a\neq a^*}\zeta_n^{\boldsymbol{\nu},p}(x,a)\geq\varepsilon_1^p(x)-\frac{\varepsilon_1^p(x)(A-1+1/2\sqrt{A-1})}{A-1/2+1/2\sqrt{A-1}}=\frac{\varepsilon_1^p(x)}{2A-1+\sqrt{A-1}}.$$

While at $n\geq 2$, assuming $\zeta_{n-1}^{\boldsymbol{\nu},p}(x,a^*)\geq\frac{\varepsilon_{n-1}^p(x)}{2A-1+\sqrt{A-1}}$ at every state x, using once again the definition of Ξ_p and (5.6) yields that

$$\begin{aligned}\zeta_n^{\boldsymbol{\nu},p}(x,a^*)&=\sum_{x',a'}\zeta_{n-1}^{\boldsymbol{\nu},p}(x',a')p_n(x|x',a')-\sum_{a\neq a^*}\zeta_n^{\boldsymbol{\nu},p}(x,a)\\&\geq\sum_{x',a'}\frac{\varepsilon_{n-1}^p(x')}{2A-1+\sqrt{A-1}}p_n(x|x',a')-\frac{\varepsilon_n^p(x)(A-1+1/2\sqrt{A-1})}{A-1/2+1/2\sqrt{A-1}}\\&=\varepsilon_n^p(x)-\frac{\varepsilon_n^p(x)(A-1+1/2\sqrt{A-1})}{A-1/2+1/2\sqrt{A-1}}=\frac{\varepsilon_n^p(x)}{2A-1+\sqrt{A-1}},\end{aligned}$$

where the second equality holds by the definition of $\varepsilon_n^p(x)$ in (5.4). We have proved that $\xi^{\boldsymbol{\nu},p}\in(\mathcal{M}^p)^-$ for any $\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}$, implying the following lemma. Let E^p be the $NS(A-1)\times NS(A-1)$ diagonal matrix where $E^p(n,x,a,n,x,a)=\varepsilon_n^p(x)$ for all n,x, and $a\neq a^*$, and note that this matrix is non-singular by the positivity of $\varepsilon_n^p(x)$.

Lemma 5.4. *For an MDP with dynamics p and uniformly reachable states, it holds that*

$$\frac{E^p\mathbb{1}_{NS(A-1)}}{A-1/2+1/2\sqrt{A-1}}+\frac{E^p\mathbb{B}^{NS(A-1)}}{2A-1+\sqrt{A-1}}\subset(\mathcal{M}^p)^-.$$

Essentially, we constructed a set of policies, given by $\{\pi\in\Pi\colon\mu^{\pi,p}=\zeta^{\boldsymbol{\nu},p}$ for some $\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}\}$, for any member of which, the probability at step n of being in some state x

and executing some given action a is lower bounded by $\varepsilon_n^p(x)/(2A-1+\sqrt{A-1})$. The set of occupancy measures induced by these policies corresponds to the $NS(A-1)$-dimensional ellipsoid described above.[4] Observing the value of the loss function at a mix (convex combination) of any policy and a randomly sampled policy from this set allows one to construct an estimator for the gradient of the function at the occupancy measure of the former, as will be shown in subsequent sections.

We close this preliminary section by proving the following auxiliary lemma.

Lemma 5.5. *Let $F\colon \mathbb{R}^{NSA} \to \mathbb{R}$ be a differentiable function, and define $\mathbb{F}\colon \mathbb{R}^{NS(A-1)} \to \mathbb{R}$ such that $\mathbb{F}(\xi) \coloneqq F(\Xi_p(\xi))$. With $\mu^1, \mu^2, \mu^3 \in \mathcal{M}^p$, it holds that*

$$\langle \nabla F(\mu^1), \mu^2 - \mu^3 \rangle = \left\langle \nabla \mathbb{F}\left(\Lambda_p^{-1}(\mu^1)\right), \Lambda_p^{-1}(\mu^2) - \Lambda_p^{-1}(\mu^3) \right\rangle.$$

Proof. Recall that Λ_p is the restriction of Ξ_p to $(\mathcal{M}^p)^-$. Notice first that since $\Xi_p(\xi) = B^p\xi + \boldsymbol{\beta}^p$, the definition of $\mathbb{F}$ implies that $\nabla\mathbb{F}(\xi) = \left(B^p\right)^\intercal \nabla F(\Xi_p(\xi))$, implying in turn that $\nabla\mathbb{F}\left(\Lambda_p^{-1}(\mu^1)\right) = \left(B^p\right)^\intercal \nabla F(\mu^1)$. Next, note that $\mu^2 - \mu^3 = B^p\left(\Lambda_p^{-1}(\mu^2) - \Lambda_p^{-1}(\mu^3)\right)$. Using these two facts, the lemma follows immediately:

$$\begin{aligned}
\langle \nabla F(\mu^1), \mu^2 - \mu^3 \rangle &= \left\langle \nabla F(\mu^1), B^p\left(\Lambda_p^{-1}(\mu^2) - \Lambda_p^{-1}(\mu^3)\right) \right\rangle \\
&= \left\langle \left(B^p\right)^\intercal \nabla F(\mu^1), \Lambda_p^{-1}(\mu^2) - \Lambda_p^{-1}(\mu^3) \right\rangle \\
&= \left\langle \nabla \mathbb{F}\left(\Lambda_p^{-1}(\mu^1)\right), \Lambda_p^{-1}(\mu^2) - \Lambda_p^{-1}(\mu^3) \right\rangle.
\end{aligned}$$

□

5.3 Entropic Regularization and Ellipsoidal Gradient Estimation

Returning to our problem setting, we outline in this section a first solution strategy based on the gradient estimation technique of Flaxman et al. (2005). To minimize clutter, we will sometimes drop the dependence on p (the MDP's dynamics) from the notation introduced in the previous section when clear. We start by describing the approach adopted by Moreno et al. (2024a) in the easier setting where the learner observes $\nabla F^t(\mu^\pi)$ upon choosing policy π at round t.

[4] We use the term "ellipsoid" to refer to both the surface and the solid.

For some distribution q over a finite set $\mathcal{S}$, we define $\phi(q) \coloneqq \sum_{x\in\mathcal{S}} q(x)\log q(x)$ as the negative entropy of q. Let $\dot{A} \in \{A, A-1\}$; for any $\xi \in \mathbb{R}^{NS\dot{A}}$, we define the norm

$$\|\xi\|_{\infty,1} \coloneqq \max_{n\in[N]} \sum_{x,a\in[S]\times[\dot{A}]} |\xi_n(x,a)|\,,$$

whose dual norm is given by $\|\xi\|_{1,\infty} \coloneqq \sum_{n\in[N]} \max_{x,a\in[S]\times[\dot{A}]} |\xi_n(x,a)|$. For $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$, define

$$\psi(\mu) \coloneqq \sum_{n=1}^{N} \left(\phi(\mu_n) - \phi(\rho_n^{\mu})\right),$$

which is the sum of the negative conditional entropy of the state-action distributions $(\mu_n)_{n\in[N]}$ (see, e.g., Neu et al., 2017, Section 3). For $\mu, \mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ such that $\mu_n(x,a) = 0$ whenever $\mu'_n(x,a) = 0$, this function induces the Bregman divergence

$$\Gamma(\mu,\mu') \coloneqq \sum_{n=1}^{N} \left(D(\mu_n,\mu'_n) - D(\rho_n^{\mu},\rho_n^{\mu'})\right)$$

as states Proposition 4.3 of Moreno et al. (2024a), where $D(\cdot,\cdot)$ is the KL-divergence. Moreover, the same proposition states that ψ is 1-strongly convex w.r.t. the norm $\|\cdot\|_{\infty,1}$ when restricted to the domain $\mathcal{M}^p$.

Moreno et al. (2024a) use an online mirror descent (OMD) update to generate their predictions at each round, taking $\mathcal{M}^p$ as their decision space and ψ as the regularizer. In particular, for some learning rate $\eta > 0$, they set

$$\mu^{t+1} \leftarrow \operatorname*{arg\,min}_{\mu\in\mathcal{M}^p} \eta\left\langle \nabla F^t(\mu^t), \mu\right\rangle + \Gamma(\mu,\mu^t)\,.$$

They show that these iterates, specifically, their corresponding policies,[5] can be expressed in closed form via a simple recursion:

Lemma 5.6 (Moreno et al., 2024a, Theorem 4.1)**.** *Let $\eta > 0$, $\boldsymbol{z} \in \mathbb{R}^{N,S,A}$, $\pi' \in \Pi$, and*

$$\pi^* \coloneqq \operatorname*{arg\,min}_{\pi\in\Pi} \eta\left\langle \boldsymbol{z}, \mu^{\pi}\right\rangle + \Gamma(\mu^{\pi},\mu^{\pi'})\,.$$

Then,

$$\pi_n^*(a|x) = \frac{\pi'_n(a|x)\exp\left(-\eta\widetilde{Q}_n(x,a)\right)}{\sum_{a'}\pi'_n(a'|x)\exp\left(-\eta\widetilde{Q}_n(x,a')\right)},$$

[5] Recall that $\mathcal{M}^p = \{\mu^{\pi,p} : \pi \in \Pi\}$.

where $\widetilde{Q} \in \mathbb{R}^{NSA}$ is an auxiliary vector (a regularized action-value function) defined recursively as follows: firstly, we have that $\widetilde{Q}_N(x,a) := -\boldsymbol{z}_N(x,a)$; then, for $1 \le n \le N-1$,

$$\widetilde{Q}_n(x,a) := -\boldsymbol{z}_n(x,a) + \sum_{x'} p_{n+1}(x'|x,a)\left(-\frac{D\big(\pi^*_{n+1}(\cdot|x'),\pi'_{n+1}(\cdot|x')\big)}{\eta} + \sum_{a'}\pi^*_{n+1}(a'|x')\widetilde{Q}_{n+1}(x',a')\right).$$

In our case, we do not have access to $\nabla F^t(\mu^t)$, so we resort to constructing a one-point estimate to use as a surrogate. Let $k\colon \mathcal{S} \to \mathbb{R}$ be a convex function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set satisfying $\mathbb{B}^d \subseteq \mathcal{S}$. Fix some $\delta \in (0,1)$ and define $\widehat{k}\colon \mathcal{S} \to \mathbb{R}$ as a smoothed approximation of k: $\widehat{k}(x) := \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^d}[k((1-\delta)x+\delta\boldsymbol{v})]$ for $x \in \mathcal{S}$. Then, Lemma 1 in Flaxman et al. (2005) implies that

$$\nabla\widehat{k}(x) = \frac{(1-\delta)d}{\delta}\mathbb{E}_{\boldsymbol{u}\in\mathbb{S}^d}[k((1-\delta)x+\delta\boldsymbol{u})\boldsymbol{u}],$$

recalling that $\mathbb{S}^d$ denotes the unit Euclidean sphere in R^d. Hence, for some $\boldsymbol{u}$ uniformly sampled from $\mathbb{S}^d$, $\frac{(1-\delta)d}{\delta}k((1-\delta)x+\delta\boldsymbol{u})\boldsymbol{u}$ can be used as a one-point stochastic surrogate for the gradient. Here, δ is used to control the trade-off between the magnitude (hence, the variance) of the gradient estimate and the bias incurred by relying on the smoothed approximation $\widehat{k}$.

Applying this idea to our problem is not immediate since $\mathbb{B}^d \not\subseteq \mathcal{M}^p$ as discussed before. Hence, we will rely on the lower dimensional representation of the set of occupancy measures introduced in the previous section. Firstly, at every round t, we define $\mathbb{F}^t$ as the mapping $\xi \mapsto F^t(\Lambda_p(\xi))$ for $\xi \in (\mathcal{M}^p)^-$. We will focus instead on estimating the gradient of $\mathbb{F}^t$ at $\Lambda_p^{-1}(\mu^t)$. Via Lemma 5.5, this will suffice for our purposes even if it does not allow recovering $\nabla F^t(\mu^t)$; as will be shown in the analysis, we only need information about $\nabla F^t(\mu^t)$ in the direction of $\mu^t - \mu$ for any comparator $\mu \in \mathcal{M}^p$.

Recalling the definition of $\varepsilon_n^p(x)$ in (5.4) and that of the diagonal matrix E^p defined before Lemma 5.4, let $\kappa := 1/(A - 1/2 + 1/2\sqrt{A-1})$ and define the smoothed function $\widehat{\mathbb{F}}^t\colon (\mathcal{M}^p)^- \to \mathbb{R}$ such that

$$\widehat{\mathbb{F}}^t(\xi) := \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)\xi+\delta\xi^{\boldsymbol{v},p}\right)\right] = \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)\xi+\delta\kappa E^p(\mathbb{1}_{NS(A-1)}+\boldsymbol{v}/2)\right)\right]$$

for $\xi \in (\mathcal{M}^p)^-$, where the equality follows from the definition of $\xi^{\boldsymbol{v},p}$ in (5.5). This is a well-defined function as $\mathbb{F}^t$ is evaluated on a convex combination of ξ and a point in the ellipsoid asserted by Lemma 5.4 to belong to $(\mathcal{M}^p)^-$. Similarly to before, we have via the

chain rule and Corollary 6.8 in (Hazan, 2021) that

$$\nabla\widehat{\mathbb{F}}^t(\xi) = \frac{2(1-\delta)}{\delta\kappa} NS(A-1)\,\mathbb{E}_{\boldsymbol{u}\in\mathbb{S}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)\xi+\delta\xi^{\boldsymbol{u},p}\right)\left(E^p\right)^{-1}\boldsymbol{u}\right],$$

assuming E^p is invertible, which holds under uniform reachability. Then, with $\boldsymbol{u}^t$ sampled uniformly from $\mathbb{S}^{NS(A-1)}$, we define

$$\begin{aligned} g^t &:= \frac{2(1-\delta)}{\delta\kappa} NS(A-1)\mathbb{F}^t\left((1-\delta)\Lambda_p^{-1}(\mu^t)+\delta\xi^{\boldsymbol{u}^t,p}\right)\left(E^p\right)^{-1}\boldsymbol{u}^t \\ &= \frac{2(1-\delta)}{\delta\kappa} NS(A-1)F^t\left((1-\delta)\mu^t+\delta\zeta^t\right)\left(E^p\right)^{-1}\boldsymbol{u}^t, \end{aligned} \tag{5.7}$$

where $\zeta^t := \zeta^{\boldsymbol{u}^t,p} = \Lambda_p(\xi^{\boldsymbol{u}^t,p})$, and the equality holds since

$$\mathbb{F}^t\left((1-\delta)\Lambda_p^{-1}(\mu^t)+\delta\xi^{\boldsymbol{u}^t,p}\right) = F^t\left(\Lambda_p\left((1-\delta)\Lambda_p^{-1}(\mu^t)+\delta\xi^{\boldsymbol{u}^t,p}\right)\right) = F^t\left((1-\delta)\mu^t+\delta\zeta^t\right)$$

using the definition of $\mathbb{F}^t$ and the fact that Λ_p is an affine function. Via the arguments laid before, we have that

$$\mathbb{E}_{\boldsymbol{u}^t}\left[g^t\right] = \nabla\widehat{\mathbb{F}}^t\left(\Lambda_p^{-1}(\mu^t)\right). \tag{5.8}$$

In Algorithm 5.1, we use the OMD update of Moreno et al. (2024a) except that we employ g^t in place of the gradient $\nabla F^t(\mu^t)$. To be precise, g^t is firstly expanded into an NSA-dimensional vector simply by placing 0 in the coordinates corresponding to the action a^*. Note that constructing g^t only requires querying F^t once at the mixture $(1-\delta)\mu^t+\delta\zeta^t$. This means, however, that our choice of policy at round t corresponds to this mixture occupancy measure and not to μ^t; this discrepancy is accounted for in the analysis. We will rely on the following assumption, also present in (Moreno et al., 2024a), when analyzing the regret of Algorithm 5.1.

Assumption 5.7. *At every round $t\in[T]$, the loss function F^t is L-Lipschitz in the $\|\cdot\|_1$ norm for some common $L>0$; that is, for any $\mu,\mu'\in(\Delta_{\mathcal{X}\times\mathcal{A}})^N$, $|F^t(\mu)-F^t(\mu')|\le L\|\mu-\mu'\|_1$.*

Recalling the definition of $\varepsilon_n^p(x)$ in (5.4), define $\varepsilon := \min_{n,x}\varepsilon_n^p(x)$, which is strictly positive if the states of the MDP are uniformly reachable (see Definition 5.1). The following theorem provides a bound on the regret of Algorithm 5.1.

Algorithm 5.1 Bandit OMD-CURL with entropic regularization

input: MDP with transition dynamics p such that $\min_{n,x} \varepsilon_n^p(x) > 0$, learning rate $\eta > 0$, exploration parameter $\delta \in (0,1]$

initialization: $\mu^1 \leftarrow \arg\min_{\mu \in \mathcal{M}} \psi(\mu)$

for $t = 1, \ldots, T$ **do**

draw $\boldsymbol{u}^t \in \mathbb{S}^{NS(A-1)}$ uniformly at random

$\zeta^t \leftarrow \zeta^{\boldsymbol{u}^t, p}$

$\widehat{\mu}^t \leftarrow (1-\delta)\mu^t + \delta\zeta^t$

$\pi_n^t(a|x) \leftarrow \widehat{\mu}^t(x,a) / \sum_{a \in \mathcal{A}} \widehat{\mu}^t(x,a)$

output π^t and observe $F^t(\widehat{\mu}^t)$

$g^t \leftarrow \frac{2(1-\delta)}{\delta\kappa} NS(A-1) F^t(\widehat{\mu}^t) \big(E^p\big)^{-1} \boldsymbol{u}^t$

construct $\mathring{g}^t \in \mathbb{R}^{NSA}$ as $\mathring{g}_n^t(x,a) \leftarrow g_n^t(x,a)$ for $a \neq a^*$ and $\mathring{g}_n^t(x,a^*) \leftarrow 0$

$\mu^{t+1} \leftarrow \arg\min_{\mu \in \mathcal{M}^p} \eta \langle \mathring{g}^t, \mu \rangle + \Gamma(\mu, \mu^t)$ ▷ see Lemma 5.6 for a closed-from solution

end for

Theorem 5.8. *Under Assumption 5.7 and uniform reachability ($\varepsilon > 0$), Algorithm 5.1 with* $\eta = \sqrt{\frac{2\varepsilon^2\delta^2 N \log(A)}{9N^5 S^2 A^4 T}}$ *and* $\delta = \min\left\{1, \sqrt{\frac{3N^3SA^2\sqrt{2\log(A)T}}{4\varepsilon LNT}}\right\}$ *satisfies*

$$R_T \leq \max\left\{4\big(2\log A\big)^{1/4}\sqrt{\frac{3LS}{\varepsilon}}AN^2T^{3/4}, \frac{6}{\varepsilon}N^3SA^2\sqrt{2\log(A)T}\right\}.$$

This bound features a $T^{3/4}$ term, as typical for regret guarantees proved for OMD with the one-point gradient estimator we relied on. Its bigger drawback is the dependence on $1/\varepsilon$, a quantity that depends on the transition dynamics of the MDP. In particular, ε lower bounds $\varepsilon_n^p(x)$ for any n and x, which as we recall, is a uniform lower bound on the probability of visiting state x at step n for any policy within the set $\{\pi \in \Pi\colon \mu^\pi = \zeta^{\boldsymbol{v},p}$ for some $\boldsymbol{v} \in \mathbb{B}^{NS(A-1)}\}$. These quantities can be arbitrarily small, especially at later steps in the episode. We expand more on this issue after the proof.

Proof. Let $\mu^* \in \arg\min_{\mu \in \mathcal{M}^p} \sum_{t=1}^T F^t(\mu)$. Since $\widehat{\mu}^t = \mu^{\pi^t}$, it holds that

$$R_T = \mathbb{E}\sum_{t=1}^T \big(F^t(\mu^{\pi^t}) - F^t(\mu^*)\big) = \mathbb{E}\sum_{t=1}^T \big(F^t(\widehat{\mu}^t) - F^t(\mu^*)\big).$$

Define $\widehat{F}^t\colon (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ such that

$$\widehat{F}^t(\mu) \coloneqq \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[F^t((1-\delta)\mu + \delta\zeta^{\boldsymbol{\nu},p})\right],$$

where, recalling from the previous section, $\zeta^{\boldsymbol{\nu},p} = \Lambda_p(\xi^{\boldsymbol{\nu},p})$ with $\xi^{\boldsymbol{\nu},p}$ as defined in (5.5). Next, using the Lipschitz smoothness of F^t and the fact that $\|\mu\|_1 \le N$ for any $\mu \in \mathcal{M}^p$, we have that

$$\begin{aligned} F^t(\widehat{\mu}^t) - \widehat{F}^t(\mu^t) &= F^t((1-\delta)\mu^t + \delta\zeta^t) - \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[F^t((1-\delta)\mu^t + \delta\zeta^{\boldsymbol{\nu},p})\right] \\ &\le \delta L \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\|\zeta^t - \zeta^{\boldsymbol{\nu},p}\|_1 \le 2\delta LN \end{aligned}$$

and that

$$\begin{aligned} \widehat{F}^t(\mu^*) - F^t(\mu^*) = \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[F^t((1-\delta)\mu^* + \delta\zeta^{\boldsymbol{\nu},p})\right] - F^t(\mu^*) &\le \delta L \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\|\zeta^{\boldsymbol{\nu},p} - \mu^*\|_1 \\ &\le 2\delta LN. \end{aligned}$$

Hence, we obtain that

$$R_T \le \mathbb{E}\sum_{t=1}^{T}\left(\widehat{F}^t(\mu^t) - \widehat{F}^t(\mu^*)\right) + 4\delta LNT.$$

Note that $\widehat{\mathbb{F}}^t = \widehat{F}^t \circ \Lambda_p$, which holds since Λ_p is an affine function and $\mathbb{F}^t = F^t \circ \Lambda_p$:

$$\begin{aligned} \widehat{\mathbb{F}}^t(\xi) &= \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)\xi + \delta\xi^{\boldsymbol{\nu},p}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[F^t\left(\Lambda_p\left((1-\delta)\xi + \delta\xi^{\boldsymbol{\nu},p}\right)\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\nu}\in\mathbb{B}^{NS(A-1)}}\left[F^t\left((1-\delta)\Lambda_p(\xi) + \delta\zeta^{\boldsymbol{\nu},p}\right)\right] = \widehat{F}^t\left(\Lambda_p(\xi)\right). \end{aligned}$$

Hence, the convexity of F^t and Lemma 5.5 imply that

$$\begin{aligned} R_T &\le \mathbb{E}\sum_{t=1}^{T}\left\langle \nabla\widehat{F}^t(\mu^t), \mu^t - \mu^*\right\rangle + 4\delta LNT \\ &= \mathbb{E}\sum_{t=1}^{T}\left\langle \nabla\widehat{\mathbb{F}}^t\left(\Lambda_p^{-1}(\mu^t)\right), \Lambda_p^{-1}(\mu^t) - \Lambda_p^{-1}(\mu^*)\right\rangle + 4\delta LNT. \end{aligned} \tag{5.9}$$

For round $t \in [T]$, let $\mathcal{F}_t \coloneqq \sigma(u^1,\dots,u^t)$ denote the σ-algebra generated by $u^1,\dots,u^t$; and let $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_0$ being the trivial σ-algebra. Building upon the last

inequality, we have that

$$\begin{aligned}
R_T &\leq \mathbb{E}\sum_{t=1}^{T}\left\langle \nabla\widehat{\mathbb{F}}^t\left(\Lambda_p^{-1}(\mu^t)\right), \Lambda_p^{-1}(\mu^t) - \Lambda_p^{-1}(\mu^*)\right\rangle + 4\delta LNT \\
&= \mathbb{E}\sum_{t=1}^{T}\left\langle \mathbb{E}_{\boldsymbol{u}^t}\left[g^t\right], \Lambda_p^{-1}(\mu^t) - \Lambda_p^{-1}(\mu^*)\right\rangle + 4\delta LNT \\
&= \mathbb{E}\sum_{t=1}^{T}\left\langle \mathbb{E}_t\left[g^t\right], \Lambda_p^{-1}(\mu^t) - \Lambda_p^{-1}(\mu^*)\right\rangle + 4\delta LNT \\
&= \mathbb{E}\sum_{t=1}^{T}\left\langle g^t, \Lambda_p^{-1}(\mu^t) - \Lambda_p^{-1}(\mu^*)\right\rangle + 4\delta LNT \\
&= \mathbb{E}\sum_{t=1}^{T}\left\langle \mathring{g}^t, \mu^t - \mu^*\right\rangle + 4\delta LNT,
\end{aligned}$$

where the first equality follows from (5.8); the second equality holds since conditioned on $\mathcal{F}_{t-1}$, $\boldsymbol{u}^t$ is the only source of randomness in g^t and is sampled identically and independently in every round; the third equality uses the tower rule, linearity of expectation, and the fact that μ^t is measurable with respect to $\mathcal{F}_{t-1}$; and the fourth equality uses that Λ_p only expands its input by recovering the coordinates involving action a^*, and that the corresponding coordinates in $\mathring{g}^t$ are 0.

Thanks to the definition of μ^t, Theorem 6.10 in (Orabona, 2023) implies that

$$\sum_{t=1}^{T}\left\langle \mathring{g}^t, \mu^t - \mu^*\right\rangle \leq \frac{\Gamma(\mu^*, \mu^1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathring{g}^t\|_{1,\infty}^2.$$

For the first term, we have that

$$\Gamma(\mu^*, \mu^1) = \psi(\mu^*) - \psi(\mu^1) - \left\langle \nabla\psi(\mu^1), \mu^* - \mu^1\right\rangle \leq \psi(\mu^*) - \psi(\mu^1) \leq -\psi(\mu^1) \leq N\log(A),$$

where the first equality holds since Γ is the Bregman divergence induced by ψ, the first inequality follows from the convexity of ψ and the fact that $\mu^1 = \arg\min_{\mu\in\mathcal{M}}\psi(\mu)$, the second inequality uses the negativity of ψ, and the last inequality holds via Lemma D.2 in

(Moreno et al., 2024a). While for the second term, it holds for every round t that

$$\begin{aligned}
\|\mathring{g}^t\|_{1,\infty}^2 &= \Big(\frac{2(1-\delta)}{\delta\kappa} NS(A-1)F^t(\widehat{\mu}^t)\Big)^2 \Big\| \big(E^p\big)^{-1} \boldsymbol{u^t} \Big\|_{1,\infty}^2 \\
&\leq \Big(\frac{3}{\delta} N^2 SA^2\Big)^2 \Big\| \big(E^p\big)^{-1} \boldsymbol{u^t} \Big\|_{1,\infty}^2 \\
&= \frac{9}{\delta^2} N^4 S^2 A^4 \Big(\sum_{n=1}^{N} \max_{x,a} \big| \boldsymbol{u}_n^t(x,a) \big| / \varepsilon_n^p(x) \Big)^2 \\
&\leq \frac{9}{\delta^2\varepsilon^2} N^4 S^2 A^4 \Big(\sum_{n=1}^{N} \max_{x,a} \big| \boldsymbol{u}_n^t(x,a) \big| \Big)^2 \\
&\leq \frac{9}{\delta^2\varepsilon^2} N^5 S^2 A^4 \Big(\sum_{n=1}^{N} \max_{x,a} \big| \boldsymbol{u}_n^t(x,a) \big|^2 \Big) \\
&\leq \frac{9}{\delta^2\varepsilon^2} N^5 S^2 A^4 \|\boldsymbol{u}^t\|_2^2 \leq \frac{9}{\delta^2\varepsilon^2} N^5 S^2 A^4,
\end{aligned}$$

where the first inequality uses that $\kappa = 1/(A - 1/2 + 1/2\sqrt{A-1})$, $\delta \leq 1$, and $F^t(\widehat{\mu}^t) \leq N$; the second inequality uses that $\varepsilon = \min_{n,x} \varepsilon_n^p(x)$; and the third inequality holds via the Cauchy-Schwarz inequality. Putting everything together, we have shown that

$$R_T \leq \frac{N\log(A)}{\eta} + \frac{9\eta}{2\delta^2\varepsilon^2} N^5 S^2 A^4 T + 4\delta LNT.$$

Using $\eta = \sqrt{\frac{2\varepsilon^2\delta^2 N\log(A)}{9N^5S^2A^4T}}$ and $\delta = \min\left\{1, \sqrt{\frac{3N^3SA^2\sqrt{2\log(A)T}}{4\varepsilon LNT}}\right\}$ yields that

$$\begin{aligned}
R_T &\leq \sqrt{\frac{18\log(A)}{\delta^2\varepsilon^2} N^6 S^2 A^4 T} + 4\delta LNT \\
&= \frac{3}{\delta\varepsilon} N^3 SA^2 \sqrt{2\log(A)T} + 4\delta LNT \\
&\leq \max\left\{ 4\big(2\log A\big)^{1/4} \sqrt{\frac{3LS}{\varepsilon}} AN^2 T^{3/4}, \frac{6}{\varepsilon} N^3 SA^2 \sqrt{2\log(A)T} \right\}.
\end{aligned}$$

□

At each round t, the point we query is $\widehat{\mu}^t$, whose $NS(A-1)$-dimensional representation is given by $\Lambda_p^{-1}(\widehat{\mu}^t) = (1-\delta)\Lambda_p^{-1}(\mu^t) + \delta\kappa E^p(\mathbb{1}_{NS(A-1)} + \boldsymbol{u}^t/2)$, which belongs to the surface of the (standard axis-aligned) ellipsoid centered at the point $(1-\delta)\Lambda_p^{-1}(\mu^t) + \delta\kappa E^p \mathbb{1}_{NS(A-1)}$ and whose semi-axis corresponding to some n, x, and $a \neq a^*$ is of length $\delta\kappa\varepsilon_n^p(x)/2$. From the definition of the gradient estimate g^t, the shorter theses semi-axes are, the larger the magnitude of g^t will be, which affects our regret guarantee through the term $\mathbb{E}\|g^t\|_{1,\infty}^2$

featured in the analysis of the stability (variance) term of the OMD routine we employ; in particular, the effect of the lengths of the semi-axes manifests itself through the factor

$$\frac{4}{\delta^2\kappa^2}\mathbb{E}\left\|\left(E^p\right)^{-1}\boldsymbol{u}^t\right\|_{1,\infty}^2.$$

Somewhat sloppily, we bound this quantity by $4N/(\delta^2\varepsilon^2\kappa^2)$, seeing that in any case, the undesirable dependence on the reciprocals of the $\varepsilon_n^p(x)$ terms is unavoidable at this junction. And although we control the value of δ, it must remain small enough to offset the effect of the bias introduced by our reliance on the smoothed functions $\widehat{F}^t$ in the algorithm design and the analysis.

One tool we did not use is the local norm analysis of OMD, see, for example, Lemma 6.16 in (Orabona, 2023). This tool, often used when analyzing bandit algorithms based on online convex optimization methods, would have allowed us to replace the norm $\|\cdot\|_{1,\infty}$ in the aforementioned problematic factor with the local norm $\|\cdot\|_{(\nabla^2\psi(\mu^t))^{-1}}$, where $\|x\|_{(\nabla^2\psi(\mu^t))^{-1}} \coloneqq \sqrt{x^\intercal(\nabla^2\psi(\mu^t))^{-1}x}$. In our case, however, there is no immediate property of the gradient estimate that can be exploited by using the local norm; neither the orientation of the axes of our ellipsoid nor their length are related to $(\nabla^2\psi(\mu^t))^{-1}$. This then inspires an alternative approach: instead of sampling the perturbation at every round from a shifted version of a common ellipsoid as we have hitherto done, we can attempt instead to sample at round t from an ellipsoid defined using the local norm at μ^t. This is better facilitated via the more advanced machinery of self-concordant regularization, which is a standard technique in bandit convex (and linear) optimization (see, e.g., Abernethy et al., 2008; Hazan and Levy, 2014; Saha and Tewari, 2011), and is the subject of the next section.

5.4 Self-Concordant Regularization and (Dikin) Ellipsoidal Gradient Estimation

We have used the set $(\mathcal{M}^p)^-$, the preimage of $\mathcal{M}^p$ under the map Ξ_p (or Λ_p), to represent in $\mathbb{R}^{NS(A-1)}$ the set of valid occupancy measures. A more concise characterization, given by Lemma 5.2, is that

$$(\mathcal{M}^p)^- = \left\{\xi \in \mathbb{R}^{NS(A-1)} \colon B^p\xi \geq -\boldsymbol{\beta}^p\right\};$$

in other words, $(\mathcal{M}^p)^-$ is a polyhedral set formed by the constraints $B^p(n,x,a,\cdot,\cdot,\cdot)^\intercal\xi + \boldsymbol{\beta}_n^p(x,a) \geq 0$ for $n,x,a \in [N]\times\mathcal{X}\times\mathcal{A}$. Moreover, Lemma 5.3 asserts that $\operatorname{int}(\mathcal{M}^p)^-$, the interior of $(\mathcal{M}^p)^-$, is not empty assuming all states are uniformly reachable (Definition 5.1).

Consider then the function $\psi_{\text{lb}}\colon \operatorname{int}(\mathcal{M}^p)^- \to \mathbb{R}$ defined as

$$\psi_{\text{lb}}(\xi) := -\sum_{n,x,a} \log\big(B(n,x,a,\cdot,\cdot,\cdot)^\intercal \xi + \boldsymbol{\beta}_n(x,a)\big).$$

Via Corollary 3.1.1 in (Nemirovski, 2004), ψ_{lb} is a ϑ-self-concordant barrier (see Definition 3.1.1 in Nemirovski, 2004) for $(\mathcal{M}^p)^-$ with $\vartheta = NSA$. Our approach in this section is to perform OMD directly on the set $(\mathcal{M}^p)^-$ with ψ_{lb} as the regularizer. Self-concordant regularization (and logarithmic barrier regularization in particular) has been used for online learning in MDPs in different settings (Lee et al., 2020; Cohen et al., 2021; Van der Hoeven et al., 2023). The closest work is that of Cohen et al. (2021), who study reinforcement learning with aggregate (trajectory-wise) bandit feedback. Our next algorithm is essentially an analogue of their second algorithm (if the MDP is known), but deals with general convex loss functions (as opposed to just linear functions), and with minor differences in the representation of the set of occupancy measures and the formulation of the regularizer.

For $\xi \in \operatorname{int}(\mathcal{M}^p)^-$ and $y \in \mathbb{R}^{NS(A-1)}$, define the local norm $\|y\|_\xi := \sqrt{y^\intercal \nabla^2 \psi_{\text{lb}}(\xi) y}$. This is indeed a norm since the fact that $(\mathcal{M}^p)^-$ is bounded implies via Property II in (Nemirovski, 2004, Section 2.2) that the Hessian of ψ_{lb} is non-singular everywhere. Its dual norm is denoted as $\|y\|_{\xi,*} := \sqrt{y^\intercal (\nabla^2 \psi_{\text{lb}}(\xi))^{-1} y}$. The Dikin ellipsoid of radius r at $\xi \in \operatorname{int}(\mathcal{M}^p)^-$ is given by

$$\mathcal{E}_r(\xi) := \{y \in \mathbb{R}^{NS(A-1)} \colon \|y-\xi\|_\xi \le r\} = \xi + r(\nabla^2 \psi_{\text{lb}}(\xi))^{-1/2} \mathbb{B}^{NS(A-1)}.$$

Via Property I in (Nemirovski, 2004, Section 2.2), $\mathcal{E}_1(\xi) \subseteq (\mathcal{M}^p)^-$ for any $\xi \in \operatorname{int}(\mathcal{M}^p)^-$. As opposed to the somewhat laborious construction undertaken in the previous section, this result immediately provides us with an ellipsoid entirely contained inside $(\mathcal{M}^p)^-$ for every $\xi \in \operatorname{int}(\mathcal{M}^p)^-$.

For $\xi, y \in \operatorname{int}(\mathcal{M}^p)^-$, we denote by $D_{\psi_{\text{lb}}}(y,\xi) := \psi_{\text{lb}}(y) - \psi_{\text{lb}}(\xi) - \langle y-\xi, \nabla\psi_{\text{lb}}(\xi)\rangle$ the Bregman divergence between y and ξ with respect to ψ_{lb}. From the previous section, we recall the definition $\mathbb{F}^t := F^t \circ \Lambda_p$. As alluded to above, our OMD updates will take the form

$$\xi^{t+1} \leftarrow \operatorname*{arg\,min}_{\xi \in (\mathcal{M}^p)^-} \eta \langle g^t, \xi\rangle + D_{\psi_{\text{lb}}}(\xi, \xi^t),$$

where g^t will be chosen as a surrogate for $\nabla \mathbb{F}^t(\xi^t)$. Differently from the previous section, we redefine the smoothed approximation $\widehat{\mathbb{F}}^t\colon (\mathcal{M}^p)^- \to \mathbb{R}$ such that

$$\widehat{\mathbb{F}}^t(\xi) := \mathbb{E}_{\boldsymbol{v} \in \mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\big((1-\delta)\xi + \delta\big(\xi^t + (\nabla^2 \psi_{\text{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\big)\big)\right].$$

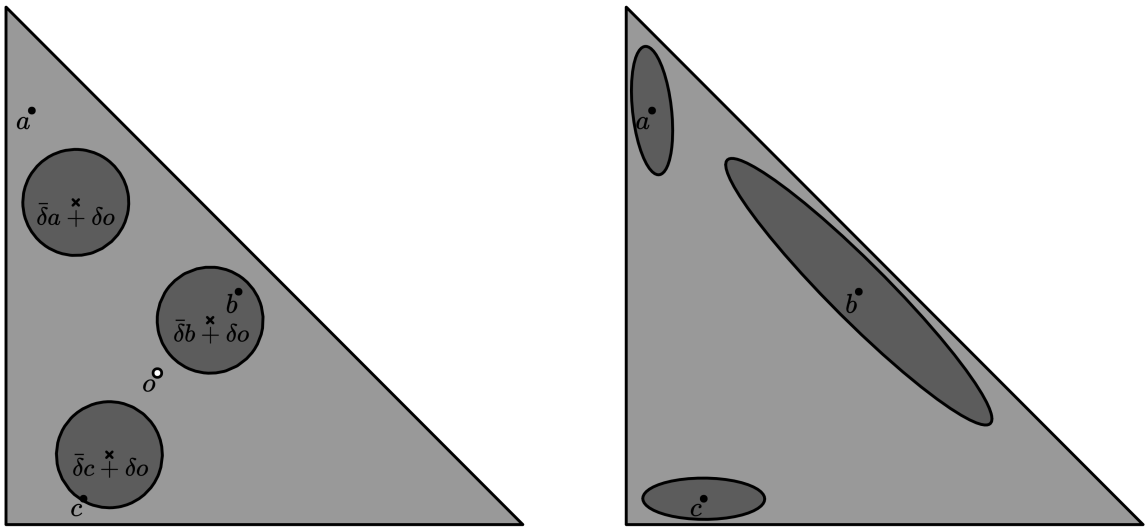

Fig. 5.1 This figure provides a comparison between the sampling approach used in (5.7), represented on the left, and that used in (5.11), represented on the right. The simplified domain here is $\{x \in [0,1]^2 : \|x\|_1 \leq 1\}$. Both approaches are illustrated at three points: a, b, and c. In the first approach, with some $\delta \in (0,1)$ and $\bar{\delta} := 1-\delta$, we sample from a circle of radius $\delta/(2+\sqrt{2})$ centered at a convex combination between the point of interest and $o := \big(1/(2+\sqrt{2}), 1/(2+\sqrt{2})\big)$. In the second approach, we consider the barrier $-\log(1-x_1-x_2) - \sum_{i=1,2}\log(x_i)$ and sample from the Dikin ellipsoid (of a certain common radius) induced by this function at each point.

This is well-defined since we are evaluating $\mathbb{F}^t$ on a convex combination of the argument ξ and a point inside the ellipsoid $\mathcal{E}_1(\xi^t)$, which is a subset of $(\mathcal{M}^p)^-$ as cited before. Invoking as before the chain rule and Corollary 6.8 in (Hazan, 2021) yields that

$$\nabla\widehat{\mathbb{F}}^t(\xi) = \frac{(1-\delta)}{\delta} NS(A-1)\,\mathbb{E}_{\boldsymbol{u}\in\mathbb{S}^{NS(A-1)}}\Big[\mathbb{F}^t\Big((1-\delta)\xi+\delta\big(\xi^t+(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}\big)\Big)(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{1/2}\boldsymbol{u}\Big]. \tag{5.10}$$

Hence, with $\boldsymbol{u}^t$ sampled uniformly from $\mathbb{S}^{NS(A-1)}$, we pick

$$g^t := \frac{(1-\delta)}{\delta} NS(A-1)\mathbb{F}^t\big(\xi^t + \delta(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\big)(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{1/2}\boldsymbol{u}^t \tag{5.11}$$

such that $\mathbb{E}_{\boldsymbol{u}^t}\big[g^t\big] = \nabla\widehat{\mathbb{F}}^t(\xi^t)$. We summarize this approach in Algorithm 5.2, and provide in Figure 5.1 a graphical comparison with the approach of the previous section on a simple decision set. Using the local norm analysis of OMD, we derive a regret bound, stated in Theorem 5.11, for Algorithm 5.2 eliminating the undesirable dependence on ε featured in the bound of Theorem 5.8, even if we lose the appealing closed-form solution property enjoyed by Algorithm 5.1. Before stating the theorem we collect a few standard properties and auxiliary results concerning self-concordant barriers and their use as regularizers.

Algorithm 5.2 Bandit OMD-CURL with logarithmic barrier regularization

input: domain $(\mathcal{M}^p)^-$ with non-empty interior, learning rate $\eta > 0$, exploration parameter $\delta \in (0,1]$

initialization: $\xi^1 \leftarrow \arg\min_{\xi \in \operatorname{int}(\mathcal{M}^p)^-} \psi_{\mathrm{lb}}(\xi)$

for $t = 1, \ldots, T$ **do**

draw $\boldsymbol{u}^t \in \mathbb{S}^{NS(A-1)}$ uniformly at random

$\widehat{\xi}^t \leftarrow \xi^t + \delta(\nabla^2 \psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t$

$\widehat{\mu}^t \leftarrow \Lambda_p(\widehat{\xi}^t)$

$\pi_n^t(a|x) \leftarrow \widehat{\mu}^t(x,a) / \sum_{a \in \mathcal{A}} \widehat{\mu}^t(x,a)$

output π^t and observe $F^t(\widehat{\mu}^t)$

$g^t \leftarrow \frac{(1-\delta)}{\delta} NS(A-1) F^t(\widehat{\mu}^t)(\nabla^2 \psi_{\mathrm{lb}}(\xi^t))^{1/2}\boldsymbol{u}^t$

$\xi^{t+1} \leftarrow \arg\min_{\xi \in \operatorname{int}(\mathcal{M}^p)^-} \eta \langle g^t, \xi \rangle + D_{\psi_{\mathrm{lb}}}(\xi, \xi^t)$

end for

For $x, y \in \operatorname{int}(\mathcal{M}^p)^-$, it holds via Property I in (Nemirovski, 2004, Section 2.2) that

$$(1 - \|y - x\|_x)^2 \nabla^2 \psi_{\mathrm{lb}}(x) \preccurlyeq \nabla^2 \psi_{\mathrm{lb}}(y) \preccurlyeq \frac{1}{(1 - \|y - x\|_x)^2} \nabla^2 \psi_{\mathrm{lb}}(x) \tag{5.12}$$

whenever $\|y - x\|_x < 1$. We state the following auxiliary lemma, which will be used to assert the proximity between ξ^t and ξ^{t+1} for our algorithm. Proving this 'stability' is a crucial step in the local norm analysis.

Lemma 5.9. *Let $x \in int(\mathcal{M}^p)^-$ and $\ell \in \mathbb{R}^{NS(A-1)}$ be such that $\|\ell\|_{x,*} \leq \frac{1}{16}$, and define*

$$y := \operatorname*{arg\,min}_{\xi \in int(\mathcal{M}^p)^-} \langle \ell, \xi \rangle + D_{\psi_{lb}}(\xi, x).$$

Then, $y \in \mathcal{E}_{1/2}(x)$.

Proof. For $\xi \in \operatorname{int}(\mathcal{M}^p)^-$, let

$$g(\xi) := \langle \ell, \xi \rangle + D_{\psi_{\mathrm{lb}}}(\xi, x) = \langle \ell, \xi \rangle + \psi_{\mathrm{lb}}(\xi) - \psi_{\mathrm{lb}}(x) - \langle \xi - x, \nabla \psi_{\mathrm{lb}}(x) \rangle.$$

Note that g is a self-concordant function on $\operatorname{int}(\mathcal{M}^p)^-$ (Item (ii) in Nemirovski, 2004, Proposition 2.1.1), whose Hessian (hence, local norms and Dikin ellipsoids) coincides with that of ψ_{lb} everywhere. Moreover, g is below bounded thanks to $(\mathcal{M}^p)^-$ being a bounded set, which implies that g attains its minimum on $\operatorname{int}(\mathcal{M}^p)^-$ (Property VI in

Nemirovski, 2004, Section 2.2). This minimum is also unique via strict convexity. Hence, y is well-defined.

The rest of the proof is similar to the proof of Lemma 13 in (Wei and Luo, 2018) and Lemma 9 in (Van der Hoeven et al., 2023). Thanks to the strict convexity of g, to show that $y \in \mathcal{E}_{1/2}(x)$ it suffices to show that for any ξ on the boundary of $\mathcal{E}_{1/2}(x)$, $g(x) \le g(\xi)$; this is because $x \in \mathcal{E}_{1/2}(x)$ and $y = \arg\min_{\xi \in (\mathcal{M}^p)^-} g(\xi)$. For any such ξ on the boundary of $\mathcal{E}_{1/2}(x)$, Taylor's theorem implies that there exists some z on the line segment between x and ξ such that

$$\begin{aligned}
g(\xi) - g(x) &= \langle \xi - x, \nabla g(x) \rangle + \frac{1}{2}(\xi - x)^\intercal \nabla^2 g(z)(\xi - x) \\
&= \langle \xi - x, \ell \rangle + \frac{1}{2}(\xi - x)^\intercal \nabla^2 \psi_{\mathrm{lb}}(z)(\xi - x) \\
&\ge \langle \xi - x, \ell \rangle + \frac{1}{8}(\xi - x)^\intercal \nabla^2 \psi_{\mathrm{lb}}(x)(\xi - x) \\
&= \langle \xi - x, \ell \rangle + \frac{1}{8}\|\xi - x\|_x^2 \\
&\ge -\|\xi - x\|_x \|\ell\|_{x,*} + \frac{1}{8}\|\xi - x\|_x^2 \\
&= -\frac{1}{2}\|\ell\|_{x,*} + \frac{1}{32} \ge 0,
\end{aligned}$$

where the second equality holds since $\nabla^2 g = \nabla^2 \psi_{\mathrm{lb}}$ and $\nabla g(x) = \ell + \nabla \psi_{\mathrm{lb}}(x) - \nabla \psi_{\mathrm{lb}}(x) = \ell$, the first inequality holds via (5.12) and the fact that $z \in \mathcal{E}_{1/2}(x)$, the second inequality holds via the definition of a dual norm, the last equality holds since ξ is on the boundary of $\mathcal{E}_{1/2}(x)$, and the last inequality holds via the assumption that $\|\ell\|_{x,*} \le \frac{1}{16}$. □

For $x \in \operatorname{int}(\mathcal{M}^p)^-$, the Minkowski function of $(\mathcal{M}^p)^-$ with the pole at x is defined as

$$\pi_x(y) := \inf\{t > 0 \colon x + t^{-1}(y - x) \in (\mathcal{M}^p)^-\}$$

for $y \in (\mathcal{M}^p)^-$. The following lemma is used in the analysis to handle the bias term of the standard OMD regret guarantee, which is slightly more involved in this case considering that the comparator need not belong to the interior of $(\mathcal{M}^p)^-$, where ψ_{lb} is defined (and finite).

Lemma 5.10. *Let $x \in \operatorname{int}(\mathcal{M}^p)^-$, $y \in (\mathcal{M}^p)^-$, $\delta \in (0,1)$, and $z := (1-\delta)y + \delta x$. Then,*

$$\psi_{lb}(z) \le \psi_{lb}(x) + NSA \log \delta^{-1}.$$

Further, let $\dot{x} \coloneqq \arg\min_{x \in int(\mathcal{M}^p)^-} \psi_{lb}(x)$ *and* $\dot{z} \coloneqq (1-\delta)y + \delta\dot{x}$. *Then,*

$$D_{\psi_{lb}}(\dot{z}, \dot{x}) \leq NSA \log \delta^{-1}.$$

Proof. Since ψ_{lb} is an NSA-self-concordant barrier for $\mathcal{M}^p$, Property II in (Nemirovski, 2004, Section 3.2) implies that

$$\psi_{\text{lb}}(z) \leq \psi_{\text{lb}}(x) + NSA \log \frac{1}{1-\pi_x(z)}.$$

On the other hand,

$$x + (1-\delta)^{-1}(z-x) = x + (1-\delta)^{-1}((1-\delta)y + \delta x - x) = x + y - x = y \in (\mathcal{M}^p)^-,$$

implying that $\pi_x(z) \leq 1-\delta$. Hence, $\psi_{\text{lb}}(z) \leq \psi_{\text{lb}}(x) + NSA \log \delta^{-1}$.

Next, we note that the optimality of $\dot{x}$ implies that

$$D_{\psi_{\text{lb}}}(\dot{z}, \dot{x}) = \psi_{\text{lb}}(\dot{z}) - \psi_{\text{lb}}(\dot{x}) - \langle \dot{z} - \dot{x}, \nabla\psi_{\text{lb}}(\dot{x}) \rangle \leq \psi_{\text{lb}}(\dot{z}) - \psi_{\text{lb}}(\dot{x}),$$

which concludes the proof when combined with the first part. □

We are now ready to state and proof our regret bound for Algorithm 5.2. Besides eliminating the dependence on ε, it features a slightly better dependence on N and A compared to the bound of Theorem 5.8, but a slightly worse dependence on S. Note that we still rely on the uniform reachability assumption (see Definition 5.1) to have that the interior of $\mathcal{M}^p$ is non-empty; nevertheless, this condition can be enforced without loss of generality by modifying the adopted representation of the occupancy measures as described in Section 5.2.

Theorem 5.11. *Under Assumption 5.7 and uniform reachability, Algorithm 5.2 with* $\eta = \frac{\delta}{16}\sqrt{\frac{\log T}{N^3 SAT}}$ *and* $\delta = \min\left\{\sqrt{\frac{17}{4L}}\frac{N^{3/4}S^{3/4}A^{3/4}(\log T)^{1/4}}{T^{1/4}}, 1\right\}$ *satisfies*

$$R_T \leq \max\left\{4\sqrt{17L}N^{7/4}(SAT)^{3/4}(\log T)^{1/4}, 34\sqrt{N^5S^3A^3T\log T}\right\} + 2LN.$$

Proof. Firstly, we assert that the iterates ξ^t are well defined; similar to what was argued in the proof of Lemma 5.9, the functions $\psi_{\text{lb}}(\cdot)$ and $\eta\langle g^t, \cdot\rangle + D_{\psi_{\text{lb}}}(\cdot, \xi^t)$ are self-concordant on int $(\mathcal{M}^p)^-$ (Item (ii) in Nemirovski, 2004, Proposition 2.1.1) and bounded from below thanks to $(\mathcal{M}^p)^-$ being a bounded set, implying via Property VI in (Nemirovski, 2004, Section 2.2) that each of these functions attains its minimum on int $(\mathcal{M}^p)^-$, which is also

unique via strict convexity. Also note that indeed $\widehat{\mu}^t \in \mathcal{M}^p$ since $\widehat{\xi}^t \in (\mathcal{M}^p)^-$ as we argued before presenting the algorithm.

Let $\mu^* \in \arg\min_{\mu \in \mathcal{M}^p} \sum_{t=1}^T F^t(\mu)$, and define $\xi^* := (1-\dot{\delta})\Lambda_p^{-1}(\mu^*) + \dot{\delta}\xi^1$, where $\dot{\delta} \in (0,1)$ is a constant to be specified later. To start with, we have that

$$R_T = \mathbb{E}\sum_{t=1}^T \left(F^t(\mu^{\pi^t}) - F^t(\mu^*)\right) = \mathbb{E}\sum_{t=1}^T \left(F^t(\widehat{\mu}^t) - F^t(\mu^*)\right) = \mathbb{E}\sum_{t=1}^T \left(\mathbb{F}^t(\widehat{\xi}^t) - \mathbb{F}^t\left(\Lambda_p^{-1}(\mu^*)\right)\right).$$

Next, we derive that

$$\begin{aligned}
&\mathbb{F}^t(\widehat{\xi}^t) - \widehat{\mathbb{F}}^t(\xi^t) \\
&\quad = \mathbb{F}^t\left(\xi^t + \delta(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\right) - \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left(\xi^t + \delta(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right] \\
&\quad \le L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left\|\Lambda_p\left(\xi^t + \delta(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\right) - \Lambda_p\left(\xi^t + \delta(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1 \\
&\quad = \delta L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left\|\Lambda_p\left((\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\right) - \Lambda_p\left((\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1 \\
&\quad = \delta L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left\|\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\right) - \Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1 \\
&\quad \le \delta L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\left\|\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t\right)\right\|_1 + \left\|\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1\right] \\
&\quad \le 2\delta LN,
\end{aligned}$$

where the first inequality uses the Lipschitz smoothness of F^t and the fact that $\mathbb{F}^t = F^t \circ \Lambda_p$; the second and third equalities use the fact that Λ_p is an affine function; and the last inequality holds since both $\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{u}^t$, $\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v} \in \mathcal{E}_1(\xi^t) \subset (\mathcal{M}^p)^-$, and that for any $\xi \in (\mathcal{M}^p)^-$, $\Lambda_p(\xi) \in \mathcal{M}^p$ and therefore satisfies $\|\Lambda_p(\xi)\|_1 \le N$. We similarly derive that

$$\begin{aligned}
&\widehat{\mathbb{F}}^t(\xi^*) - \mathbb{F}^t\left(\Lambda_p^{-1}(\mu^*)\right) \\
&\quad = \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)\xi^* + \delta\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right)\right] - \mathbb{F}^t\left(\Lambda_p^{-1}(\mu^*)\right) \\
&\quad = \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[\mathbb{F}^t\left((1-\delta)(1-\dot{\delta})\Lambda_p^{-1}(\mu^*) + (1-\delta)\dot{\delta}\xi^1 + \delta\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right)\right] \\
&\qquad\qquad - \mathbb{F}^t\left(\Lambda_p^{-1}(\mu^*)\right) \\
&\quad \le L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left\|(1-\delta)(1-\dot{\delta})\mu^* + (1-\delta)\dot{\delta}\Lambda_p(\xi^1) + \delta\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right) - \mu^*\right\|_1 \\
&\quad \le L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[|\delta\dot{\delta} - \delta - \dot{\delta}|\|\mu^*\|_1 + (1-\delta)\dot{\delta}\left\|\Lambda_p(\xi^1)\right\|_1 + \delta\left\|\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1\right] \\
&\quad \le L\mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^{NS(A-1)}}\left[(\delta+\dot{\delta})\|\mu^*\|_1 + \dot{\delta}\left\|\Lambda_p(\xi^1)\right\|_1 + \delta\left\|\Lambda_p\left(\xi^t + (\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{-1/2}\boldsymbol{v}\right)\right\|_1\right] \\
&\quad \le 2\delta LN + 2\dot{\delta}LN.
\end{aligned}$$

Hence, using also the convexity of $\widehat{\mathbb{F}}^t$, we obtain that

$$\begin{aligned} R_T &\leq \mathbb{E}\sum_{t=1}^{T}\left(\widehat{\mathbb{F}}^t(\xi^t) - \widehat{\mathbb{F}}^t(\xi^*)\right) + 4\delta LNT + 2\dot{\delta}LNT \\ &\leq \mathbb{E}\sum_{t=1}^{T}\left\langle \nabla\widehat{\mathbb{F}}^t(\xi^t), \xi^t - \xi^*\right\rangle + 4\delta LNT + 2\dot{\delta}LNT\,. \end{aligned} \tag{5.13}$$

As in the proof of the previous theorem, for round $t \in [T]$, let $\mathcal{F}_t \coloneqq \sigma(u^1,\dots,u^t)$ denote the σ-algebra generated by $u^1,\dots,u^t$; and let $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_0$ being the trivial σ-algebra. We then have that

$$\widehat{\mathbb{F}}^t(\xi^t) = \mathbb{E}_{\boldsymbol{u}^t}\left[g^t\right] = \mathbb{E}_t\left[g^t\right],$$

where the first equality follows from (5.10) and the the second equality holds since conditioned on $\mathcal{F}_{t-1}$, $\boldsymbol{u}^t$ is the only source of randomness in g^t and is sampled identically and independently in every round. Using that $\xi^t - \xi^*$ is measurable with respect to $\mathcal{F}_{t-1}$, we then obtain that

$$R_T \leq \mathbb{E}\sum_{t=1}^{T}\left\langle \mathbb{E}_t g^t, \xi^t - \xi^*\right\rangle + 4\delta LNT + 2\dot{\delta}LNT = \mathbb{E}\sum_{t=1}^{T}\left\langle g^t, \xi^t - \xi^*\right\rangle + 4\delta LNT + 2\dot{\delta}LNT\,.$$

Via the definition of ξ^t and the fact that $\xi^* \in \operatorname{int}(\mathcal{M}^p)^-$, Lemma 6.16 in (Orabona, 2023) implies that

$$\sum_{t=1}^{T}\left\langle g^t, \xi^t - \xi^*\right\rangle \leq \frac{D_{\psi_{\mathrm{lb}}}(\xi^*,\xi^1)}{\eta} + \sum_{t=1}^{T}\frac{\eta}{2}\|g_t\|^2_{\zeta^t,*},$$

where ζ^t lies on the line segment between ξ^t and ξ^{t+1}. We firstly observe that

$$\begin{aligned} \|g_t\|^2_{\xi^t,*} &= \left(\frac{(1-\delta)}{\delta}NS(A-1)F^t(\widehat{\mu}^t)\right)^2 \cdot (\boldsymbol{u}^t)^{\intercal}(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{1/2}\left(\nabla^2\psi_{\mathrm{lb}}(\xi^t)\right)^{-1}(\nabla^2\psi_{\mathrm{lb}}(\xi^t))^{1/2}\boldsymbol{u}^t \\ &= \left(\frac{(1-\delta)}{\delta}NS(A-1)F^t(\widehat{\mu}^t)\right)^2 \leq \frac{1}{\delta^2}N^4S^2A^2\,, \end{aligned}$$

where we have used that $F^t(\widehat{\mu}^t) \leq N$. Hence, if

$$\eta \leq \frac{\delta}{16N^2SA}, \tag{5.14}$$

then $\eta\|g_t\|_{\xi^t,*} \le 1/16$. Consequently, Lemma 5.9 (with $x = \xi^t$, $y = \xi^{t+1}$, and $\ell = \eta g_t$) would assert that $\xi^{t+1} \in \mathcal{E}_{1/2}(\xi^t)$; and hence, $\zeta^t \in \mathcal{E}_{1/2}(\xi^t)$. It would then hold via (5.12) that

$$\|g_t\|^2_{\zeta^t,*} \le \frac{1}{(1-\|\zeta^t-\xi^t\|_{\xi_t})^2}\|g_t\|^2_{\xi^t,*} \le 4\|g_t\|^2_{\xi^t,*},$$

On the other hand, via Lemma 5.10 and the definitions of ξ^1 and ξ^*, we have that

$$D_{\psi_{\mathrm{lb}}}(\xi^*,\xi^1) \le NSA\log\dot{\delta}^{-1}.$$

Hence, conditioned on (5.14), we obtain the following regret bound

$$R_T \le \frac{NSA\log\dot{\delta}^{-1}}{\eta} + \frac{2\eta}{\delta^2}N^4S^2A^2T + 4\delta LNT + 2\dot{\delta}LNT. \tag{5.15}$$

Setting

$$\dot{\delta} = \frac{1}{T}, \quad \eta = \frac{\delta}{16}\sqrt{\frac{\log T}{N^3SAT}}, \quad \text{and} \quad \delta = \min\left\{\sqrt{\frac{17}{4L}}\frac{N^{3/4}S^{3/4}A^{3/4}(\log T)^{1/4}}{T^{1/4}}, 1\right\};$$

we obtain that

$$\begin{aligned} R_T &\le \frac{NSA\log T}{\eta} + \frac{2\eta}{\delta^2}N^4S^2A^2T + 4\delta LNT + 2LN \\ &\le \frac{17}{\delta}\sqrt{N^5S^3A^3T\log T} + 4\delta LNT + 2LN \\ &\le \max\left\{4\sqrt{17L}N^{7/4}(SAT)^{3/4}(\log T)^{1/4}, 34\sqrt{N^5S^3A^3T\log T}\right\} + 2LN. \end{aligned} \tag{5.16}$$

If $T \ge NSA\log(T)$, then our choice of η indeed satisfies (5.14):

$$\eta = \frac{\delta}{16}\sqrt{\frac{\log T}{N^3SAT}} \le \frac{\delta}{16N^2SA}.$$

Otherwise, we can fall back to the trivial regret bound

$$R_T \le NT \le N^2SA\log(T),$$

which is dominated by the bound in (5.16); hence, the theorem follows. □

5.5 Future Directions

A natural extension to the work presented here is handling the case when the MDP is not known in advance but has to be estimated progressively from observed trajectories. One such extension to Algorithm 5.1 is provided in (Moreno et al., 2025), adapting to the bandit setting a technique used in (Moreno et al., 2024a). This technique consists in using the mirror descent algorithm over an evolving constraint set corresponding at each round to the set of occupancy measures associated with the current estimate of the MDP's dynamics. The achieved regret bound is $\widetilde{\mathcal{O}}\big(\sqrt{L(L+1)/\varepsilon}S^{5/4}A^{5/4}N^3T^{3/4}\big)$, which naturally suffers from a similar dependence on ε to that in the bound of Theorem 5.8.

Extending Algorithm 5.2 in the same manner seems unfortunately more challenging. A more promising approach in this direction would be adapting the techniques of Cohen et al. (2021), who address RL with trajectory-wise feedback via a similar algorithm. A notable element in their algorithm is an increasing learning rate schedule, adapting which to our setting would allow (also in the known MDP case) running our algorithm over a controllably modified MDP, which can be made to satisfy uniform reachability readily without modifying the representation of the decision set, and can be designed to exert some minimal control over the conditioning of the Hessian matrices $\nabla^2\psi_{\mathrm{lb}}(\xi^t)$.

Other interesting directions include proving high probability regret guarantees, improving the dependence on the dimension (i.e., the dependence on N, S, and A), and working towards designing practical approaches in the bandit convex MDP setting enjoying the optimal $\sqrt{T}$ rate.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04. Association for Computing Machinery, 2004.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, volume 3, pages 263–274, 2008.

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

Zeyuan Allen-Zhu, Sebastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 186–194. PMLR, 2018.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 26, 2013.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35. PMLR, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 373–382. PMLR, 2021.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 107–132. PMLR, 2011.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002b.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1753–1800. PMLR, 2023.

D. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific optimization and computation series. Athena Scientific, 2019.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26. PMLR, 2011.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 41.1–41.14. PMLR, 2012.

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *Journal of the ACM (JACM)*, 68(4):1–35, 2021.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122, 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23. PMLR, 2012.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 2020.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

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

Nicolò Cesa-Bianchi, Yoav Freund, David P Helmbold, David Haussler, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 382–391, 1993.

Nicolo 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.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 465–481. PMLR, 2017.

Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *arXiv preprint arXiv:2406.16802*, 2024.

Seok-Ho Chang, Pamela C. Cosman, and Laurence B. Milstein. Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011. doi: 10.1109/TCOMM.2011.072011.100049.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. In *Advances in Neural Information Processing Systems*, volume 34, pages 3401–3412. Curran Associates, Inc., 2021.

Houshuang Chen, Yuchen He, and Chihao Zhang. On interpolating experts and multi-armed bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6776–6802. PMLR, 2024.

Austin Coffman, Ana Bušić, and Prabir Barooah. A unified framework for coordination of thermostatically controlled loads. *Automatica*, 152:111002, 2023.

Alon Cohen, Tamir Hazan, and Tomer Koren. Tight bounds for bandit combinatorial optimization. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 629–642. PMLR, 2017.

Alon Cohen, Haim Kaplan, Tomer Koren, and Yishay Mansour. Online markov decision processes with aggregate bandit feedback. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1301–1329. PMLR, 2021.

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

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 93–104. PMLR, 2013.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR, 12–15 Jul 2023.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7288–7295, 2021.

Khaled Eldowa and Andrea Paudice. General tail bounds for non-smooth stochastic mirror descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3205–3213. PMLR, 2024.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information-theoretic regret bounds for bandits with fixed expert advice. In *2023 IEEE Information Theory Workshop (ITW)*, pages 30–35, 2023a.

Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 46122–46133. Curran Associates, Inc., 2023b.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 25(353):1–36, 2024.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34:28511–28521, 2021.

Emmanuel Esposito, Federico Fusco, Dirk Van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 34776–34788, 2022.

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

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394. Society for Industrial and Applied Mathematics, 2005.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online newton method for bandit convex optimisation, 2024.

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.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 489–497. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1259–1277. PMLR, 2020.

Adityanand Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

D. Haussler, J. Kivinen, and M.K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998. doi: 10.1109/18.705569.

Elad Hazan. Introduction to Online Convex Optimization, December 2021. arXiv preprint arXiv:1909.05207.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Elad Hazan and Yuanzhi Li. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2681–2691. PMLR, 2019.

Xiaowei Hu, Prashanth L.A., András György, and Csaba Szepesvari. (bandit) convex optimization with biased noisy gradient oracles. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 819–828, Cadiz, Spain, 09–11 May 2016. PMLR.

Shinji Ito. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Improved regret bounds for bandit combinatorial optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 28631–28643. Curran Associates, Inc., 2022.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 2020.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems*, volume 36, pages 30918–30978. Curran Associates, Inc., 2023.

Rémi Jézéquel, Dmitrii M. Ostrovskii, and Pierre Gaillard. Efficient and near-optimal online portfolio selection. *arXiv preprint arXiv:2209.13932*, 2022.

Satyen Kale. Multiarmed bandits with limited expert advice. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 107–122. PMLR, 13–15 Jun 2014.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Proceedings of a Symposium on the Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

Robert Kleinberg and Tom Leighton. The Value of Knowing a Demand Curve: Bounds on Regret for Online Posted-Price Auctions . In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, page 594. IEEE Computer Society, 2003.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2):245–272, 2010.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, volume 27, pages 613–621, 2014.

Wojciech Kotłowski. On minimaxity of follow the leader strategy in the stochastic setting. *Theoretical Computer Science*, 742:50–65, 2018. Algorithmic Learning Theory.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *The Journal of Machine Learning Research*, 21(1):5402–5446, 2020.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

Tor Lattimore. Bandit convex optimisation. *arXiv preprint arXiv:2402.06535*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jonathan Wilder Lavington, Sharan Vaswani, and Mark Schmidt. Improved policy optimization for online imitation learning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 1146–1173. PMLR, 2022.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Advances in Neural Information Processing Systems*, volume 33, pages 15522–15533. Curran Associates, Inc., 2020.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, 1989.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1739–1776. PMLR, 2018.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems*, volume 34, pages 22931–22942. Curran Associates, Inc., 2021.

Haipeng Luo, Hanghang Tong, Mengxiao Zhang, and Yuheng Zhang. Improved high-probability regret for adversarial bandits with time-varying feedback graphs. In *International Conference on Algorithmic Learning Theory*, pages 1074–1100. PMLR, 2023.

Maryam Majzoubi, Chicheng Zhang, Rajan Chari, Akshay Krishnamurthy, John Langford, and Aleksandrs Slivkins. Efficient contextual bandits with continuous actions. In *Advances in Neural Information Processing Systems*, volume 33, pages 349–360. Curran Associates, Inc., 2020.

Anuran Makur and Lizhong Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56:103–156, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.

H. Brendan McMahan and Matthew J. Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory*, pages 244–256. Omnipress, 2010.

Alberto Maria Metelli, Matteo Papini, Pierluca D'Oro, and Marcello Restelli. Policy optimization as online learning with mediator feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8958–8966, 2021.

Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38 (2):226–252, 2019.

Bianca Marin Moreno, Margaux Bregere, Pierre Gaillard, and Nadia Oudjane. Efficient model-based concave utility reinforcement learning through greedy mirror descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2206–2214. PMLR, 2024a.

Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, and Nadia Oudjane. Metacurl: Non-stationary concave utility reinforcement learning. *arXiv preprint arXiv:2405.19807*, 2024b.

Bianca Marin Moreno, Khaled Eldowa, Pierre Gaillard, Margaux Brégère, and Nadia Oudjane. Online episodic convex reinforcement learning. *arXiv preprint arXiv:2505.07303*, 2025.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 4489–4502. Curran Associates, Inc., 2022a.

Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in maximum state entropy exploration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16223–16239. PMLR, 2022b.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250): 1–42, 2023.

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

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1360–1375. PMLR, 2015a.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2023.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Pure exploration under mediators' feedback. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023. (draft).

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Maxim Raginsky. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

Kota Srinivas Reddy, P. N. Karthik, Nikhil Karamchandani, and Jayakrishnan Nair. Best arm identification in bandits with limited precision sampling. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1466–1471, 2023.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 2019.

Chloé Rouyer and Yevgeny Seldin. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 47(4):2815–2839, 2022.

Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 636–642. PMLR, 2011.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Yevgeny Seldin and Gábor Lugosi. A lower bound for multi-armed bandits with expert advice. In *The 13th European Workshop on Reinforcement Learning (EWRL)*, 2016.

Yevgeny Seldin, Peter Auer, John Shawe-taylor, Ronald Ortner, and François Laviolette. Pac-bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 1683–1691. Curran Associates, Inc., 2011.

Yevgeny Seldin, Koby Crammer, and Peter Bartlett. Open problem: Adversarial multiarmed bandits with limited advice. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 1067–1072. PMLR, 2013.

Rajat Sen, Karthikeyan Shanmugam, and Sanjay Shakkottai. Contextual bandits with stochastic experts. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 852–861. PMLR, 2018.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.

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, 12 1933.

Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

Dirk van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1285–1321. PMLR, 2023.

Nuri Mert Vural, Hakan Gokcesu, Kaan Gokcesu, and Suleyman S Kozat. Minimax optimal algorithms for adversarial bandit problem with multiple plays. *IEEE Transactions on Signal Processing*, 67(16):4383–4398, 2019.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

L.A. Wolsey and G.L. Nemhauser. *Integer and Combinatorial Optimization*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 1999.

Tom Zahavy, Brendan O' Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*, volume 34, pages 25746–25759. Curran Associates, Inc., 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020.

Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27574–27590. PMLR, 2022.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A

Supplementary Material for Chapter 2

A.1 Proof of Theorem 2.7

Theorem 2.7. *Suppose the policy set conforms to the multi-task structure. Then, for any algorithm and $T \geq \frac{K}{4\log(4/3)}$, there exists a sequence of losses such that $R_T \geq \frac{1}{18}\sqrt{KT}$.*

Proof. In the following, we will overload the notation and denote by $x_{i,\theta}$—which belongs to $\{x_{i,j}\}_{j=1}^{q}$—the outcome chosen by policy θ in section i (i.e. we have that $\theta(x_{i,\theta}) = 1/M$). For each policy θ, we construct an environment μ_θ such that for any outcome x, $\mu_\theta(x) \coloneqq 1/2 - \Delta\mathbb{I}\{x \in \mathrm{Supp}(\theta)\}$, where $0 \leq \Delta \leq 1/4$ is to be tuned later. Moreover, we will also use the following variations of each environment. For $i \in [M]$, let μ_θ^{-i} be an environment such that for any outcome x,

$$\mu_\theta^{-i}(x) \coloneqq \begin{cases} \frac{1}{2}, & \text{if } x \in \{x_{i,j}\}_{j=1}^{q} \\ \mu_\theta(x), & \text{otherwise.} \end{cases}$$

In words, μ_θ^{-i} is identical to μ_θ everywhere except in game i, where all outcomes are assigned a mean loss of $1/2$. For any policy θ, we have that

$$\begin{aligned} \overline{R}_T(\mu_\theta) &= \Delta\,\mathbb{E}_{\mu_\theta}\sum_{t=1}^{T}\sum_{i=1}^{M}(\theta(x_{i,\theta}) - \vartheta_t(x_{i,\theta})) \\ &= \frac{\Delta}{M}\mathbb{E}_{\mu_\theta}\sum_{t=1}^{T}\sum_{i=1}^{M}(1 - \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}) \\ &= \frac{\Delta}{M}\sum_{i=1}^{M}(T - N_{\mu_\theta}(i,\theta;T)), \end{aligned}$$

where for an environment μ, a policy θ, and a section $i \in [M]$, we define $N_\mu(i,\theta;T) := \mathbb{E}_\mu \sum_{t=1}^T \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}$. In words, this counts the expected number of times (under μ) that the chosen policy agrees with θ in section i. Next, we use that for any $i \in [M]$, $N_{\mu_\theta}(i,\theta;T) - N_{\mu_\theta^{-i}}(i,\theta;T) \le T\delta\big(P_{\mu_\theta^{-i}}, P_{\mu_\theta}\big)$ together with Pinsker's inequality to get that

$$\overline{R}_T(\mu_\theta) \ge \frac{\Delta}{M}\sum_{i=1}^M \left(T - N_{\mu_\theta^{-i}}(i,\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_\theta^{-i}} \,\big\|\, P_{\mu_\theta}\big)}\right). \tag{A.1}$$

For bounding the KL-divergence term, we start from Lemma 2.4:

$$\begin{aligned}
D\big(P_{\mu_\theta^{-i}} \,\big\|\, P_{\mu_\theta}\big) &= \sum_{\theta'\in\Theta} N_{\mu_\theta^{-i}}(\theta';T) \sum_{x\in\mathcal{X}} \theta'(x) d\big(\mu_\theta^{-i}(x) \,\big\|\, \mu_\theta(x)\big) \\
&= \sum_{\theta'\in\Theta} N_{\mu_\theta^{-i}}(\theta';T)\theta'(x_{i,\theta}) d\big(\mu_\theta^{-i}(x_{i,\theta}) \,\big\|\, \mu_\theta(x_{i,\theta})\big) \\
&= \frac{1}{M}\sum_{\theta'\in\Theta} \mathbb{I}\{x_{i,\theta'} = x_{i,\theta}\} N_{\mu_\theta^{-i}}(\theta';T) d\Big(\frac{1}{2} \,\Big\|\, \frac{1}{2} - \Delta\Big) \\
&\le \frac{c\Delta^2}{M}\sum_{\theta'\in\Theta} \mathbb{I}\{x_{i,\theta'} = x_{i,\theta}\} N_{\mu_\theta^{-i}}(\theta';T) \\
&= \frac{c\Delta^2}{M}\mathbb{E}_{\mu_\theta^{-i}}\sum_{t=1}^T \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\} = \frac{c\Delta^2}{M} N_{\mu_\theta^{-i}}(i,\theta;T)\,,
\end{aligned}$$

where the second equality holds since $x_{i,\theta}$ is the only outcome that does not have the same mean loss in the two environments, and the inequality holds for $\Delta \le 1/4$ with $c := 8\log(4/3)$. Plugging back into (A.1), we get that

$$\overline{R}_T(\mu_\theta) \ge \frac{\Delta}{M}\sum_{i=1}^M \left(T - N_{\mu_\theta^{-i}}(i,\theta;T) - T\Delta\sqrt{\frac{c}{2M} N_{\mu_\theta^{-i}}(i,\theta;T)}\right). \tag{A.2}$$

For what follows, we introduce an extra bit of notation. For each $i \in [M]$, we let $\sim_i$ denote an equivalence relation on the policy set such that for $\theta, \theta' \in \Theta$,

$$\theta \sim_i \theta' \iff \forall s \in [M]\backslash\{i\}, x_{s,\theta} = x_{s,\theta'}\,.$$

In words, two policies are equivalent according to $\sim_i$ if they agree everywhere outside of section i. Denote the set of all equivalence classes of $\sim_i$ by $\Theta/\sim_i$, which contains q^{M-1} classes, each containing q policies corresponding to the possible outcome choices in section i. For any $Y \in \Theta/\sim_i$, notice that if $\theta, \theta' \in Y$, then μ_θ^{-i} and $\mu_{\theta'}^{-i}$ denote the same environment, which will be referred to in the sequel as μ_Y^{-i}. Now, notice that for any

section i,

$$\sum_{\theta\in\Theta} N_{\mu_\theta^{-i}}(i,\theta;T) = \sum_{Y\in\Theta/\sim_i}\sum_{\theta\in Y} N_{\mu_Y^{-i}}(i,\theta;T) = \sum_{Y\in\Theta/\sim_i} \mathbb{E}_{\mu_Y^{-i}} \sum_{t=1}^{T} \underbrace{\sum_{\theta\in Y} \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}}_{=1} = q^{M-1}T\,,$$

whereas $\sum_{\theta\in\Theta}\sqrt{N_{\mu_\theta^{-i}}(i,\theta;T)} \le \sqrt{\sum_{\theta\in\Theta} 1^2}\sqrt{\sum_{\theta\in\Theta} N_{\mu_\theta^{-i}}(i,\theta;T)} = q^M\sqrt{\frac{T}{q}}$. Hence, we conclude that

$$\begin{aligned}
\sup_\mu \overline{R}_T(\mu) &\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\overline{R}_T(\mu_\theta) \\
&\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\frac{\Delta}{M}\sum_{i=1}^{M}\left(T - N_{\mu_\theta^{-i}}(i,\theta;T) - T\Delta\sqrt{\frac{c}{2M}N_{\mu_\theta^{-i}}(i,\theta;T)}\right) \\
&\ge \frac{\Delta}{M}\sum_{i=1}^{M}\left(T - \frac{1}{|\Theta|}q^M T\left(\frac{1}{q} + \Delta\sqrt{\frac{cT}{2qM}}\right)\right) \\
&= \Delta T\left(1 - \frac{1}{q} - \Delta\sqrt{\frac{cT}{2K}}\right) \overset{q\ge 2}{\ge} \Delta T\left(\frac{1}{2} - \Delta\sqrt{\frac{cT}{2K}}\right).
\end{aligned}$$

Plugging $\Delta := \frac{1}{4}\sqrt{\frac{2K}{cT}}$ into the previous display proves the theorem after observing that $\frac{1}{16}\sqrt{\frac{2}{c}} \ge \frac{1}{18}$. Lastly, notice that the condition imposed on T ensures that indeed $\Delta \le 1/4$. □

A.2 Proof of Theorem 2.9

Theorem 2.9. *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, under linear bandit feedback, we have that for any algorithm and $T \ge \frac{N}{8\varepsilon^2}$, there exists a sequence of losses (bounded in $[0,1]$) such that $R_T \ge \frac{1}{64\sqrt{2\log(16T)}}\sqrt{NT}$.*

Proof. Building on the result of Proposition 2.8, we follow the technique used in the proof of Theorem 5 in (Cohen et al., 2017). For the class of linear bandit environments specified in Section 2.6, we define

$$\widehat{R}_T(\mu) := \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))(\mu(x) + Z_t) = \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))\mu(x)$$

$$\widetilde{R}_T(\mu) := \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))\,\mathrm{clip}(\mu(x) + Z_t)\,,$$

where $\mathrm{clip}(a) := \max\{\min\{a,1\},0\}$. Notice that $\mathbb{E}\widehat{R}_T(\mu) \geq \overline{R}_T(\mu)$, and that $\sup_{(\ell_t)_t} R_T \geq \sup_\mu \mathbb{E}\widetilde{R}_T(\mu)$ considering sequences of losses $(\ell_t)_t$ bounded in $[0,1]$. We also define the event $A_\mu := \{\forall\, t \in [T],\, x \in \mathcal{X}\colon\ \mathrm{clip}(\mu(x)+Z_t) = \mu(x)+Z_t\}$. We will consider again the environments $\{\mu_\theta\}_{\theta\in\Theta}$ used in the proof of Proposition 2.8, recalling that $\mu_\theta(x) := 1/2 + \Delta\,((1-\varepsilon)/N - \mathbb{I}\{x = x_\theta\})$ for some $0 \leq \Delta \leq 1/2$. For any θ, we have that

$$\mathbb{E}\widehat{R}_T(\mu_\theta) = \mathbb{E}\big[\widehat{R}_T(\mu_\theta)\mathbb{I}\{A_{\mu_\theta}\}\big] + \mathbb{E}\big[\widehat{R}_T(\mu_\theta)\mathbb{I}\{A^c_{\mu_\theta}\}\big] \leq \mathbb{E}\big[\widetilde{R}_T(\mu_\theta)\big] + \Delta\varepsilon T\mathbb{P}\big(A^c_{\mu_\theta}\big), \tag{A.3}$$

where we have used the fact that $\widetilde{R}_T(\mu_\theta)$ and $\widehat{R}_T(\mu_\theta)$ are identical when A_{μ_θ} occurs, and that $\widehat{R}_T(\mu_\theta)$ is uniformly bounded by $\Delta\varepsilon T$ (see proof of Proposition 2.8). Assuming we enforce that $\Delta \leq 1/4$, the event $\{\mathrm{clip}(\mu(x)+Z_t) \neq \mu(x)+Z_t\}$ cannot hold for any outcome unless $|Z_t| > 1/4$. Hence, using a union bound and the fact that $Z_t \sim \mathcal{N}(0,\sigma^2)$, we get that

$$\mathbb{P}\big(A^c_{\mu_\theta}\big) \leq \sum\nolimits_t \mathbb{P}\big(|Z_t| > 1/4\big) \leq 2T\exp\left(\frac{-(1/4)^2}{2\sigma^2}\right).$$

Combining this with (A.3) and the fact that $\mathbb{E}\widehat{R}_T(\mu_\theta) \geq \overline{R}_T(\mu_\theta)$ allows us to conclude that

$$\sup_{(\ell_t)_t} R_T \geq \sup_\mu \mathbb{E}\widetilde{R}_T(\mu) \geq \frac{1}{N}\sum\nolimits_\theta \mathbb{E}\widetilde{R}_T(\mu_\theta) \geq \frac{1}{N}\sum\nolimits_\theta \overline{R}_T(\mu_\theta) - 2\Delta\varepsilon T^2\exp\left(\frac{-(1/4)^2}{2\sigma^2}\right).$$

Setting $\Delta := \frac{\sigma}{2\varepsilon}\sqrt{\frac{N}{T}}$, we obtain from the proof of Proposition 2.8 that $\frac{1}{N}\sum_\theta \overline{R}_T(\mu_\theta) \geq \frac{\sigma}{8}\sqrt{NT}$. Hence, choosing $\sigma := 1/(4\sqrt{2\log(16T)})$ entails that the required bound. Notice that the condition $T \geq \frac{N}{8\varepsilon^2}$ suffices to ensure that $\Delta \leq 1/4$. □

Appendix B

Supplementary Material for Chapter 3

B.1 Auxiliary Results

Lemma B.1. *If Algorithm 3.1 is run with $q \in (0,1)$, learning rate $\eta > 0$, and non-negative loss estimates that satisfy $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$ for all $t = 1, \dots, T$, then its regret satisfies*

$$R_T \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{\ell}_t(i)^2 \right].$$

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^{T} \ell_t(i)$ be an action that minimizes the cumulative loss, and let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^*. Recall that for $t \in [T]$, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$, and notice that p_t is measurable with respect to the σ-algebra generated by $I_1, \dots, I_{t-1}$. Hence, using that

$$\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i \in V} p_t(i)\ell_t(i) \qquad \text{and} \qquad \mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t,$$

we have, via the tower rule and the linearity of expectation, that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(i^*) = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \ell_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t \rangle\right],$$

from which we can obtain the desired result by using Lemma B.2 (which holds even if the loss $\widehat{\ell}_t$ at each round $t \in [T]$ depends on the prediction p_t made at that round). □

Lemma B.2. *Let $q \in (0,1)$, $\eta > 0$, and $(y_t)_{t=1}^{T}$ be an arbitrary sequence of non-negative loss vectors in $\mathbb{R}^K$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_K and the q-Tsallis*

regularizer ψ_q over this sequence of losses. That is, $p_1 = \arg\min_{p\in\Delta_K} \psi_q(p)$, and for $t \in [T]$,

$$p_{t+1} = \arg\min_{p\in\Delta_K} \eta \sum_{s=1}^{t} \langle y_s, p\rangle + \psi_q(p)\,.$$

Then for any $u \in \Delta_K$,

$$\sum_{t=1}^{T} \langle p_t - u, y_t\rangle \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i\in V} p_t(i)^{2-q}\, y_t(i)^2\,.$$

Proof. By Theorem 28.5 in (Lattimore and Szepesvári, 2020), we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t - u, y_t\rangle &\le \frac{\psi_q(u) - \psi_q(p_1)}{\eta} + \sum_{t=1}^{T} \Big(\langle p_t - p_{t+1}, y_t\rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \Big) \\
&= \frac{K^{1-q} - 1}{(1-q)\eta} + \sum_{t=1}^{T} \Big(\langle p_t - p_{t+1}, y_t\rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \Big) \\
&\le \frac{K^{1-q}}{(1-q)\eta} + \sum_{t=1}^{T} \Big(\langle p_t - p_{t+1}, y_t\rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \Big),
\end{aligned}$$

where $D_{\psi_q}(\cdot,\cdot)$ is the Bregman divergence based on ψ_q. For bounding each summand in the second term, we follow a similar argument to that used in Theorem 30.2 in (Lattimore and Szepesvári, 2020). Namely, for each $i \in V$ and round $t \in [T]$, define $\overline{y}_t(i) := \mathbb{I}\{p_{t+1}(i) \le$

$p_t(i)\}y_t(i)$. We then have that

$$\begin{aligned}
&\langle p_t - p_{t+1}, y_t\rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&\qquad \le \langle p_t - p_{t+1}, \overline{y}_t\rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&\qquad = \frac{1}{\eta}\langle p_t - p_{t+1}, \eta\overline{y}_t\rangle - \frac{1}{2\eta}\|p_{t+1} - p_t\|^2_{\nabla^2\psi_q(z_t)} \\
&\qquad \le \frac{\eta}{2}\|\overline{y}_t\|^2_{(\nabla^2\psi_q(z_t))^{-1}} \\
&\qquad = \frac{\eta}{2q}\sum_{i\in V} z_t(i)^{2-q}\,\overline{y}_t(i)^2 \\
&\qquad = \frac{\eta}{2q}\sum_{i\in V} \big(\gamma_t p_{t+1}(i) + (1-\gamma_t)p_t(i)\big)^{2-q}\,\overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q}\sum_{i\in V} p_t(i)^{2-q}\,\overline{y}_t(i)^2 + \gamma_t\frac{\eta}{2q}\sum_{i\in V}\big(p_{t+1}(i)^{2-q} - p_t(i)^{2-q}\big)\,\overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q}\sum_{i\in V} p_t(i)^{2-q}\,\overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q}\sum_{i\in V} p_t(i)^{2-q}\,y_t(i)^2,
\end{aligned}$$

where $z_t := \gamma_t p_{t+1} + (1-\gamma_t)p_t$ for some $\gamma_t \in [0,1]$; the first inequality holds due to the non-negativity of the losses, the second inequality is an application of the Fenchel-Young inequality, the second equality holds since the Hessian of ψ_q is a diagonal matrix with $(\nabla^2\psi_q(x))_{i,i} = qx(i)^{q-2}$, the third inequality is an application of Jensen's inequality (since $q \in (0,1)$), and the fourth inequality holds since $\overline{y}_t(i) = 0$ for any i such that $p_{t+1}(i)^{2-q} > p_t(i)^{2-q}$. □

Lemma B.3. *Let a and b be positive integers such that $2 \le a \le b$, and let $n := \lceil \log_2 a \rceil$. Then,*

$$\sum_{r=0}^{n-1}\sqrt{2^r \log\big(e^2 b 2^{-r}\big)} \le \frac{\sqrt{2\pi} + 2\sqrt{2-\log 2}}{\log 2}\sqrt{a\log\Big(\frac{e^2 b}{a}\Big)}.$$

Proof. Since $n \le \log_2(2b)$ and $2^r\log\big(e^2 b 2^{-r}\big)$ is monotonically increasing in r for r taking values in $[0, \log_2(eb)]$, we can bound the sum by an integral:

$$\sum_{r=0}^{n-1}\sqrt{2^r \log\big(e^2 b 2^{-r}\big)} \le \int_0^n \sqrt{2^r \log\big(e^2 b 2^{-r}\big)}\,\mathrm{d}r\,.$$

We proceed via a change of variable; let $x := e^2 b 2^{-r}$, and note that $\mathrm{d}r = -\frac{\mathrm{d}x}{x \log 2}$. We then have that

$$\begin{aligned}
\int_0^n \sqrt{2^r \log\left(e^2 b 2^{-r}\right)}\,\mathrm{d}r &= \sqrt{e^2 b}\int_0^n \sqrt{\frac{2^r}{e^2 b}\log\left(e^2 b 2^{-r}\right)}\,\mathrm{d}r \\
&= -\frac{e\sqrt{b}}{\log 2}\int_{e^2 b}^{e^2 b 2^{-n}} \sqrt{\frac{\log x}{x^3}}\,\mathrm{d}x = \frac{e\sqrt{b}}{\log 2}\int_{e^2 b 2^{-n}}^{e^2 b} \sqrt{\frac{\log x}{x^3}}\,\mathrm{d}x \\
&= \frac{e\sqrt{b}}{\log 2}\left[-\sqrt{2\pi}\cdot \mathrm{erfc}\left(\sqrt{(\log x)/2}\right) - 2\sqrt{(\log x)/x}\right]_{e^2 b 2^{-n}}^{e^2 b} \\
&\le \frac{e\sqrt{b}}{\log 2}\left(\sqrt{2\pi}\cdot \mathrm{erfc}\left(\sqrt{\log(e^2 b 2^{-n})/2}\right) + 2\sqrt{\frac{2^n \log(e^2 b 2^{-n})}{e^2 b}}\right),
\end{aligned}$$

where $\mathrm{erfc}(x) := 1 - \frac{2}{\sqrt{\pi}}\int_0^x \exp(-z^2)\,\mathrm{d}z$ is the complementary Gaussian error function, which is always positive. By (Chang et al., 2011, Theorem 1), we have that $\mathrm{erfc}(x) \le \exp(-x^2)$. Consequently,

$$\begin{aligned}
\int_0^n \sqrt{2^r \log\left(e^2 b 2^{-r}\right)}\,\mathrm{d}r &\le \frac{e\sqrt{b}}{\log 2}\left(\sqrt{2\pi}\sqrt{\frac{2^n}{e^2 b}} + 2\sqrt{\frac{2^n \log(e^2 b 2^{-n})}{e^2 b}}\right) \\
&= \frac{\sqrt{2^n}}{\log 2}\left(\sqrt{2\pi} + 2\sqrt{\log(e^2 b 2^{-n})}\right) \\
&\le \frac{\sqrt{2a}}{\log 2}\left(\sqrt{2\pi} + 2\sqrt{\log\left(\frac{e^2 b}{2a}\right)}\right) \\
&\le \frac{\sqrt{2\pi} + 2\sqrt{2 - \log 2}}{\log 2}\sqrt{a \log\left(\frac{e^2 b}{a}\right)},
\end{aligned}$$

where in the second inequality we used once again the fact that $2^r \log\left(e^2 b 2^{-r}\right)$ is monotonically increasing in r for $r \in [0, \log_2(eb)]$ to replace n with $\log_2(a)+1$, and the last inequality holds since $b/a \ge 1$. □

B.2 Proofs of Section 3.3

In this section, we provide the proof of Theorem 3.3, which is restated below.

Theorem 3.3. *Let $G_1, \ldots, G_T$ be a sequence of strongly observable undirected feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If*

Algorithm 3.1 is run with input

$$q = \frac{1}{2}\left(1 + \frac{\log(K/\alpha)}{\sqrt{\log(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}},$$

and loss estimates (3.6), then its regret satisfies $R_T \leq 6\sqrt{e\alpha T\left(2 + \log(K/\alpha)\right)}$.

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^{T} \ell_t(i)$ and $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be its indicator vector. Whenever J_t is nonempty, let $j_t \in V$ be the only action such that $J_t = \{j_t\}$. Similarly to (Zimmert and Lattimore, 2019), let $z_t := \mathbb{I}\{J_t \neq \emptyset\}\mathbb{I}\{I_t \in N_t(j_t)\}\frac{1-\ell_t(j_t)}{1-p_t(j_t)}$ and define new losses $\widetilde{\ell}_t(i) := \widehat{\ell}_t(i) + z_t$ for each time step $t \in [T]$ and each action $i \in V$. Since $p_t, \mathbf{e}_{i^*} \in \Delta_K$, we have that $\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ for every $t \in [T]$. Then, using the fact that $\mathbb{E}_t\left[\widehat{\ell}_t\right] = \ell_t$, we get that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right],$$

where the first equality holds via the same arguments made in the proof of Lemma B.1. If we consider the optimization step of Algorithm 3.1, computing the same inner product over the new losses $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_T$ for some $p \in \Delta_K$ gives

$$\left\langle \sum_{s=1}^{t} \widetilde{\ell}_s, p \right\rangle = \sum_{s=1}^{t} z_s + \left\langle \sum_{s=1}^{t} \widehat{\ell}_s, p \right\rangle,$$

where the sum $\sum_{s=1}^{t} z_s$ is constant with respect to p. This implies that the objective functions in terms of either $(\widehat{\ell}_t)_{t\in[T]}$ and $(\widetilde{\ell}_t)_{t\in[T]}$, respectively, are minimized by the same probability distributions. However, notice that, unlike $(\widehat{\ell}_t)_{t\in[T]}$, the loss vectors in $(\widetilde{\ell}_t)_{t\in[T]}$ are always non-negative. Consequently, similar to the proof of Lemma B.1, we may apply Lemma B.2 to upper bound the regret of Algorithm 3.1 in terms of the losses $(\widetilde{\ell}_t)_{t\in[T]}$. Doing so gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\left[\widetilde{\ell}_t(i)^2\right]\right]. \tag{B.1}$$

We can bound the second term by observing that $\widetilde{\ell}_t(j_t) = 1$ whenever $J_t \neq \emptyset$. Therefore,

$$\begin{aligned}\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big] &\le 2\sum_{i\in V\setminus J_t} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widehat{\ell}_t(i)^2\Big] + 2\,\mathbb{E}_t\big[z_t^2\big]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1\\ &\le 2\sum_{i\in V\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 2\,\mathbb{E}_t\big[z_t^2\big]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1\\ &\le 2\sum_{i\in V\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 3\,,\end{aligned}$$

where the second inequality holds because $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$ for all $i \notin J_t$, and the third inequality follows from the fact that

$$\mathbb{E}_t\big[z_t^2\big]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} = \mathbb{I}\{J_t \neq \emptyset\}\,\frac{\big(1-\ell_t(j_t)\big)^2}{1-p_t(j_t)}\sum_{i\in V\setminus J_t} p_t(i)^{2-q} \le 1\,.$$

We can handle the remaining sum by separating it over nodes $i \in S_t$, which satisfy $P_t(i) = 1 - p_t(i)$ because of strong observability, and those in $\overline{S}_t = V \setminus S_t$. In the first case, any node $i \in S_t \setminus J_t$ has $p_t(i) \le 1/2$ and thus

$$\sum_{i\in S_t\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} = \sum_{i\in S_t\setminus J_t}\frac{p_t(i)^{2-q}}{1-p_t(i)} \le 2\sum_{i\in S_t\setminus J_t} p_t(i)^{2-q} \le 2\,.$$

while in the second case we have that $\sum_{i\in\overline{S}_t} p_t(i)^{2-q}/P_t(i) \le \alpha^q$ by Lemma 3.1 with $U = \overline{S}_t$ and $b = 1-q$. Overall, we have shown that

$$\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big] \le 2\sum_{i\in\overline{S}_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 7 \le 2\alpha^q + 7 \le 9\alpha^q\,. \tag{B.2}$$

Plugging back into (B.1), we obtain that

$$\begin{aligned}R_T &\le \frac{K^{1-q}}{\eta(1-q)} + \frac{9\eta}{2q}\alpha^q T\\ &= 3\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}\,T}\\ &\le 6\sqrt{e\alpha T\big(2+\log(K/\alpha)\big)}\,,\end{aligned}$$

where the equality is due to our choice of η, and the last inequality follows as in the proof of Theorem 3.2 together with our choice of q. □

B.3 Proofs of Section 3.4

In this section, we provide the proof of Theorem 3.4, which is restated below.

Theorem 3.4. *Let* $C := 4\sqrt{6e}\frac{\sqrt{\pi}+\sqrt{4-2\log 2}}{\log 2}$. *Then, the regret of Algorithm 3.2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T} \alpha_t\Big(2+\log\Big(\frac{K}{\overline{\alpha}_T}\Big)\Big)+\log_2 \overline{\alpha}_T}\,.$$

Proof. Notice that if $\overline{\alpha}_T = 1$, the initial guess is correct and the algorithm will never restart. Moreover, since in this case we have that $\alpha_t = 1$ for all t, the theorem follows trivially from the regret bound of Theorem 3.3. Hence, we can assume for what follows that $\overline{\alpha}_T > 1$. Let $i^* \in \arg\min_{i\in[K]} \sum_{t=1}^{T} \ell_t(i)$ and $n := \lceil \log_2 \overline{\alpha}_T \rceil$. Note that the maximum value of r that the algorithm can reach is $n-1$. To see this, observe that Lemma 3.1 implies that for any r and t, $H_t(q_r) \le \alpha_t^{q_r}$. Consequently, for any $t \ge T_r$,

$$\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} \le \frac{1}{T}\sum_{s=T_r}^{t} \alpha_s \le \overline{\alpha}_T \le 2^n.$$

For $t \in [T]$, let r_t be the value of r at round t. Without loss of generality, we assume that r takes each value in $\{0,\dots,n-1\}$ for at least two rounds. Additionally, we define $T_n = T+2$ for convenience. We start by decomposing the regret over the n intervals (each corresponding to a value of r in $\{0,\dots,n-1\}$) and bounding the instantaneous regret with 1 for each step in which we restart (i.e., at the last step of each but the last interval):

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\big(\ell_t(I_t)-\ell_t(i^*)\big)\right] \\ &\le \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t)-\ell_t(i^*)\big)\right] + n - 1 \\ &\le \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t)-\ell_t(i^*)\big)\right] + \log_2 \overline{\alpha}_T\,. \end{aligned} \tag{B.3}$$

For what follows, let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^* and let $\widetilde{\ell}_t$ be as defined in the proof of Theorem 3.3. Fix $r \in \{0,\dots,n-1\}$, we proceed by bounding the regret in the

interval $[T_r, T_{r+1}-2]$:

$$
\begin{aligned}
&\mathbb{E}\Big[\textstyle\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t)-\ell_t(i^*)\big)\Big] \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t=r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\Big\}\big(\ell_t(I_t)-\ell_t(i^*)\big)\Bigg] \\
&\overset{(a)}{=} \mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t=r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\Big\}\langle p_t-\mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\Bigg] \\
&\overset{(b)}{=} \mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t=r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\Big\}\langle p_t-\mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{t=T_r}^{T_{r+1}-2}\langle p_t-\mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\Bigg] \\
&\overset{(c)}{\le} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\Bigg[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\Bigg] \\
&\overset{(d)}{=} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t=r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\Big\}\mathbb{E}_t\Big[\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\Big]\Bigg] \\
&\overset{(e)}{\le} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{I}\Big\{r_t=r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\Big\}(2H_t(q_r)+7)\Bigg] \\
&= \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\Bigg[\sum_{t=T_r}^{T_{r+1}-2}(2H_t(q_r)+7)\Bigg], \qquad \text{(B.4)}
\end{aligned}
$$

where (a) follows since $\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i\in V} p_t(i)\ell_t(i)$, $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$, and the indicator at round t is measurable with respect to $\sigma(I_1,\dots,I_{t-1})$, that is, the σ-algebra generated by $I_1,\dots,I_{t-1}$; (b) follows since $\langle p_t-\mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t-\mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ holds by the definition of $\widetilde{\ell}_t$; (c) is an application of Lemma B.2, justifiable with the same argument leading to (B.1) in the proof of Theorem 3.3; (d) uses once again that the indicator at round t is measurable with respect to $\sigma(I_1,\dots,I_{t-1})$; finally, (e) follows via (B.2). Define $T_{r:r+1} = T_{r+1}-T_r-1$, and notice that

$$
\begin{aligned}
\textstyle\sum_{t=T_r}^{T_{r+1}-2} H_t(q_r) &= \frac{T_{r:r+1}}{T_{r:r+1}}\textstyle\sum_{t=T_r}^{T_{r+1}-2}\big(H_t(q_r)^{1/q_r}\big)^{q_r} \\
&\le T_{r:r+1}\Big(\frac{1}{T_{r:r+1}}\textstyle\sum_{t=T_r}^{T_{r+1}-2} H_t(q_r)^{1/q_r}\Big)^{q_r} \\
&\le T_{r:r+1}\Big(\frac{T}{T_{r:r+1}}2^{r+1}\Big)^{q_r} \\
&\le 2T\big(2^r\big)^{q_r},
\end{aligned}
$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0,1)$, and the second follows from the restarting condition of Algorithm 3.2. Next, we plug this inequality back into (B.4), and then, similar to the proof of Theorem 3.3, we use the definitions of η_r and q_r and bound the resulting expression to get that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\left(\ell_t(I_t)-\ell_t(i^*)\right)\right] &\le \frac{K^{1-q_r}}{\eta_r(1-q_r)}+\frac{11\eta_r}{2q_r}T\left(2^r\right)^{q_r}\\ &\le 2\sqrt{11eT2^r\left(2+\log\left(K2^{-r}\right)\right)} \le 4\sqrt{3eT2^r\log\left(e^2K2^{-r}\right)}.\end{aligned}$$

We then sum this quantity over r and use Lemma B.3 to get that

$$\begin{aligned}\mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\left(\ell_t(I_t)-\ell_t(i^*)\right)\right] &\le 4\sqrt{3eT}\sum_{r=0}^{n-1}\sqrt{2^r\log\left(e^2K2^{-r}\right)}\\ &\le 4\sqrt{6e}\frac{\sqrt{\pi}+\sqrt{4-2\log 2}}{\log 2}\sqrt{\overline{\alpha}_T T\left(2+\log\left(K/\overline{\alpha}_T\right)\right)},\end{aligned}$$

which, together with (B.3), concludes the proof. □

B.4 Proof of the Lower Bound

In this section, we prove the lower bound provided in Section 3.5, which we restate below. As remarked before, parts of the proof are adapted from the proof of Theorem 2.7.

Theorem 3.5. *Pick any $K \ge 2$ and any α such that $2 \le \alpha \le K$. Then, for any algorithm and for all $T \ge \frac{\alpha \log_\alpha K}{4\log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1,\dots,G_T$ such that $\alpha(G_t)=\alpha$ for all $t=1,\dots,T$ and*

$$R_T \ge \frac{1}{18\sqrt{2}}\sqrt{\alpha T\log_\alpha K}.$$

Proof. Once again, we define $M := \log_\alpha K$, which we assume for now to be an integer; we discuss in the end how to extend the proof to the case when it is not. The proof will be divided into five parts I–V. We begin by formalizing the class of environments described in Section 3.5 and stating two useful lemmas.

I. Preliminaries

We remind the reader that we identify each action in V with a vector $a = \big(a(1), \dots, a(M)\big) \in [\alpha]^M$. We will focus on a set of M undirected graphs $\mathcal{G} = \{G^i\}_{i=1}^M$, where G^i consists of α isolated cliques (with self-loops) $\{C_{i,j}\}_{j=1}^{\alpha}$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. As remarked before, all these graphs have independence number α. For convenience, we also use actions in V as functions from $\mathcal{G}$ to $[\alpha]$, with $a(G^i) = a(i)$.

An environment is identified by a function $\mu\colon [\alpha] \times \mathcal{G} \to [0,1]$ such that at every round t, after having drawn a graph G_t from the uniform distribution over $\mathcal{G}$ (denoted with $U_{\mathcal{G}}$), the environment latently draws for each $j \in [\alpha]$ and $G \in \mathcal{G}$, a Bernoulli random variable $\gamma_t(j;G)$ with mean $\mu(j;G_t)$. Subsequently, for defining the loss of action $a \in V$ at round t, we simply set $\ell_t(a) = \gamma_t(a(G_t);G_t)$, whose expectation, conditioned on G_t, is $\mu(a(G_t);G_t)$. To simplify the notation, we use $\mu(a;G)$ as shorthand for $\mu(a(G);G)$ and $\gamma_t(a;G)$ as shorthand for $\gamma_t(a(G);G)$. Denote by A_t the action picked by the player at round t, which is chosen prior to observing G_t. We will focus on the following notion of stochastic regret, which we define for environment μ as:

$$\overline{R}_T(\mu) \coloneqq \max_{a \in V} \mathbb{E}_\mu \Big[\sum_{t=1}^{T} (\ell_t(A_t) - \ell_t(a)) \Big],$$

where $\mathbb{E}_\mu[\cdot]$ denotes the expectation with respect to the sequence of losses and graphs generated by environment μ, as well as the randomness in the choices of the player. We can use the tower rule to rewrite this expression as

$$\begin{aligned}
\overline{R}_T(\mu) &= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \Big[\mathbb{E}_\mu \Big[\mathbb{E}_\mu \Big[\ell_t(A_t) - \ell_t(a) \,\Big|\, G_t, A_t \Big] \,\Big|\, A_t \Big] \Big] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \Big[\mathbb{E}_\mu \Big[\mu(A_t;G_t) - \mu(a;G_t) \,\Big|\, A_t \Big] \Big] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \Big[\sum_{i=1}^{M} U_{\mathcal{G}}(G^i) (\mu(A_t;G^i) - \mu(a;G^i)) \Big] \\
&= \max_{a \in V} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_\mu \Big[\sum_{t=1}^{T} (\mu(A_t;G^i) - \mu(a;G^i)) \Big]. \qquad \text{(B.5)}
\end{aligned}$$

For a fixed algorithm, one can show via standard arguments that

$$\sup_{(\ell_t)_{t=1}^T, (G_t)_{t=1}^T} R_T \geq \sup_{\mu} \overline{R}_T(\mu).$$

Hence, it suffices for our purposes to prove a lower bound for the right-hand side of this inequality.

In the following, we will have to be more precise about the probability measure with respect to which the expectation in (B.5) is defined. Let $\boldsymbol{\lambda}_t \in \{0,1\}^{K/\alpha}$ denote the vector of losses observed by the player at round t, which corresponds to the losses of the actions connected to A_t assuming that a systematic ordering of the actions makes it clear which coordinate of $\boldsymbol{\lambda}_t$ belongs to which action. Let $\mathbf{1}_{K/\alpha}$ and $\mathbf{0}_{K/\alpha}$ be the K/α dimensional[1] vectors of all ones and all zeros respectively. Clearly, we have that $\boldsymbol{\lambda}_t = \gamma_t(A_t; G_t)\mathbf{1}_{K/\alpha} = \ell_t(A_t)\mathbf{1}_{K/\alpha}$, which is a binary random variable taking values in $\{\mathbf{0}_{K/\alpha}, \mathbf{1}_{K/\alpha}\}$. Let $\mathbb{P}_\mu^\lambda$ be the probability distribution of $\boldsymbol{\lambda}_t$ in environment μ. Notice then that we have that

$$\mathbb{P}_\mu^\lambda(\gamma_t = \mathbf{1}_{K/\alpha} \,|\, G_t = G, A_t = a) = \mu(a; G)\,. \tag{B.6}$$

Let $H_t := (A_1, G_1, \boldsymbol{\lambda}_1, \ldots, A_t, G_t, \boldsymbol{\lambda}_t) \in (V \times \mathcal{G} \times \{0,1\}^{K/\alpha})^t$ be the interaction trajectory after t steps. The policy π adopted by the player can be modelled as a sequence of probability kernels $\{\pi_t\}_{t=1}^T$ each mapping the trajectory so far to a distribution over the actions, i.e., A_t is sampled from $\pi_t(\cdot \,|\, H_{t-1})$. An environment μ and a policy π (implicit in the notation, and fixed throughout the rest of the proof) together define a distribution $\mathbb{P}_\mu$ over the set of possible trajectories of T steps such that:

$$\mathbb{P}_\mu(H_T) = \prod_{t=1}^{T} \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}_\mu^\lambda(\lambda_t \,|\, G_t, A_t)\,.$$

Similarly to Chapter 2, if P and Q are two distributions defined on the same space, we let $D(P \,\|\, Q)$ and $\delta(P, Q)$ be the KL-divergence and the total variation distance respectively between P and Q. Further, we also denote by $d(p \,\|\, q)$ the KL-divergence between two Bernoulli random variables with means p and q. The following lemma (an analogue of Lemma 2.4) provides an expression for the KL-divergence between two the probability distributions associated to two environments.

Lemma B.4. *For a fixed policy, let μ and μ' be two environments as described above. Then,*

$$D(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{a \in V} N_\mu(a; T) d\big(\mu(a; G^i) \,\big\|\, \mu'(a; G^i)\big)\,,$$

where $N_\mu(a; T) := \mathbb{E}_\mu\Big[\sum_{t=1}^T \mathbb{I}\{A_t = a\}\Big]$.

[1] Note that $K/\alpha = \alpha^{M-1}$ is an integer since $M(\geq 1)$ was assumed to be an integer.

Proof. The proof is similar to that of Lemma 15.1 in (Lattimore and Szepesvári, 2020). Namely, we have in our case that

$$\begin{aligned}
D(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) &= \mathbb{E}_\mu\left[\log \frac{\mathbb{P}_\mu(H_T)}{\mathbb{P}_{\mu'}(H_T)}\right] \\
&= \mathbb{E}_\mu\left[\log \frac{\prod_{t=1}^T \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}_\mu^\lambda(\lambda_t \,|\, G_t, A_t)}{\prod_{t=1}^T \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}_{\mu'}^\lambda(\lambda_t \,|\, G_t, A_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\log \frac{\mathbb{P}_\mu^\lambda(\lambda_t \,|\, G_t, A_t)}{\mathbb{P}_{\mu'}^\lambda(\lambda_t \,|\, G_t, A_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\log \frac{\mathbb{P}_\mu^\lambda(\lambda_t \,|\, G_t, A_t)}{\mathbb{P}_{\mu'}^\lambda(\lambda_t \,|\, G_t, A_t)} \,\middle|\, G_t, A_t\right] \,\middle|\, A_t\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[D(\mathbb{P}_\mu^\lambda(\cdot \,|\, G_t, A_t) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \,|\, G_t, A_t)) \,\middle|\, A_t\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\sum_{i=1}^M U_{\mathcal{G}}(G^i) D(\mathbb{P}_\mu^\lambda(\cdot \,|\, G^i, A_t) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \,|\, G^i, A_t))\right] \\
&= \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \mathbb{E}_\mu\left[D(\mathbb{P}_\mu^\lambda(\cdot \,|\, G^i, A_t) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \,|\, G^i, A_t))\right] \\
&= \frac{1}{M} \sum_{i=1}^M \sum_{a \in V} N_\mu(a; T) D(\mathbb{P}_\mu^\lambda(\cdot \,|\, G^i, a) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \,|\, G^i, a)) \\
&= \frac{1}{M} \sum_{i=1}^M \sum_{a \in V} N_\mu(a; T) d\big(\mu(a; G^i) \,\big\|\, \mu'(a; G^i)\big),
\end{aligned}$$

where the last equality holds via (B.6). □

The following standard lemma, adapted from (Lattimore and Szepesvári, 2020), will be used in the sequel.

Lemma B.5. *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$. Let $a < b$ and $X : \Omega \to [a, b]$ be an $\mathcal{F}$-measurable random variable. Then,*

$$\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b-a) \sqrt{\frac{1}{2} D(P \,\|\, Q)}.$$

Proof. We have, by Exercise 14.4 in (Lattimore and Szepesvári, 2020), that

$$\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b-a) \delta(P, Q),$$

from which the lemma follows by applying Pinsker's inequality. □

II. Choosing the environments

We will construct a collection of environments $\{\mu_a\}_{a\in V}$, each associated to an action, such that for any $i \in [M]$ and $j \in [\alpha]$,

$$\mu_a(j; G^i) := \frac{1}{2} - \varepsilon \mathbb{I}\{a(i) = j\},$$

where $0 < \varepsilon \leq \frac{1}{4}$ will be tuned later. In words, for a fixed graph, environment μ_a gives a slight advantage to actions that are connected to a in that graph, and thus agree with a in the corresponding game. Additionally, for every $a \in V$ and $i \in [M]$, we define environment μ_a^{-i} to be such that for any $s \in [M]$ and $j \in [\alpha]$,

$$\mu_a^{-i}(j; G^s) := \begin{cases} \frac{1}{2}, & \text{if } s = i \\ \mu_a(j; G^s), & \text{otherwise.} \end{cases}$$

Similar to the proof of Theorem 2.7, we will define, for every $i \in [M]$, an equivalence relation $\sim_i$ on the arms such that

$$a \sim_i a' \iff \forall s \in [M] \setminus \{i\}, a'(s) = a(s),$$

for any $a, a' \in V$. This means that two arms are equivalent according to $\sim_i$ if and only if their choices of base actions coincide in all games that are different from i. Let $V/\sim_i$ be the set of equivalence classes of $\sim_i$. It is easy to see that $V/\sim_i$ contains exactly α^{M-1} equivalence classes, and that each class consists of α actions, each corresponding to a different choice of base action in game i. Notice then that for an equivalence class $W \in V/\sim_i$, all environments μ_a^{-i} with $a \in V$ are indeed identical. In the sequel, this environment will also be referred to as μ_W^{-i}.

III. Lower-bounding the regret of a single environment

Note that in environment μ_a, we have that $a = \arg\min_{a' \in V} \sum_{i=1}^{M} \mu_a(a'; G^i)$. Consequently, starting from (B.5) we get that

$$\begin{aligned}
\overline{R}_T(\mu_a) &= \sum_{i=1}^{M} \frac{1}{M} \mathbb{E}_{\mu_a}\Big[\sum_{t=1}^{T} (\mu_a(A_t; G^i) - \mu_a(a; G^i))\Big] \\
&= \sum_{i=1}^{M} \frac{1}{M} \mathbb{E}_{\mu_a}\Big[\sum_{t=1}^{T} \left(\frac{1}{2} - \varepsilon \mathbb{I}\{A_t(i) = a(i)\} - \left(\frac{1}{2} - \varepsilon\right)\right)\Big] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^{M} \mathbb{E}_{\mu_a}\Big[\sum_{t=1}^{T} (1 - \mathbb{I}\{A_t(i) = a(i)\})\Big] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^{M} \big(T - N_{\mu_a}(i, a; T)\big),
\end{aligned}$$

where for environment μ, action a, and game i, $N_\mu(i, a; T) := \mathbb{E}_\mu\big[\sum_{t=1}^{T} \mathbb{I}\{A_t(i) = a(i)\}\big]$ is the expected number of times in environment μ that the action chosen by the policy agrees with action a in game i. Next, we use Lemma B.5 to obtain that

$$\overline{R}_T(\mu_a) \geq \frac{\varepsilon}{M} \sum_{i=1}^{M} \left(T - N_{\mu_a^{-i}}(i, a; T) - T\sqrt{\frac{1}{2} D\big(\mathbb{P}_{\mu_a^{-i}} \,\big\|\, \mathbb{P}_{\mu_a}\big)}\right). \tag{B.7}$$

For bounding the KL-divergence term, we start from Lemma B.4:

$$\begin{aligned}
D\big(\mathbb{P}_{\mu_a^{-i}} \,\big\|\, \mathbb{P}_{\mu_a}\big) &= \frac{1}{M}\sum_{s=1}^{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(\mu_a^{-i}(a';G^s) \,\big\|\, \mu_a(a';G^s)\big) \\
&= \frac{1}{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(\mu_a^{-i}(a';G^i) \,\big\|\, \mu_a(a';G^i)\big) \\
&= \frac{1}{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(1/2 \,\big\|\, 1/2-\varepsilon\mathbb{I}\{a'(i)=a(i)\}\big) \\
&= \frac{1}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\} N_{\mu_a^{-i}}(a';T)\, d\big(1/2 \,\big\|\, 1/2-\varepsilon\big) \\
&\le \frac{c\varepsilon^2}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\} N_{\mu_a^{-i}}(a';T) \\
&= \frac{c\varepsilon^2}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\}\, \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^{T} \mathbb{I}\{A_t=a'\} \\
&= \frac{c\varepsilon^2}{M}\, \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^{T} \mathbb{I}\{A_t(i)=a(i)\} \\
&= \frac{c\varepsilon^2}{M} N_{\mu_a^{-i}}(i,a;T)\,,
\end{aligned}$$

where the second equality holds since the two environments only differ in G^i, and the inequality holds for $\varepsilon \le \frac{1}{4}$ with $c := 8\log\frac{4}{3}$. Plugging back into (B.7) gets us that

$$\overline{R}_T(\mu_a) \ge \frac{\varepsilon}{M}\sum_{i=1}^{M}\Big(T - N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M} N_{\mu_a^{-i}}(i,a;T)}\Big). \tag{B.8}$$

IV. Summing up

Fix $i \in [M]$. Notice that for $W \in V/\sim_i$,

$$\sum_{a\in W} \mathbb{I}\{A_t(i)=a(i)\} = 1$$

since each action in W corresponds to a different choice of base action in game i. Hence,

$$\begin{aligned}
\frac{1}{\alpha^M}\sum_{a\in V} N_{\mu_a^{-i}}(i,a;T) &= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W} N_{\mu_a^{-i}}(i,a;T)\\
&= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W} N_{\mu_W^{-i}}(i,a;T)\\
&= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\mathbb{E}_{\mu_W^{-i}}\left[\sum_{t=1}^{T}\sum_{a\in W}\mathbb{I}\{A_t(i)=a(i)\}\right]\\
&= \frac{1}{\alpha^M}\alpha^{M-1}T = \frac{T}{\alpha}.
\end{aligned}$$

Using this together with (B.8) allows us to conclude that

$$\begin{aligned}
\sup_{\mu}\overline{R}_T(\mu) &\geq \frac{1}{\alpha^M}\sum_{a\in V}\overline{R}_T(\mu_a)\\
&\geq \frac{1}{\alpha^M}\sum_{a\in V}\frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M}N_{\mu_a^{-i}}(i,a;T)}\right)\\
&\geq \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - \frac{1}{\alpha^M}\sum_{a\in V}N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M\alpha^M}\sum_{a\in V}N_{\mu_a^{-i}}(i,a;T)}\right)\\
&= \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - \frac{T}{\alpha} - \varepsilon T\sqrt{\frac{cT}{2M\alpha}}\right)\\
&= \varepsilon T\left(1 - \frac{1}{\alpha} - \varepsilon\sqrt{\frac{cT}{2M\alpha}}\right)\\
&\geq \varepsilon T\left(\frac{1}{2} - \varepsilon\sqrt{\frac{cT}{2M\alpha}}\right),
\end{aligned}$$

where the third inequality holds due to the concavity of the square root, and the last inequality holds by our assumption that $\alpha \geq 2$. Setting $\varepsilon = \frac{1}{4}\sqrt{\frac{2M\alpha}{cT}}$ yields that

$$\sup_{\mu}\overline{R}_T(\mu) \geq \frac{1}{16}\sqrt{\frac{2}{c}}\cdot\sqrt{\alpha TM} \geq \frac{1}{18}\sqrt{\alpha TM} = \frac{1}{18}\sqrt{\alpha T\log_\alpha K},$$

whereas it holds that $\varepsilon \leq \frac{1}{4}$ thanks to the assumption made on T in the statement of the theorem.

V. The case when $\log_\alpha K$ is not an integer

If M is not an integer,[2] we can use the same construction as before for the first $\alpha^{\lfloor M \rfloor}$ actions and force the remaining actions to behave identically to some action in the construction. That is, we can designate a certain action such that, in all environments, all the excess actions receive the same loss as this action and are connected to it, to each other, and to every action that happens to share an edge with this designated action in a given graph (in other words, we are expanding the designated action into a clique). This way, the independence number of all the graphs in the construction is still α, and the excess actions do not provide any extra utility to the learner; playing one of them is exactly like playing the designated action, and the construction does not hide this from the player. We can then obtain the same bound as before but in terms of $\lfloor M \rfloor$, thus costing us an extra $1/\sqrt{2}$ factor to recover the desired bound (using that $\lfloor M \rfloor \geq M/2$). □

[2] Note that M is never smaller than 1 since $\alpha \leq K$.

Appendix C

Supplementary Material for Chapter 4

C.1 Auxiliary Results

Lemma C.1. *Let $q \in (0,1)$, $b > 0$, $c > 1$, and $(y_t)_{t=1}^T$ be a sequence of loss vectors in $\mathbb{R}^N$ satisfying $y_t(i) \geq -b$ for all $t \in [T]$ and $i \in [N]$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_N and the q-Tsallis regularizer ψ_q over this sequence of losses; that is, $p_1 = \arg\min_{p \in \Delta_N} \psi_q(p)$, and for $t \in [T]$,*

$$p_{t+1} = \operatorname*{arg\,min}_{p \in \Delta_N} \eta \sum_{s=1}^{t} \langle y_s, p \rangle + \psi_q(p),$$

assuming the learning rate η satisfies $0 < \eta \leq \frac{q}{(1-q)b}\left(1 - c^{\frac{q-1}{2-q}}\right)$. Then for any $u \in \Delta_N$,

$$\sum_{t=1}^{T} \langle p_t - u, y_t \rangle \leq \frac{N^{1-q} - 1}{(1-q)\eta} + \frac{\eta c}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} p_t(i)^{2-q}\, y_t(i)^2.$$

Proof. Let $p'_{t+1} := \arg\min_{p \in \mathbb{R}^N_{\geq 0}} \langle p, y_t \rangle + D_{\psi_q}(p, p_t)$, where $D_{\psi_q}(\cdot,\cdot)$ denotes the Bregman divergence based on ψ_q. Via Lemma 7.14 in Orabona (2023) we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t - u, y_t \rangle &\leq \frac{\psi_q(u) - \psi_q(p_1)}{\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2 \\
&\leq \frac{N^{1-q} - 1}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2,
\end{aligned}$$

where z_t lies on the line segment between p_t and p'_{t+1}. A simple derivation shows that

$$p'_{t+1}(i) = p_t(i)\left(\frac{1}{1+\eta\frac{1-q}{q}y_t(i)p_t(i)^{1-q}}\right)^{\frac{1}{1-q}},$$

for each $i \in [N]$. On the other hand, it holds that

$$\eta\frac{1-q}{q}y_t(i)p_t(i)^{1-q} \geq -\eta\frac{1-q}{q}bp_t(i)^{1-q} \geq -\eta\frac{1-q}{q}b \geq c^{\frac{q-1}{2-q}} - 1,$$

where the first inequality uses that $y_t(i) \geq -b$ (and that $p_t(i), \eta > 0$), the second uses that $p_t(i) \leq 1$, and the third uses that $\eta \leq \frac{q}{(1-q)b}\left(1 - c^{\frac{q-1}{2-q}}\right)$. This entails that $p'_{t+1}(i) \leq c^{\frac{1}{2-q}}p_t(i)$, which implies that $z_t(i) \leq c^{\frac{1}{2-q}}p_t(i)$ concluding the proof. □