Model Selection for Contextual Bandits and Reinforcement Learning

Aldo Pacchiano

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-164
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-164.html

June 25, 2021

Model Selection for Contextual Bandits and Reinforcement Learning

by

Aldo Pacchiano Camacho

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Co-chair
Professor Michael Jordan, Co-chair
Professor Paul Grigas

Spring 2021

Model Selection for Contextual Bandits and Reinforcement Learning

Abstract

Model Selection for Contextual Bandits and Reinforcement Learning

by

Aldo Pacchiano Camacho

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter Bartlett, Co-chair

Professor Michael Jordan, Co-chair

In many domains ranging from internet commerce, to robotics and computational biology, many algorithms have been developed that make decisions with the objective of maximizing a reward, while learning how to make better decisions in the future. In hopes of realizing this objective a vast literature focused on the study of Bandits and Reinforcement Learning algorithms has arisen. Although in most practical applications, precise knowledge of the nature of the problem faced by the learner may not be known in advance most of this work has chiefly focused on designing algorithms with provable regret guarantees that work under specific modeling assumptions. Less work has been spent on the problem of model selection where the objective is to design algorithms that can select in an online fashion the best suitable algorithm among a set of candidates to deal with a specific problem instance.

In this thesis we provide a comprehensive set of algorithmic approaches to the problem of model selection in stochastic contextual bandits and reinforcement learning. We propose and analyze two distinct approaches to the problem. First, we introduce Stochastic CORRAL, an algorithm that successfully combines an adversarial EXP3 or CORRAL master with multiple stochastic bandit algorithms. Second, we introduce three distinct stochastic master algorithms: Explore-Commit-Eliminate (ECE), Regret Balancing, and Regret Bound Balancing and Elimination (RBBE) that recover the rates of Stochastic CORRAL under an EXP3 and a CORRAL master but with the advantage the model selection guarantees of RBBE extend to the setting of contextual linear bandits with adversarial contexts. We complement our algorithmic results with

a variety of lower bounds designed to explore the theoretical limits of model selection in Contextual Bandits and Reinforcement Learning.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

As the last year of my PhD comes to a close, I cannot avoid stopping to think about all the experiences and the people behind them that have shaped me throughout this time. First I would like to thank my parents and my sister for believing in me and giving me the steadfast support that has propelled me throughout the years. I wouldn't be writing this thesis without it.

The years that I have spent during my PhD have been a fun and exciting ride, full of stimulating intellectual discussions, and generally a jovial and friendly atmosphere. I want to thank my advisors Peter Bartlett and Michael Jordan for structuring their research groups in a way that not only helped me grow academically, but allowed for a fair bit of comradery with other fellow students, fomenting research collaborations and friendships that will hopefully transcend my time at UC Berkeley. I want to particularly emphasize my gratitude towards Peter, who believed enough in the proposition that I would be a good student to help me transfer to UC Berkeley barely a few months after starting my PhD at a different institution. I am extremely happy that transfer was possible. To Mike, thanks so much for letting me be part of your group and for nudging all of us to pursue interesting research directions.

To Mark Rowland and Adrian Weller, thanks for giving me the opportunity to collaborate with you when I was still quite new to the research world. It gave me the confidence that in research a bit of persistence can go a far way. A particularly warm thank you to fellow PhD student Niladri Chatterji not only for being a great collaborator but also a great friend. The time we worked on 'Online Learning with Kernel Losses' [17] was a particularly important growth experience. Despite being quite raw, we came up with a research problem which we tackled, solved and published. Very many thanks to to Krzysztof Choromanski, for the multitude of fun projects and papers that we have worked on, and for introducing me to the other members of the Klab such as Jack Parker-Holder and Yunhao Tang.

Great many thanks to the rest of the Berkeley crowd, Kush Bhatia, Yeshwanth Cherapanamjeri, Ghassen Jerfel, Nicolas Flammarion, Yian Ma, Eric Mazumdar, Juan Carlos Pedromo and Nilesh Tripuraneni for the many collaborations and fun times during these years. I want to make a special mention of Jonathan Lee, who I started collaborating with when he was an undergraduate student at UC Berkeley and who is currently a PhD student at Stanford University. Jonathan's diligence and enthusiasm for research have been the basis of one of the most pleasurable research collaborations that I participated in during my PhD. Another very special thank you to Silvia Chiappa and Heinrich Jiang for introducing me to the world of ML Fairness. I have had lots of great and interesting discussions with both of you. I hope we can continue collaborating in the future.

Among other collaborators I want to thank Yasin Abbasi-Yadkori, Emma Brunskill, Christoph Dann, Miro Dudik, Jakob Foerster, Claudio Gentile, Mohammad Ghavamzadeh, Tor Lattimore, Thodoris Lykouris, My Phan, Robert Schapire and Csaba Szepesvari. I have learned a lot from each and every one of you. I very much appreciate all the time you have spent in our shared discussions. To everyone else that I have collaborated with whose name couldn't fit in this small note, thank you so much.

Chapter 1

Introduction

Recent advances in machine learning and compute have made the presence of learning algorithms ubiquitous in many aspects of modern life. The vast majority of internet powered services ranging from social networks, email providers, marketplaces and search engines rely on online algorithms to produce engaging content to users while allowing to further learn about their preferences. The use of these algorithmic techniques is not limited to the internet domain. In applications where machine learning is starting to have a profound impact, such as computational biology, economics, or robotics, the use of methods that allow to simultaneously learn about the system's objective while at the same time produce near optimal solutions has also become of paramount importance.

The need for developing algorithms than can simultaneously learn while taking near optimal decisions has led to the formalism of online learning and more specifically to the study of Bandits and Reinforcement Learning problems. There is a vast literature revolving around the objective of designing algorithms with provable regret guarantees for Bandits and Reinforcement learning problems. The vast majority of these works have focused on settings with specific model assumptions. This has given rise to a rich literature that includes among others algorithms such as UCB [9] for the Multi Armed bandit problem, OFUL [1] for linear bandits and UCBVI [10] for tabular Reinforcement Learning.

In many applications, the appropriate model may not be known in advance by the learner. For example, when faced with a contextual bandit instance with multiple arms and changing contexts, the learner may be unaware if there is a linear reward structure underlying the rewards or if these rewards instead follow a simpler multi armed bandit model. In this case, selecting the appropriate algorithm adapted to the true structure of the problem is of paramount importance. For example if the learner was facing a problem over T time-steps with K arms and contexts of dimension d, and the problem structure was truly linear, the learner would greatly benefit from using a linear bandits algorithm instead of a multi armed bandit one. This is true for two reasons, first because the reward achievable by an algorithm leveraging the linear structure of the contexts could be much higher than that reachable by an algorithm that ignores it, and second because if the problem is truly linear a linear bandit algorithm such as OFUL will accrue regret (up to logarithmic factors) of order $\widetilde{\mathcal{O}}(d\sqrt{T})$ whereas UCB would instead collect a regret of order $\widetilde{\mathcal{O}}(\sqrt{KT})$, a quantity that could be substantially higher than the former in case $d \ll \sqrt{K}$. Conversely, if the problem does not satisfy a linear model, it may be prejudicial for the learner to use a linear bandit algorithm because if the problem is not truly linear, a linear bandit algorithm such as OFUL may incur a linear regret[1].

It is thus important to develop methods that can deal with model uncertainty by allowing to select in an online fashion among multiple algorithms the one that is best adapted to the problem instance at hand. The problem of model selection in contextual bandits and reinforcement learning focuses on this task. In this thesis we shall focus on the study of the problem of model selection for stochastic contextual bandits and reinforcement learning.

The algorithms we develop in this work follow the template of [5] where a 'master' algorithm is placed on top of a couple of 'base' algorithms. At the beginning of each round the master selects which base algorithm to 'listen to' during that time-step effectively treating the base algorithms as arms to be pulled by the master. The difficulty in using existing algorithms such as UCB or EXP3 [13] as a master lies in the non-stationary nature of the rewards collected by a learning base algorithm. The master needs to be sufficiently smart to recognize when a base algorithm is simply performing poorly because it still in the early stages of learning from the case where poor performance is the result of model misspecification.

The problem of model selection in online decision-making environments with limited-information feedback (which includes both bandits and reinforcement learning), has been an active area of recent research as witnessed by a proliferation of recent works (e.g., [4, 5, 7, 12, 18, 26, 27, 28, 43]). Broadly speaking previous works on model selection can be split into two types of methods,

1. **Adversarial Master.** Methods that make use of an adversarial algorithm such as EXP3 to select the base algorithm to listen to.

2. **Statistical Test.** Methods that perform a statistical test to detect when a base algorithm is misspecified.

[1] We henceforth refer to an algorithm that is not well adapted to the environment where it is played to be misspecified.

Adversarial Master. In the first group are the so-called *corraling* algorithms. These algorithms satisfy regret guarantees of the form $\mathcal{O}(d_\star^\alpha T^\beta)$ for $\alpha \geq 1, \beta < 1$, where $d_\star$ depends on the complexity of the best model class or algorithm adapted to the problem at hand. The original CORRAL Algorithm of [5] uses an adversarial master algorithm that can be combined with many base algorithms (both stochastic and adversarial) whenever these base algorithms satisfy a stability guarantee. Unfortunately, in order to show that a base algorithm can be combined with the corraling master to satisfy a valid model selection regret guarantee, it is necessary to verify this stability condition is satisfied, something that has to be done on a case-by-case basis. There exist other related approaches in the literature that use adversarial corralling master algorithms to performing model selection. For example [7] make use of a Tsallis-INF adversarial master in an algorithm they show is able to obtain gap-dependent guarantees for stochastic bandit problems. Unfortunately the model selection guarantees achievable by their approach depend not on the complexity of the optimal model class, but on the size of the largest model class under consideration. This means that whenever the rates of the input base algorithms are of the form $\{d_i T^\alpha\}_{i=1}^M$, where $d_1 \leq \cdots \leq d_M$, their master algorithm satisfies a regret guarantee scaling with d_M instead of $d_{i_\star}$, a quantity that could be substantially smaller.

Statistical Test In the second group of approaches that use a statistical test to detect misspecification, model selection regret guarantees have been shown under strong eigenvalue conditions on the context distribution in the setting of stochastic linear contextual bandits setting. When the contextual information is stochastic, [27] obtain model selection guarantees of the form $\mathcal{O}(d_*^{1/3} T^{2/3})$ under an action-averaged eigenvalue condition, and [18] match the optimal guarantee when choosing between multi-armed bandits and contextual bandits under a stronger universal eigenvalue condition that ensures that contexts corresponding to all arms are sufficiently diverse. The results of [27] leverage the fact that it is possible to estimate the optimal value under the optimal model at a rate of $\sqrt{d}/n$ finding the optimal policy under the complex model (which has estimation error rate d/n). Both of these algorithms work by keeping a collection of active base learners, and playing a low complexity algorithm/model in the active set. When enough information is obtained to conclude that a higher complexity model would be more adequate to describe the observed data, they eliminate the low complexity model from the active set, and proceed to play a more complex one.

Model selection in Reinforcement Learning. Though there has been some work on offline feature selection and model selection for RL given a batch of data (see e.g. [24, 31, 35, 55]), there has been very little work specifically on online model selection in reinforcement learning. Prior work provided PAC results for online feature selection for factored tabular MDPs [30]. More recent work provides PAC bounds [48] for model selection in online RL when the optimal value is known: however, unlike contextual bandits [27, 40], there are no known algorithms for estimating the optimal value faster than identifying the optimal policy in RL settings.

Main Results and Organization of the Thesis

The remainder of the thesis is organized into five chapters. The subject of the first four is to introduce a variety of procedures for model selection that chart the approach space outlined in the previous discussion. We propose and analyze four model selection algorithmic techniques. The first of which, Stochastic CORRAL allows to combine an adversarial master with multiple base algorithms satisfying high probability regret guarantees. Two of the three remaining algorithmic approaches proposed in this thesis, Explore-Commit-Eliminate and Regret Bound Balancing are based on the principle of a statistical test to detect misspecification among the base algorithms. The two of them use different exploration schedules. The remaining approach, Simple Regret Balancing is an algorithmic technique that without the need of an adversarial master nor an elimination procedure, yields model selection guarantees among base algorithms with the same putative regret guarantee[2]. In the last chapter we describe future directions of research. Below we provide an overview of the results in each of the coming chapters.

Stochastic CORRAL In Chapter 2 we expand on the CORRAL algorithm from [5] and introduce Stochastic CORRAL, an algorithmic approach that successfully combines an adversarial EXP3 or CORRAL master with multiple stochastic bandit algorithms. Our approach is flexible enough that it does not require the algorithm designer to have any information about the inner workings of the base algorithms, other than knowing they may satisfy (if not misspecified) a high probability regret guarantee. In this chapter we also present two minimax lower bounds showing,

A) It is impossible to distinguish between logarithmic and square root base learners.

B) Knowledge of the target regret guarantee is necessary for perfect model selection.

Work in this chapter is joint with My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore and Csaba Szepesvari. It is based on the paper [54].

[2]Just as in every other approach we propose, not all of these algorithms has to be well adapted to the environment at hand.

Explore-Commit-Eliminate In Chapter 3 we introduce the Explore-Commit-Eliminate Algorithm (ECE). This algorithm makes use of a simple misspecification detection procedure that allows it to weed out base algorithms that are not well adapted to the problem at hand. In contrast with Stochastic CORRAL, the set of 'active' base algorithms shrinks as time progresses. ECE uses an exploration schedule reminiscent of $\epsilon-$greedy approaches to the Multi Armed Bandit problem. The model selection regret guarantees achievable by ECE recover the EXP3 rates of Stochastic CORRAL. Additionally, ECE can be shown to achieve gap-dependent model selection regret guarantees, a result that to our knowledge is not possible with Stochastic CORRAL. Work on this chapter is joint with Jonathan Lee, Vidya Muthukumar, Weihao Kong and Emma Brunskill. It is based on the paper [43].

Regret Balancing In Chapter 4 we describe an approach to model selection based on the principle of regret balancing (equating the empirical regret guarantees of multiple base algorithms). We show this idea can yield surprisingly simple algorithms with meaninfgul model selection guarantees for stochastic contextual bandits and reinforcement learning. We complement our results with a lower bound showing that any 'perfect' model selection procedure must be doing a form of regret balancing. Work on this chapter is joint with Yasin Abbasi-Yadkori and My Phan. It is based on the paper [4].

Regret Bound Balancing and Elimination In Chapter 5 we expand on the basic principle of regret balancing described in Chapter 4 and introduce Regret Bound Balancing and Elimination (RBBE), an algorithmic procedure that makes use of the same statistical test as ECE to eliminate misspecified base algorithms, but that follows an exploration schedule dictated by a regret balancing condition. The model selection regret guarantees achievable by RBBE recover the CORRAL master rates of Stochastic CORRAL. Similar to ECE, RBBE achieves gap-dependent model selection regret guarantees. We also show that when applied to the problem of model selection for linear stochastic bandits RBBE is versatile enough to also cover cases where the context information is generated by an adversarial environment. Work on this chapter is joint with Christoph Dann, Claudio Gentile and Peter Bartlett. It is based on the paper [53].

All of our algorithms recover meaningful model selection rates in several applications, including linear bandits and MDPs with nested function classes, linear bandits with unknown misspecification, and OFUL applied to linear bandits with different confidence parameters.

Chapter 2

Stochastic CORRAL

2.1 Introduction

Bandit algorithms have been applied in a variety of decision making and personalization problems in industry. There are many specialized algorithms each designed to perform well in specific environments. For example, algorithms are designed to exploit low variance [8], extra context information and linear reward structure [1, 20, 44], sparsity [2, 15], etc. The exact properties of the current environment however might not be known in advance, and we might not know which algorithm is going to perform best.

Model selection in contextual bandits aims to solve this problem. More formally, the learner is tasked to solve a bandit problem for which the appropriate bandit algorithm to use is not known in advance. Despite this limitation, the learner does have access to M different algorithms $\{\mathcal{B}_i\}_{i=1}^M$, one of which $\mathcal{B}_{i_\star}$ is promised to be adequate for the problem the learner wishes to solve. We use regret to measure the learner's performance[1]. The problem's objective is to design algorithms that would minimize regret.

Adapted and misspecified algorithms We say that an algorithm is *adapted* to the environment at hand if it satisfies a valid regret guarantee. Let's illustrate this with an example in the setting of linear bandits with finitely many arms. In this problem the learner has access to K arms. Each arm $i \in [K]$ is associated with a feature vector $z_i \in \mathbb{R}^d$, and the reward of arm $i \in [K]$ follows a linear model of the form $r_i = \langle z_i, \theta_\star \rangle + \xi_i$ where ξ_i is conditionally zero mean and $\theta_\star$ is an unknown parameter. An algorithm such as LinUCB [19] achieves a regret guarantee

[1] We will define regret more formally in the following section.

of order $\widetilde{\mathcal{O}}(\sqrt{d \log^3(K) T})$ where $\widetilde{\mathcal{O}}$ hides logarithmic factors in T. In contrast, the UCB algorithm [9] yields a regret guarantee of order $\widetilde{\mathcal{O}}(\sqrt{KT})$. In this case, both algorithms are well adapted to the problem of linear bandits with finitely many actions, but LinUCB's regret guarantee may be substantially smaller than UCB's regret upper bound if d is much smaller than K. If an algorithm is not well adapted, we say it is *misspecified*. For the sake of exposition let's assume we are in a similar setting as above, where the learner has access to K arms each of which is associated with a feature vector $z_i \in \mathbb{R}^d$. Instead of assuming a linear model as before, let's instead assume that $r_i = (\langle z_i, \theta_\star \rangle)^2 + \xi_i$ is quadratic. In this case, there is no reason to believe LinUCB can yield a valid regret guarantee since the underlying linearity assumption of LinUCB is violated. We say that in this case LinUCB is misspecified. Consider an instance of LinUCB that instead uses matrix features of the form $z_i z_i^\top$. In this case the quadratic reward is again a linear function of the feature vectors since $(\langle z_i, \theta_\star \rangle)^2 = \langle z_i z_i^\top, \theta_\star \theta_\star^\top \rangle$. Thus this version of LinUCB with quadratic features is adapted.

We will assume that all algorithms $\mathcal{B}_i$ for $i \in [M]$ are associated with a putative regret guarantee $U_i(t, \delta)$ *known* by the learner and holding with probability $1 - \delta$ for all $t \in [\mathbb{N}]$ *if algorithm i is adapted* to the environment at hand. If the learner knew the identity of the best adapted algorithm $i_\star$, it would be able to incur regret of order $U_{i_\star}(T, \delta)$ by playing $\mathcal{B}_{i_\star}$. The learner's objective in the model selection problem is to design a procedure that would allow a learner to incur in regret that is competitive with the regret upper bound $U_{i_\star}(t, \delta)$ of the best adapted algorithm among those in $\{\mathcal{B}_i\}_{i=1}^M$, so that the regret incurred by the learner up to time T scales as a function of T, the parameters defining $\mathcal{B}_{i_\star}$ and possibly M. From now on we will refer to each of the M algorithms in $\mathcal{B}_{i_\star}$ as a *base algorithm*. We will alert the reader if we have a specific set of M algorithms in mind. In any other case, when we talk about the set of base algorithms we simply mean a set of M algorithms the learner is hoping to model select from.

The authors of [49] were perhaps the first to address the bandit model-selection problem, with a variant of EXP4 master algorithm that works with UCB or EXP3 base algorithms. These results are improved by [5] via the CORRAL algorithm. The CORRAL algorithm follows the master-base template that we discussed in Chapter 1. It makes use of a CORRAL master based on a Log-Barrier Online Mirror Descent algorithm controlling which of the base algorithms to play at any given round. Let p_t be the probability distribution over the M base algorithms given by the CORRAL master. The learner will then sample an algorithm index $j_t \in [M]$ with $j_t \sim p_t$. and play the action prescribed by $\mathcal{B}_{j_t}$ to collect a reward r_t. All algorithms $\{\mathcal{B}_i\}_{i=1}^M$ are then updated using an importance weighted version of r_t regardless of whether they were selected by the master or not.

Unfortunately, this means that in order to use a base algorithm in CORRAL, this needs to be compatible with this importance weighting modification of the rewards. For example, to use UCB as a base, we would need to manually re-derive UCB's confidence intervals and modify its regret analysis to be compatible with importance weighted feedback. The authors show that a base algorithm can be safely combined with the CORRAL master to yield model selection guarantees provided it satisfies a stability condition (see Definition 3 in [5]). Verifying that an algorithm satisfies such stability condition is a cumbersome process that requires a detailed analysis of the algorithm's internal workings. In this work we instead focus on devising a black-box procedure that can solve the model selection problem for a general class of stochastic contextual bandit algorithms.

Contributions. We focus on the problem of bandit model-selection in stochastic environments. Our contributions are as follows:

- We introduce Stochastic CORRAL, a two step per round algorithm and an accompanying base "smoothing" wrapper that can be shown to satisfy model selection guarantees when combined with any set of M stochastic contextual bandit algorithms that satisfy a high probability regret guarantee when adapted. We also show model selection regret guarantees for Stochastic CORRAL with two distinct adversarial master algorithms, CORRAL [5] and EXP3.P [13]. Our approach is more general than that of the original CORRAL algorithm [5] because instead of requiring each base algorithm to be individually modified to satisfy a certain stability condition, our version of the CORRAL algorithm provides the algorithm designer with a generic black-box wrapper that allows to do model selection over any set of M base algorithm with high probability regret guarantees. Stochastic CORRAL has another important difference with respect to the original CORRAL algorithm: instead of importance weighted feedback, the reward r_t is sent to algorithm $\mathcal{B}_{j_t}$, and only this algorithm is allowed to update its internal state at round t. The main consequence of these properties of Stochastic CORRAL is that our model selection strategy can be used with almost any base algorithm developed for stochastic environments. When the optimal base regret is known, the CORRAL master achieves optimal regret guarantees. Under certain conditions when the optimal base regret is unknown EXP3.P can achieve better performance.

- We demonstrate the generality and effectiveness of our method by showing how it seamlessly improves existing results or addresses open questions in a variety of problems. We show applications in adapting to the misspecification level in contextual linear bandits [42], adapting to the unknown dimension in nested linear bandit classes [27], tuning the data-dependent exploration rate of bandit algorithms, and choosing feature maps in reinforcement learning. Moreover, our master algorithm can simultaneously perform different types of model selection. For example, we show how to choose both the unknown dimension and the unknown mis-specification error at the same time. This is in contrast to algorithms that specialize in a specific type of model selection such as detecting the unknown dimension [27].

- In the stochastic domain, an important question is whether a model selection procedure can inherit the $O(\log T)$ regret of a fast stochastic base algorithm. We show a lower bound for the model selection problem that scales as $\Omega(\sqrt{T})$, which implies that our result is minimax optimal. Our master algorithm requires knowledge of the best base's regret to achieve the same regret. We show that this condition is unavoidable in general: there are problems where regret of the best base scales as $O(T^x)$ for an unknown x, and the regret of any master algorithm scales as $\Omega(T^y)$ for $y > x$.

2.2 Problem Statement

Let δ_a denotes the delta distribution at a. For an integer n, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We consider the following formulation of contextual stochastic bandits. At the beginning of each time-step t, the learner observes a context $\mathcal{A}_t$ coming from a set of contexts. After this the learner will select an action a_t and then collect a reward $r_t = f(\mathcal{A}_t, a_t) + \xi_t$, a noisy quantity that will depends on the context $\mathcal{A}_t$, and the learner's action a_t, a reward function f and a $1-$subGaussian conditionally zero mean random noise random variable ξ_t. In this Chapter we will restrict ourselves to the case where contexts sets $\mathcal{A}_t$ are all subsets of a context generating set $\mathcal{A}$. This is in fact a very general scenario that captures all types of contextual bandit problems ranging from the case of changing linear contexts with linear rewards, to more general contexts and reward sets studied in works such as [25]. For simplicity we will assume the contexts $\mathcal{A}_t \subset \mathbb{R}^d$ are parametrized as a subset of action features. Our formulation allows for the action set to vary in size from round to round and even to be infinite. For example, the finite linear contextual bandit setting (where $\mathcal{A}_t = \mathcal{A}$ for all t) fits in this setting. Similarly it is easy to see the linear contextual bandit problem with i.i.d. contexts and K actions can also be written as an instance of our formulation. In the linear contextual bandit problem with K actions the learner is presented at time t with K action-vectors $\mathcal{A}_t = \{a_i\}_{i=1}^K$ with $a_i \in \mathbb{R}^d$ and the (random) return r_a of any action $a \in \mathcal{A}$ satisfies $r_a = \langle a, \theta_\star \rangle + \xi$.

In this work we focus on the setting of stochastic i.i.d. contexts. Let S be the set of all subsets of $\mathcal{A}$ and let $\mathcal{D}_S$ be a distribution over S. We assume all contexts $\mathcal{A}_t \overset{i.i.d.}{\sim} D_S$ and that $f : S \times \mathcal{A} \to \mathbb{R}$. Let $\mathcal{X} \subset \mathcal{A}$ and denote by $\Delta_{\mathcal{X}}$ to the space of distributions over $\mathcal{X}$. For any policy $\pi : \mathcal{X} \to \Delta_{\mathcal{X}}$. Let's denote by Π as the space of all policies with domain in Support($\mathcal{D}_S$). We abuse notation and denote $f(\mathcal{X}, \pi) = \mathbb{E}_{a\sim\pi}\left[f(\mathcal{X}, a)\right]$. Notice that in this case $f(\mathcal{X}, a) = f(\mathcal{X}, \delta_a)$ for all $a \in \mathcal{A}$.

In a contextual bandit problem the learner chooses policy π_t at time t, which takes context set $\mathcal{A}_t \in S$ as an input and outputs a distribution over $\mathcal{A}_t$. The learner then selects an action $a_t \sim \pi_t(\mathcal{A}_t)$ and receives a reward r_t such that $r_t = f(\mathcal{A}_t, \delta_{a_t}) + \xi_t$.

We are interested in designing an algorithm with small regret, defined as

$$R(T) = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} f(\mathcal{A}_t, \pi) - \sum_{t=1}^{T} f(\mathcal{A}_t, \pi_t)\right]. \tag{2.1}$$

If for example $U_i(T, \delta) = c d_i \sqrt{T \log(1/\delta)}$ for all $i \in [M]$ we would like our algorithm to satisfy a regret guarantee of the form $R(T) \le \mathcal{O}(M^\alpha d_{i_\star}^\beta \sqrt{T \log(1/\delta)})$ for some $\alpha \ge 0, \beta \ge 1$ and where $i_\star$ is the index of the best performing adapted base algorithm $\mathcal{B}_{i_\star}$. Crucially, we want to avoid this guarantee to depend on other $d_i > d_{i_\star}$ (if any). From now on we will refer to the policy maximizing the right hand side of the equation above as π^*. For simplicity we will also make the following assumption regarding the range of f,

Assumption 2.2.1 (Bounded Expected Rewards). *The absolute value of f is bounded by 1,*

$$\max_{\mathcal{A}', \pi} |f(\mathcal{A}', \pi)| \le 1$$

Throughout this work we assume the base algorithms we want to model select from satisfy a high probability regret bound whenever they are well adapted to their environment. We make this more precise in definition 2.2.1,

Definition 2.2.1 ((U, δ, T)−Boundedness). *Let $U : \mathbb{R} \times [0, 1] \to \mathbb{R}^+$. We say an adapted algorithm $\mathcal{B}$ is $(U, \delta, T)-$bounded if with probability at least $1 - \delta$ and for all rounds $t \in [1, T]$, its cumulative pseudo-regret is bounded above by $U(t, \delta)$: $\sum_{l=1}^t f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}_l, \pi_l) \le U(t, \delta)$.*

We assume that for all $i \in [M]$ the base algorithm $\mathcal{B}_i$ is (U_i, δ, T)-bounded for a function U_i known to the learner[2]. For example in the Multi Armed Bandit Problem with K arms the UCB algorithm is $(c\sqrt{KT\log(T/\delta)}, \delta, T)-$bounded for some universal constant $c > 0$.

Original CORRAL

We start by reproducing the pesudo-code of CORRAL [5] (see Algorithm 1) as it will prove helpful in our discussion of our main algorithm: Stochastic CORRAL. Recall that CORRAL follows the master-base template that we discussed in Chapter 1. As we have explained in the previous section we assume there are M candidate *base* algorithms and a master algorithm which we denote as $\mathcal{M}$. At time-step t the CORRAL master $\mathcal{M}$ selects one of the base algorithms in $\{\mathcal{B}_i\}_{i=1}^M$ according to a distribution $p_t \in \Delta_M$ by sampling an index $j_t \sim p_t$. The learner plays action $a_t \sim \pi_{t,j_t}(\mathcal{A}_t)$ and receives reward $r_t = f(\mathcal{A}_t, \delta_{a_t}) + \xi_t$. An importance weighted version of r_t is sent out to all base algorithms, after which all of them update their internal state.

Algorithm 1: Original CORRAL

Input: Base Algorithms $\{\mathcal{B}_j\}_{j=1}^M$, learning rate η.
Initialize: $\gamma = 1/T, \beta = e^{\frac{1}{\ln T}}, \eta_{1,j} = \eta, \rho_1^j = 2M, \underline{p}_1^j = \frac{1}{\rho_1^j}, p_1^j = 1/M$ for all $j \in [M]$.
Initialize all base algorithms.
for $t = 1, \cdots, T$ **do**
    Receive context $\mathcal{A}_t \sim \mathcal{D}_S$.
    Receive policy $\pi_{t,j}$ from $\mathcal{B}_j$ for all $j \in [M]$.
    Sample $j_t \sim p_t$.
    Play action $a_t \sim \pi_{t,j_t}(\mathcal{A}_t)$.
    Receive feedback $r_t = f(\mathcal{A}_t, \delta_{a_t}) + \xi_t$.
    Send feedback $\frac{r_t}{p_{t,j_t}}\mathbf{1}\{j = j_t\}$ to $\mathcal{B}_j$ for all $j \in [M]$.
    Update p_t, η_t, $\underline{p}_t$ and ρ_t to p_{t+1}, η_{t+1}, $\underline{p}_{t+1}$ and ρ_{t+1} via $\mathsf{CORRAL-Update}$

---

[2]Recall that in this case the upper bound on the algorithm's regret is satisfied only when $\mathcal{B}_i$ is well adapted to the environment.

Algorithm 2: Log-Barrier-OMD(p_t, ℓ_t, η_t)

Input: learning rate vector η_t, previous distribution p_t and current loss ℓ_t
Output: updated distribution p_{t+1}
Find $\lambda \in [\min_j \ell_{t,j}, \max_j \ell_{t,j}]$ such that $\sum_{j=1}^{M} \frac{1}{\frac{1}{p_t^j} + \eta_{t,j}(\ell_{t,j} - \lambda)} = 1$
Return p_{t+1} such that $\frac{1}{p_{t+1}^j} = \frac{1}{p_t^j} + \eta_{t,j}(\ell_{t,j} - \lambda)$

Algorithm 3: $\mathsf{CORRAL} - \mathsf{Update}$

Input: learning rate vector η_t, distribution p_t, lower bound $\underline{p}_t$ and current loss r_t
Output: updated distribution p_{t+1}, learning rate η_{t+1} and loss range ρ_{t+1}
Update $p_{t+1} = \text{Log-Barrier-OMD}(p_t, \frac{r_t}{p_{t,j_t}} \mathbf{e}_{j_t}, \eta_t)$.
Set $p_{t+1} = (1 - \gamma) p_{t+1} + \gamma \frac{1}{M}$.
for $j = 1, \cdots, M$ **do**
    if $\underline{p}_t^j > p_{t+1}^j$ **then**
        Set $\underline{p}_{t+1}^j = \frac{p_{t+1}^j}{2}, \eta_{t+1,j} = \beta \eta_{t,i}$,
    else
        Set $\underline{p}_{t+1}^j = \underline{p}_t^j, \eta_{t+1,j} = \eta_{t,i}$.
    Set $\rho_{t+1}^j = \frac{1}{\underline{p}_{t+1}^j}$.
Return p_{t+1}, η_{t+1}, $\underline{p}_{t+1}$ and ρ_{t+1}^j.

2.3 The Stochastic CORRAL Algorithm

In order to better describe the feedback structure of Stochastic CORRAL we abstract the master-base interaction template discussed in Chapter 1 into Algorithms 4 and 5. As we have mentioned before, one crucial difference between Stochastic CORRAL and CORRAL is that in Stochastic CORRAL only the state of the base algorithm whose action was selected is modified. In contrast in the CORRAL algorithm all the base algorithms' states are updated at every step.

To make this description more precise we introduce some notation. Each base algorithm $\mathcal{B}_i$ maintains a counter $s_{t,i}$ that keeps track of the number of times it has been updated up to time t. For any base algorithm $\mathcal{B}_j$, $\pi_{s,j}$ is the policy $\mathcal{B}_j$ uses at state s. Let $s_{t,j}$ denote the state of base j at time t. If $t_1 < t_2$ are two consecutive times when base j is chosen by the master, then base j proposed policy $\pi_{s_{t_1,j},j}$ at time t_1 and policy $\pi_{s_{t_2,j},j}$ during all times $t_1+1,\dots,t_2$ where $s_{t_2,j} = s_{t_1,j}+1$.

Algorithm 4: Master Algorithm

Input: Base Algorithms $\{\mathcal{B}_j\}_{j=1}^M$ **for** $t = 1,\cdots,T$ **do**
  Sample $j_t \sim p_t$.
  Play j_t.
  Receive feedback $r_t = r_{t,j_t}$ from playing the action prescribed by $\mathcal{B}_{j_t}$
  Update master using r_t

Algorithm 5: Base Algorithm $\mathcal{B}_j$

Initialize state counter $s = 1$ **for** $t = 1,\cdots,T$ **do**
  Receive action set $\mathcal{A}_t \sim \mathcal{D}_S$
  Choose action $a_{t,j} \sim \pi_{s,j}(\mathcal{A}_t)$
  if *selected by master (i.e. $j_t = j$)* **then**
    Play action $a_{t,j}$
    Receive feedback $r_{t,j} = f(\mathcal{A}_t, \delta_{a_{t,j}}) + \xi_t$
    Send $r_{t,j}$ to the master
    Compute $\pi_{s+1,j}$ using $r_{t,j}$
    $s \leftarrow s+1$

Regret Decomposition. Let's introduce the regret decomposition we will make use of to prove our regret guarantees. This is a similar decomposition as the one appearing in the proofs of Theorem 4,5 and 7 of [5]. We split the regret $R(T)$ of Equation 2.1 into two terms (I and II) by adding and subtracting terms $\{f(\mathcal{A}_t, \pi_{s_{t,i_\star},i_\star})\}_{t=1}^T$:

$$\begin{aligned} R(T) &= \mathbb{E}\left[\sum_{t=1}^T f(\mathcal{A}_t,\pi^*) - f(\mathcal{A}_t,\pi_t)\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^T f(\mathcal{A}_t,\pi_{s_{t,i_\star},i_\star}) - f(\mathcal{A}_t,\pi_t)\right]}_{\text{I}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T f(\mathcal{A}_t,\pi^*) - f(\mathcal{A}_t,\pi_{s_{t,i_\star},i_\star})\right]}_{\text{II}} \end{aligned} \tag{2.2}$$

Term I is the regret of the master with respect to the optimal base $\mathcal{B}_{i_\star}$, and term II is the regret of the optimal base with respect to the optimal policy π^*. Analysis of term I is largely based on the adversarial regret guarantees of the Log-Barrier-OMD in CORRAL and of the EXP3.P algorithm.

In order to bound term II we will have to further modify the feedback structure of Algorithms 4 and 5. In Algorithm 8 from Section 2.4 we introduce a smoothing procedure that allows any (U, δ, T)-bounded algorithm to be transformed into a 'smoothed' version of itself such that its conditional expected instantaneous regret is bounded with high probability during every even step. We name this procedure 'smoothing' because it is based on playing uniformly from the set of previously played policies during the smoothed algorithm's odd steps. We provide more details in section 2.4. For now, the main property we are to use from this discussion is that by smoothing a (U, δ, T)-bounded algorithm it is possible to ensure the conditional expected instantaneous regret of the smoothed algorithm is bounded above by $\frac{U(\ell,\delta)}{\ell}$ during the $\ell-$th even step. The function $\frac{U(\ell,\delta)}{\ell}$ can be shown to be decreasing (as a function of ℓ) when $U(\ell, \delta)$ is concave in ℓ. In Stochastic CORRAL the smoothing of base algorithms takes the place of the stability condition required by the CORRAL algorithm in [5].

Let's sketch some intuition behind why this decreasing instantaneous regret condition can help us bound term II. For all $i \in [M]$ let $\{p_1^i, \dots, p_T^i\}$ be the (random) probabilities used by the Stochastic CORRAL master $\mathcal{M}$ (an adversarial master algorithm) to chose base i during round t and let $\underline{p}_i = \min_t p_t^i$. Let's focus on the optimal algorithm $i_\star$ and assume $U_\star(t, \delta)$ is convex in t. Since $\frac{U_\star(t,\delta)}{t}$ is decreasing, term II is the largest when base $i_\star$ is selected the least often. For the sake of the argument let's assume that $p_t^{i_\star} = \underline{p}_{i_\star}$ $\forall t$. In this case base i will be played roughly $T\underline{p}_i$ times, and will repeat its decisions in intervals of length $\frac{1}{\underline{p}_i}$, resulting in the following bound:

Lemma 2.3.1 (informal)**.** *If regret of the optimal base is (U_*, T, δ)-bounded, then we have that*

$$\mathbb{E}\left[\text{II}\right] \leq O\left(\mathbb{E}\left[\frac{1}{\underline{p}_i} U_*(T\underline{p}_i, \delta) \log T\right] + \delta T(\log T + 1)\right).$$

We demonstrate the effectiveness of our smoothing transformation by deriving regret bounds with two master algorithms: the Log-Barrier-OMD algorithm in CORRAL (introduced by [5]) which we will henceforth refer to as the CORRAL master and EXP3.P (Theorem 3.3 in [13]) with forced exploration, a simple algorithm that ensures each base is picked with at least a (horizon dependent) constant probability

p which we will henceforth refer to as an EXP3.P master. We now state an informal version of our main result, Theorem 2.4.11.

Theorem 2.3.2 (informal version of Theorem 2.4.11)**.** *If $U_*(T,\delta) = O(c(\delta)\, T^\alpha)$ for some function $c : \mathbb{R} \to \mathbb{R}$ and constant $\alpha \in [1/2, 1)$ and $\mathcal{B}_*$ is (U,T,δ)-bounded, the regrets of Stochastic CORRAL with an EXP3.P and CORRAL masters are:*

Master	*Known α and $c(\delta)$*	*Known α, Unknown $c(\delta)$*
EXP3.P	$\widetilde{O}\left(T^{\frac{1}{2-\alpha}} c(\delta)^{\frac{1}{2-\alpha}}\right)$	$\widetilde{O}\left(T^{\frac{1}{2-\alpha}} c(\delta)\right)$
CORRAL	$\widetilde{O}\left(T^\alpha c(\delta)\right)$	$\widetilde{O}\left(T^\alpha c(\delta)^{\frac{1}{\alpha}}\right)$

The CORRAL master achieves optimal regret when α and $c(\delta)$ are known. When $c(\delta)$ is unknown and $c(\delta) > T^{\frac{(1-\alpha)\alpha}{2-\alpha}}$ (which is $T^{1/6}$ if $\alpha = 1/2$ or $\alpha = 1/3$), then using an EXP3.P master achievess better regret because $\widetilde{O}\left(T^{\frac{1}{2-\alpha}} c(\delta)\right) < \widetilde{O}\left(T^\alpha c(\delta)^{\frac{1}{\alpha}}\right)$.

We complement this result with a couple of lower bounds.

Lower bounds. Theorem 2.5.3 in Section 2.5 shows that if the regret of the best base is $O(T^x)$, in the worst case a master algorithm that does not know x can have regret $\Omega(T^y)$ with $y > x$. Theorem 2.5.2 shows that in general it is impossible for any master algorithm to achieve a regret better than $\Omega(\sqrt{T})$ even when the best base has regret $O(\log(T))$. When the regret of the best base is $O(\sqrt{T})$, CORRAL with our smoothing achieves the optimal $O(\sqrt{T})$ regret.

The detailed description of the aforementioned smoothing procedure, its properties and the regret analysis are postponed to Section 2.4. We also show some applications of our model selection results in Section 2.6.

Master Algorithms

Let's now review the different adversarial bandit algorithms that can be used as a Master in Algorithm 4.

CORRAL Master

We reproduce the CORRAL master algorithm below.

Algorithm 6: CORRAL Master

Input: Base Algorithms $\{\mathcal{B}_j\}_{j=1}^{M}$, learning rate η.
Initialize: $\gamma = 1/T, \beta = e^{\frac{1}{\ln T}}, \eta_{1,j} = \eta, \rho_1^j = 2M, \underline{p}_1^j = \frac{1}{\rho_1^j}, p_1^j = 1/M$ for all $j \in [M]$. **for** $t = 1, \cdots, T$ **do**
    Sample $i_t \sim p_t$.
    Receive feedback r_t from base $\mathcal{B}_{i_t}$.
    Update p_t, η_t, $\underline{p}_t$ and ρ_t to p_{t+1}, η_{t+1}, $\underline{p}_{t+1}$ and ρ_{t+1} using $\mathsf{CORRAL-Update}$ Algorithm 3.

EXP3.P Master

We reproduce the EXP3.P algorithm (Figure 3.1 in [14]) below. In this formulation we use $\eta = 1, \gamma = 2\beta k$ and $p = \frac{\gamma}{k}$.

Algorithm 7: EXP3.P Master

Input: Base Algorithms $\{\mathcal{B}_j\}_{j=1}^{M}$, exploration rate p.
Initialize: $p_1^j = 1/M$ for all $j \in [M]$.
for $t = 1, \cdots, T$ **do**
    Sample $i_t \sim p_t$.
    Receive feedback r_t from base $\mathcal{B}_{i_t}$.
    Compute the estimated gain for each base j: $\widetilde{r}_{t,j} = \frac{r_{t,j}\mathbf{1}_{i_t=j}+p/2}{p_{j,t}}$ and update the estimated cumulative gain $\widetilde{R}_{j,t} = \sum_{s=1}^{t} \widetilde{r}_{s,j}$. **for** $j = 1, \cdots, M$ **do**
        $p_{t+1}^j = (1-p)\frac{\exp \widetilde{R}_{j,t}}{\sum_{n=1}^{M} \exp \widetilde{R}_{n,t}} + p$

2.4 Smoothed Algorithm and Regret Analysis

Non-increasing Instantaneous Regret

We introduce a "smoothing" procedure (Algorithm 8) which, given a $(U, \delta, T)-$bounded algorithm $\mathcal{B}$ constructs a smoothed algorithm $\widetilde{\mathcal{B}}$ with the property that for some time-steps its conditional expected instantaneous regret is decreasing. For ease of presentation and instead of making use of odd and even time-steps in the definition of $\widetilde{\mathcal{B}}$ we assume each round t is split in two types of steps (Step 1 and Step 2). We will denote objects pertaining to the $t-$th round step i using a subscript t and a superscript (i). The construction of $\widetilde{\mathcal{B}}$ is simple. The smoothed algorithm maintains an internal copy of the original algorithm $\mathcal{B}$. During step 1 of round t, $\widetilde{\mathcal{B}}$ will play the action suggested by its internal copy of $\mathcal{B}$. During step 2 of round t, $\widetilde{\mathcal{B}}$ will instead sample uniformly from the set of policies previously played by the copy of $\mathcal{B}$ maintained by $\widetilde{\mathcal{B}}$ during steps of type 1 from all rounds from the start to t.

Let's define step 2 more formally. If algorithm $\mathcal{B}$ is at state s during round t, at step 2 of the corresponding time-step the smoothed strategy will pick an index q in $[1, 2, .., s]$ uniformly at random, and will then re-play the policy $\mathcal{B}$ used during step 1 of round q. Since $\mathcal{B}$ is (U, δ, T)-bounded we will show in Lemma 2.4.2 that the expected instantaneous regret of step 2 at round s is at most $U(s, \delta)/s$ with high probability.

Algorithm 8: Smoothed Algorithm

Input: Base Algorithm $\mathcal{B}$;
Let π_s be the policy of $\mathcal{B}$ in state s.
Let $\widetilde{\pi}_s^{(1)}, \widetilde{\pi}_s^{(2)}$ be the policies of $\widetilde{\mathcal{B}}$ in state s during step 1 and 2 respectively.
Initialize state counter $s = 1$.
for $t = 1, \cdots, T$ **do**

- **Step 1**
  - Receive action set $\mathcal{A}_t^{(1)} \sim D_S$
  - Let $\widetilde{\pi}_s^{(1)} = \pi_s$ from $\mathcal{B}_i$.
  - Play action $a_t^{(1)} \sim \widetilde{\pi}_s^{(1)}(\mathcal{A}_t^{(1)})$.
  - Receive feedback $r_t^{(1)} = f(\mathcal{A}_t^{(1)}, \delta_{a_t^{(1)}}) + \xi_t^{(1)}$
  - Calculate π_{s+1} of $\mathcal{B}$ using $r_t^{(1)}$.
- **Step 2**
  - Receive action set $\mathcal{A}_t^{(2)} \sim D_S$.
  - Sample $q \sim \text{Uniform}(0, \cdots, s)$; Let $\widetilde{\pi}_s^{(2)} = \pi_q$ from $\mathcal{B}$.
  - Play action $a_t^{(2)} \sim \widetilde{\pi}_s^{(2)}(\mathcal{A}_t^{(2)})$.
  - Receive feedback $r_t^{(2)} = f(\mathcal{A}_t^{(2)}, \delta_{a_t^{(2)}}) + \xi_t^{(2)}$.
- Update $\mathcal{B}$'s internal counter $s \leftarrow s + 1$

It is easy to see that if algorithm $\widetilde{\mathcal{B}}$ has been played ℓ times (including step 1 and 2 plays), the internal counter of $\mathcal{B}$ equals $\ell/2$. We will make use of this internal counter when we connect a smoothed algorithm with the Stochastic CORRAL master. We now introduce the definition of $(U, \delta, \mathcal{T}^{(2)})-$Smoothness which in short corresponds to algorithms that satisfy a high probability conditional expected regret upper bound during steps of type 2.

Definition 2.4.1 ($(U, \delta, \mathcal{T}^{(2)})-$Smoothness)**.** *Let $U : \mathbb{R} \times [0,1] \to \mathbb{R}^+$. We say a smoothed algorithm $\widetilde{\mathcal{B}}$ is $(U, \delta, \mathcal{T}^{(2)})-$smooth if with probability $1-\delta$ and for all rounds $t \in [T]$, the conditional expected instantaneous regret of Step 2 is bounded above by $U(t,\delta)/t$:*

$$\mathbb{E}_{\mathcal{A}';\sim\mathcal{D}_{\mathcal{S}},\pi_t^{(2)}=\pi_q \ s.t. \ q\sim\text{Uniform}(0,\cdots,s)}[f(\mathcal{A}',\pi^*) - f(\mathcal{A}',\pi_t^{(2)})|\widetilde{\mathcal{F}}_{t-1}] \leq \frac{U(t,\delta)}{t}, \ \forall t \in [T]. \tag{2.3}$$

Where $\widetilde{\mathcal{F}}_{t-1} = \sigma\left(\{\mathcal{A}_\ell^{(i)}, \widetilde{\pi}_\ell^{(i)}, r_\ell^{(i)}, a_\ell^{(i)}\}_{\ell\in[t-1],i\in\{1,2\}}, \cup\{\mathcal{A}_\ell^{(1)}, \widetilde{\pi}_\ell^{(1)}, r_\ell^{(i)}, a_\ell^{(1)}\}\right)$ is the sigma algebra generated by all contexts, rewards, policies and actions up to time t step 1.

During all steps of type 2 algorithm $\widetilde{\mathcal{B}}$ replays the policies it played as a result of encountering contexts $\mathcal{A}_1^{(1)}, ..., \mathcal{A}_s^{(1)}$. In Lemma 2.4.2 we will use the fact that all contexts are assumed to be generated as i.i.d. samples from the same context generating distribution $\mathcal{D}_S$ to show that $\widetilde{\mathcal{B}}$ is $(U, \delta, \mathcal{T}^{(2)})-$smooth.

With this objective in mind let's analyze a slightly more general setting. Let $\mathcal{B}$ be a $(U, \delta, T)-$bounded algorithm playing in an environment where the high probability regret upper bound U holds. Let's assume that $\mathcal{B}$ has been played for t time-steps during which it has encountered i.i.d. generated contexts $\mathcal{A}_1, \cdots, \mathcal{A}_t$ and has played actions sampled from policies $\pi_1, \cdots, \pi_t$. Similar to the definition of $\widetilde{\mathcal{F}}_{t-1}$ in Definition 2.4.1, let's define $\mathcal{F}_{t-1} = \sigma\left(\{\mathcal{A}_\ell, \widetilde{\pi}_\ell, r_\ell, a_\ell\}_{\ell\in[t-1]}\right)$, the sigma algebra generated by all contexts, rewards, policies and actions up to time $t-1$. We define the "expected replay regret" $\mathsf{Replay}(t|\mathcal{F}_{t-1})$ as:

$$\mathsf{Replay}(t|\mathcal{F}_{t-1}) = \mathbb{E}_{\mathcal{A}_1',\cdots,\mathcal{A}_t'}\left[\sum_{l=1}^t f(\mathcal{A}_l',\pi^*) - f(\mathcal{A}_l',\pi_l)\right] \tag{2.4}$$

Where $\mathcal{A}_1', \cdots, \mathcal{A}_t'$ are i.i.d. contexts from $\mathcal{D}_S$ all of them conditionally independent from $\mathcal{F}_t$. It is easy to see that the conditional instantaneous regret of a smoothed algorithm $\widetilde{\mathcal{B}}$ during round t step 2 equals the expected replay regret $\mathsf{Replay}(t|\widetilde{\mathcal{F}}_{t-1})$ of the $\mathcal{B}$ copy inside $\widetilde{\mathcal{B}}$.

As a first step in proving that $\widetilde{\mathcal{B}}$ is $(U, \delta, \mathcal{T}^{(2)})-$smooth in Lemma 2.4.2 we show the replay regret of a (U, δ, T)-bounded algorithm satisfies a high probability upper bound.

Lemma 2.4.2. *If $\mathcal{B}$ is $(U, \delta, T)-$bounded with $U(t,\delta) > 8\sqrt{t\log(\frac{t^2}{\delta})}$ and the rewards satisfy Assumption 2.2.1, then with probability at least $1-\delta$ for all $t \in [T]$ the expected replay regret of $\mathcal{B}$ satisfies:*

$$\mathsf{Replay}(t|\mathcal{F}_{t-1}) \leq 4U(t,\delta) + 2\delta t.$$

Furthermore, if $\delta \leq \frac{1}{\sqrt{T}}$ *then* $\mathsf{Replay}(t|\mathcal{F}_{t-1}) \leq 5U(t,\delta)$.

Proof. Let's condition on the event $\mathcal{E}_1$ that $\mathcal{B}$'s plays satisfy the high probability regret guarantee given by U:

$$\sum_{l=1}^{t} f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}_l, \pi_l) \leq U(t,\delta). \tag{2.5}$$

For all $t \in [T]$ and where $\mathcal{A}_1, \cdots, \mathcal{A}_t$ are the contexts algorithm $\mathcal{B}$ encountered up to time t. Since $\mathcal{B}$ is $(U,\delta,T)-$bounded it must be the case that $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$.

Let $\mathcal{A}'_1, \cdots, \mathcal{A}'_t$ be a collection of t fresh i.i.d. contexts from $\mathcal{D}_S$ independent from $\mathcal{F}_t$. We now use martingale concentration arguments to show that $\sum_{l=1}^t f(\mathcal{A}_l, \pi^*) \approx \sum_{l=1}^t f(\mathcal{A}'_l, \pi^*)$ and $\sum_{l=1}^t f(\mathcal{A}_l, \pi_l) \approx \sum_{l=1}^t f(\mathcal{A}'_l, \pi_l)$. Consider the following two martingale difference sequences:

$$\left\{M_l^1 := f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}'_l, \pi^*)\right\}_{l=1}^T$$
$$\left\{M_l^2 := f(\mathcal{A}'_l, \pi_l) - f(\mathcal{A}_l, \pi_l)\right\}_{l=1}^T$$

Since by assumption $\max_{\mathcal{A}',\pi} |f(\mathcal{A}',\pi)| \leq 1$ each term in $\{M_l^1\}$ and $\{M_l^2\}$ is bounded and satisfies $\max\left(|M_l^1|, |M_l^2|\right) \leq 2$ for all t. A simple use of Azuma-Hoeffding yields:

$$\mathbb{P}\left(\left| \sum_{l=1}^t M_l^i \right| \geq \sqrt{8t \log\left(\frac{8t^2}{\delta}\right)} \right) \leq 2\exp\left(-\frac{8t\log(\frac{8t^2}{\delta})}{8t} \right) = \frac{\delta}{4t^2}\ .$$

Summing over all t, and all $i \in \{1,2\}$, using the fact that $\sum_{t=1}^T \frac{1}{t^2} < 2$ and the union bound implies that for all t, with probability $1-\delta$,

$$\left| \sum_{l=1}^t f(\mathcal{A}_l, \pi_l) - f(\mathcal{A}'_l, \pi_l) \right| \leq \sqrt{8t\log\left(\frac{8t^2}{\delta}\right)} \tag{2.6}$$
$$\left| \sum_{l=1}^t f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}'_l, \pi^*) \right| \leq \sqrt{8t\log\left(\frac{8t^2}{\delta}\right)} \tag{2.7}$$
$$\tag{2.8}$$

Denote this event as $\mathcal{E}_2$. We shall proceed to upper bound the replay regret. Let's condition on $\mathcal{E}_1 \cap \mathcal{E}_2$. The following sequence of inequalities holds,

$$\sum_{l=1}^{t} f(\mathcal{A}'_l, \pi^*) - f(\mathcal{A}'_l, \pi_l) \stackrel{(i)}{\leq} \sum_{l=1}^{t} f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}_l, \pi_l) + \left| \sum_{l=1}^{t} f(\mathcal{A}_l, \pi_l) - f(\mathcal{A}'_l, \pi_l) \right| + \left| \sum_{l=1}^{t} f(\mathcal{A}_l, \pi^*) - f(\mathcal{A}'_l, \pi^*) \right|$$
$$\leq U(t, \delta) + 2\sqrt{8t \log\left(\frac{8t^2}{\delta}\right)}$$

For all $t \in [T]$. Inequality (i) follows by the triangle inequality while (ii) is a consequence of conditioning on $\mathcal{E}_1 \cap \mathcal{E}_2$ and invoking inequalities 2.5, 2.6 and 2.6. We conclude that with probability at least $1 - 2\delta$ and for all $t \in [T]$,

$$\sum_{l=1}^{t} f(\mathcal{A}'_l, \pi^*) - f(\mathcal{A}'_l, \pi_l) \leq U(t, \delta) + 2\sqrt{8t \log\left(\frac{8t^2}{\delta}\right)}$$

Since we have assumed that $U(t, \delta) > 8\sqrt{t \log(\frac{t^2}{\delta})}$, averaging out over the randomness in $\{\mathcal{A}'_l\}_{l=1}^{t}$ yields that conditioned on $\mathcal{E}_1$,

$$\mathsf{Replay}(t|\mathcal{F}_{t-1}) \leq 4(1 - 2\delta)U(t, \delta) + 2\delta t < 4U(t, \delta) + 2\delta t.$$

It is easy to see that in case $\delta \leq \frac{1}{\sqrt{T}}$ then $\mathsf{Replay}(t|\mathcal{F}_{t-1}) \leq 5U(t, \delta)$. □

In Propositon 2.4.3 we show that if $\mathcal{B}$ is bounded, then $\widetilde{\mathcal{B}}$ is both bounded and smooth. We will then show that several algorithms such as UCB, LinUCB, ϵ-greedy and EXP3 are (U, δ, T)-bounded for appropriate functions U. By Proposition 2.4.3 we will then conclude the smoothed versions of these algorithms are smooth.

Proposition 2.4.3. *If* $U(t, \delta) > 8\sqrt{t \log(\frac{t^2}{\delta})}$, $\delta \leq \frac{1}{\sqrt{T}}$*, the rewards satisfy Assumption 2.2.1 and* $\mathcal{B}$ *is* $(U, \delta, T)-$*bounded, then* $\widetilde{\mathcal{B}}$ *is* $(5U, \delta, \mathcal{T}^{(2)})-$*smooth and with probability at least* $1 - 3\delta$*,*

$$\sum_{l=1}^{t} \sum_{i \in \{1,2\}} f(\mathcal{A}_l^{(i)}, \pi^*) - f(\mathcal{A}_l^{(i)}, \pi_l^{(i)}) \leq 7U(t, \delta) \log(t).$$

for all $t \in [T]$.

Proof. Let $\mathcal{E}_1$ denote the event that $\widetilde{\mathcal{B}}$'s plays during steps of type 1 satisfy the high probability regret guarantee given by U:

$$\sum_{l=1}^{t} f(\mathcal{A}_l^{(1)}, \pi^*) - f(\mathcal{A}_l^{(1)}, \pi_l^{(1)}) \leq U(t, \delta). \tag{2.9}$$

for all $t \in [T]$. Since the conditional instantaneous regret of Step 2 of round t equals the average replay regret of the type 1 steps up to t, Lemma 2.4.2 implies that whenever $\mathcal{E}_2$ holds (see definition for $\mathcal{E}_2$ in the proof of Lemma 2.4.2) which occurs with probability at least $1-\delta$, the conditional expected instantaneous regret satisfies: $\mathbb{E}[f(\mathcal{A}', \pi^*) - f(\mathcal{A}', \pi_t^{(2)})|\widetilde{\mathcal{F}}_{t-1}] \leq \frac{5U(t,\delta)}{t}$ for all $t \in [T]$. This shows that $\widetilde{\mathcal{B}}$ is $(5U, \delta, \mathcal{T}^{(2)})-$smooth.

It is easy to see that if we condition on $\mathcal{E}_1 \cap \mathcal{E}_2$ the conditional expected instantaneous regret of steps of type 2 satisfy,

$$\sum_{l=1}^{t} \mathbb{E}[f(\mathcal{A}', \pi^*) - f(\mathcal{A}', \pi_l^{(2)})|\widetilde{\mathcal{F}}_{l-1}] \leq \sum_{l=1}^{t} \frac{5U(l,\delta)}{l} \leq 5U(t,\delta)\log(t) \tag{2.10}$$

For all $t \in [T]$. We now show the regret incurred by $\widetilde{\mathcal{B}}$ satisfies a high probability upper bound. To bound the regret accrued during time-steps of type 2, consider the following Martingale difference sequences,

$$\left\{M_l^1 := \mathbb{E}[f(\mathcal{A}', \pi_l^{(2)})|\widetilde{\mathcal{F}}_{l-1}] - f(\mathcal{A}_l^{(2)}, \pi_l^{(2)})\right\}_{l=1}^{T}$$
$$\left\{M_l^2 := \mathbb{E}[f(\mathcal{A}', \pi^*)|\widetilde{\mathcal{F}}_{l-1}] - f(\mathcal{A}_l^{(2)}, \pi^*)\right\}_{l=1}^{T}$$

As a result of Assumption 2.2.1, $|M_l^i| \leq 2$ for all $i \in \{1, 2\}$ and therefore a simple use of Azuma-Hoeffding's inequality ,

$$\mathbb{P}\left(\left|\sum_{l=1}^{t} M_l^i\right| \geq \sqrt{8t\log\left(\frac{8t^2}{\delta}\right)}\right) \leq 2\exp\left(-\frac{8t\log(\frac{8t^2}{\delta})}{8t}\right) = \frac{\delta}{4t^2}\ .$$

Summing over all t, applying the union bound, using the fact that $\sum_{t=1}^{T}\frac{1}{t^2} < 2$ implies that for all $t \in [T]$, with probability $1-\delta$,

$$\left|\sum_{l=1}^{t} \mathbb{E}[f(\mathcal{A}', \pi^*) - f(\mathcal{A}', \pi_l^{(2)})|\widetilde{\mathcal{F}}_{l-1}] - f(\mathcal{A}_l^{(2)}, \pi^*) - f(\mathcal{A}_l^{(2)}, \pi_l^{(2)})\right| \leq \sqrt{8t\log\left(\frac{8t^2}{\delta}\right)} \leq U(t,\delta) \tag{2.11}$$

Let's denote as $\mathcal{E}_3$ the event where Equation 2.11 holds. If $\mathcal{E}_2 \cap \mathcal{E}_3$ occur, then combining the upper bounds in 2.10 and 2.11 we conclude that,

$$\sum_{l=1}^{t} f(\mathcal{A}_l^{(2)}, \pi^*) - f(\mathcal{A}_l^{(2)}, \pi_l^{(2)}) \leq 6U(t, \delta) \log(t)$$

combining this last observation with Equation 2.9, we conclude that for all t with probability at least $1 - 3\delta$,

$$\sum_{l=1}^{t} \sum_{i \in \{1,2\}} f(\mathcal{A}_l^{(i)}, \pi^*) - f(\mathcal{A}_l^{(i)}, \pi_l^{(i)}) \leq 7U(t, \delta,) \log(t)$$

For all $t \in [T]$. The result follows. □

It remains to show how to adapt the feedback structure of the Stochastic CORRAL master to deal with the two step nature of smoothed algorithms. We reproduce the full pseudo-code of the Stochastic CORRAL master adapted to smoothed algorithms below,

Algorithm 9: Smooth Stochastic CORRAL Master Algorithm

Input: Smoothed Base Algorithms $\{\widetilde{\mathcal{B}}_j\}_{j=1}^M$, bias functions $\{b_j : \mathbb{N} \to \mathbb{R}\}_{j=1}^M$
for $t = 1, \cdots, T$ **do**
- Sample $j_t \sim p_t$.
- Play j_t for Steps 1 and 2.
- Receive feedback $r_t^{(1)}$ and $r_t^{(2)}$ from Steps 1 and 2 when executing $\widetilde{\mathcal{B}}_{j_t}$.
- Let s_{j_t} be the internal step counter of algorithm $\widetilde{\mathcal{B}}_{j_t}$ as defined in Algorithm 8.
- Update p_t using $2r_t^{(2)} - b_{j_t}(s_{j_t})$

For reasons that have to do with the analysis, Algorithm 9 has a few extra features not present in the master-base template of Algorithm 4. First, whenever the smooth stochastic corral master selects an algorithm j_t it plays it for two steps, thus coinciding with $\widetilde{\mathcal{B}}_{j_t}$'s two time step structure. Second, it updates its distribution p_t using the feedback $2r_t^{(2)} - b_{j_t}(s_{j_t})$ instead of using the sum $r_t^{(1)} + r_t^{(2)}$. Most notably, the update makes use of a bias adjustment to the reward signal that is not present in the original. The reason behind this modification will become clearer in the regret analysis.

Applications of Proposition 2.4.3

We now show the smoothed versions of several algorithms satisfy Definition 2.4.1 by showing they are $(U, \delta, T)-$bounded for an appropriate upper bound function U. We focus on algorithms for the $k-$armed bandit setting and the contextual linear bandit setting.

Lemma 2.4.4 (Theorem 3 in [1])**.** *In the case of changing and potentially infinite contexts of dimension d, LinUCB is (U, δ, T)-bounded with $U(t, \delta) = O(d\sqrt{t}\log(1/\delta))$.*

Lemma 2.4.5 (Theorem 1 in [19])**.** *In the case of finite linear contexts of size k and dimension d, LinUCB is (U, δ, T)-bounded with $U(t, \delta) = O(\sqrt{dt}\log^3(kT\log(T)/\delta))$.*

Lemma 2.4.6 (Theorem 1 in [58])**.** *In the $k-$armed adversarial bandit setting Exp3 is $(U, \delta, T)-$bounded where $U(t, \delta) = O(\sqrt{tk}\log\frac{tk}{\delta})$.*

Lemma 2.4.7. *In the stochastic $k-$armed bandit problem, if we assume the noise ξ_t is conditionally 1-sub-Gaussian, UCB is (U, δ, T)-bounded with $U(t, \delta) = O(\sqrt{tk}\log\frac{tk}{\delta})$.*

Proof. The regret of UCB is bounded as $\sum_{i:\Delta_i>0}\left(3\Delta_i + \frac{16}{\Delta_i}\log\frac{2k}{\Delta_i\delta}\right)$ (Theorem 7 of [1]) where Δ_i is the gap between arm i and the best arm. By substituting the worst-case Δ_i in the regret bound, $U(T, \delta) = O(\sqrt{Tk}\log\frac{Tk}{\delta})$. □

For the remainder of this section we focus on showing that in the stochastic $k-$armed bandit problem, the ϵ-greedy algorithm (Algorithm 1.2 of [60]) is $(U, T, \delta)-$bounded. At time t the ϵ-greedy algorithm selects with probability $\epsilon_t = \min(c/t, 1)$ an arm uniformly at random, and with probability $1 - \epsilon_t$ it selects the arm whose empirical estimate of the mean is largest so far. Let's introduce some notation: we will denote by $\mu_1, \cdots, \mu_k$ the unknown means of the K arms use the name $\widehat{\mu}_j^{(t)}$ to denote the empirical estimate of the mean of arm j after using t samples.

Without loss of generality let μ_1 be the optimal arm. We denote the sub-optimality gaps as $\Delta_j = \mu_1 - \mu_j$ for all $j \in [k]$. Let Δ_* be the smallest nonzero gap. We follow the discussion in [9] and start by showing that under the right assumptions, and for a horizon of size T, the algorithm satisfies a high probability regret bound for all $t \leq T$. The objective of this section is to prove the following Lemma:

Lemma 2.4.8. *If $c = \frac{10K\log(T^3/\gamma)}{\Delta_*^2}$[3] for some $\gamma \in (0,1)$ satisfying $\gamma \leq \frac{\Delta_j^2}{2}$, then $\epsilon-$greedy with $\epsilon_t = \frac{c}{t}$ is $(U, \delta, T)-$bounded for $\delta \leq \frac{\Delta_*^2}{T^3}$ where*

$$U(t, \delta) = \frac{30k\log(\frac{1}{\delta})}{\Delta_*^2}\left(\sum_{j=2}^{k}\frac{\Delta_j}{\Delta_*^2} + \Delta_j\right)\log(t+1).$$

[3] This choice of c is robust to multiplication by a constant.

Proof. Let $E(t) = \frac{1}{2k}\sum_{l=1}^{t}\epsilon_l$ and denote by $T_j(t)$ the random variable denoting the number of times arm j was selected up to time t. We start by analyzing the probability that a suboptimal arm $j > 1$ is selected at time t:

$$\mathbb{P}(j \text{ is selected at time } t) \leq \frac{\epsilon_t}{k} + \left(1 - \frac{\epsilon_t}{k}\right)\mathbb{P}\left(\widehat{\mu}_j^{(T_j(t))} \geq \widehat{\mu}_1^{(T_1(t))}\right) \tag{2.12}$$

Let's bound the second term.

$$\mathbb{P}\left(\widehat{\mu}_j^{(T_j(t))} \geq \widehat{\mu}_1^{(T_1(t))}\right) \leq \mathbb{P}\left(\widehat{\mu}_j^{(T_j(t))} \geq \mu_j + \frac{\Delta_j}{2}\right) + \mathbb{P}\left(\widehat{\mu}_1^{(T_1(t))} \leq \mu_1 - \frac{\Delta_j}{2}\right)$$

The analysis of these two terms is the same. Denote by $T_j^R(t)$ the number of times arm j was played as a result of a random epsilon greedy move. We have:

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_j^{(T_j(t))} \geq \mu_j + \frac{\Delta_j}{2}\right) &= \sum_{l=1}^{t}\mathbb{P}\left(T_j(t) = l \text{ and } \widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right) \\
&= \sum_{l=1}^{t}\mathbb{P}\left(T_j(t) = l|\widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right)\mathbb{P}\left(\widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right) \\
&\overset{a}{\leq} \sum_{l=1}^{t}\mathbb{P}\left(T_j(t) = l\middle|\widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right)\exp(-\Delta_j^2 t/2) \\
&\overset{b}{\leq} \sum_{l=1}^{\lfloor E(t)\rfloor}\mathbb{P}\left(T_j(t) = l\middle|\widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right) + \frac{2}{\Delta_j^2}\exp(-\Delta_j^2\lfloor E(t)\rfloor/2) \\
&\leq \sum_{l=1}^{\lfloor E(t)\rfloor}\mathbb{P}\left(T_j^R(t) = l\middle|\widehat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2}\right) + \frac{2}{\Delta_j^2}\exp(-\Delta_j^2\lfloor E(t)\rfloor/2) \\
&\leq \underbrace{\lfloor E(t)\rfloor\mathbb{P}\left(T_j(t)^R \leq \lfloor E(t)\rfloor\right)}_{(1)} + \underbrace{\frac{2}{\Delta_j^2}\exp(-\Delta_j^2\lfloor E(t)\rfloor/2)}_{(2)}
\end{aligned}$$

Inequality a is a consequence of the Azuma-Hoeffding inequality bound. Inequality b follows because $\sum_{l=E+1}^{\infty}\exp(-\alpha l) \leq \frac{1}{a}\exp(-\alpha E)$. Term (1) corresponds to the probability that within the interval $[1, \cdots, t]$, the number of greedy pulls to arm j is at most half its expectation. Term (2) is already "small". Lets proceed to bound (1). Let $\epsilon_t = \min(c/t, 1)$. with $c = \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ for some $\gamma \in (0, 1)$ satisfying $\gamma \leq \Delta_j^2$. We'll show that under these assumptions we can lower bound $E(t)$. If $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$:

$$\begin{aligned}
E(t) := \frac{1}{2k}\sum_{l=1}^{t}\epsilon_l &= \frac{5\log(T^3/\gamma)}{\Delta_*^2} + \frac{5\log(T^3/\delta)}{\Delta_*^2}\sum_{l=\log(T^3/\gamma)}^{t}\frac{1}{l} \\
&\geq \frac{5\log(T^3/\gamma)}{\Delta_*^2} + \frac{5\log(T^3/\gamma)\log(t)}{2\Delta_*^2} \\
&\geq \frac{5\log(T^3/\gamma)}{\Delta_*^2}
\end{aligned}$$

By Bernstein's inequality (see derivation of equation (13) in [9]) we can upper bound $T_j^R(t)$:

$$\mathbb{P}\left(T_j^R(t) \leq E(t)\right) \leq \exp\left(-E(t)/5\right) \tag{2.13}$$

Hence for $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$:

$$\mathbb{P}\left(T_j^R(t) \leq E(t)\right) \leq \left(\frac{\gamma}{T^3}\right)^{\frac{1}{\Delta_*^2}}$$

And therefore since $E(t) \leq T$ and $\frac{1}{\Delta_*} \geq 1$ we can upper bound (1) as:

$$\lfloor E(t)\rfloor \mathbb{P}\left(T_j^R(t) \leq \lfloor E(t)\rfloor\right) \leq \left(\frac{\gamma}{T^2}\right)^{\frac{1}{\Delta_*^2}} \leq \frac{\gamma}{T^2}$$

Now we proceed with term (2):

$$\begin{aligned}
\frac{2}{\Delta_j^2}\exp\left(-\Delta_j^2\lfloor E(t)\rfloor/2\right) &\overset{(a)}{\leq} \frac{2}{\Delta_j^2}\exp\left(-\frac{5}{2}\log\left(\frac{T^3}{\gamma}\right)\frac{\Delta_j^2}{\Delta_*^2}\right) \\
&\leq \frac{2}{\Delta_j^2}\exp\left(-\log\left(\frac{T^3}{\gamma}\right)\right) \\
&= \frac{2}{\Delta_j^2}\left(\frac{\gamma}{T^3}\right)^5 \\
&\overset{(b)}{\leq} \frac{\gamma}{T^3}
\end{aligned}$$

The first inequality (a) follows because $E(t) \geq \frac{5\log(T^3/\gamma)}{\Delta_*^2}$. Inequality (b) follows because by the assumption $\gamma \leq \frac{\Delta_j^2}{2}$ the last term is upper bounded by $\frac{\gamma}{T^3}$.

By applying the union bound over all arms $j \neq 1$ and time-steps $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$, we conclude that the probability of choosing a sub-optimal arm $j \geq 2$ at any time

time t for $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ as a **greedy choice** is upper bounded by $\frac{k\gamma}{T^2} \leq \frac{k\gamma}{T}$. In other words after $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ rounds, with probability $1 - \frac{k\gamma}{T}$ sub-optimal arms are only chosen as a result of random epsilon greedy move (occurring with probability ϵ_t).

A similar argument as the one that gave us Equation 2.13 can be used to upper bound the probability that $T_j^R(t)$ be much larger than its mean:

$$\mathbf{P}\left(T_j^R(t) \geq 3E(j)\right) \leq \exp(-E(t)/5)$$

Using this and the union bound we see that with probability more than $1 - \frac{k\gamma}{T}$ and for all $t \in [T]$ and arms $j \in [k]$, $T_j^R(t) \leq 3E(t)$. Combining this with the observation that after $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ and with probability $1 - \frac{k\gamma}{T}$ over all t simultaneously regret is only incurred by random exploration pulls (and not greedy actions), we can conclude that with probability at least $1 - \frac{2k\gamma}{T}$ simultaneously for all $t \geq \frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ the regret incurred is upper bounded by:

$$\underbrace{\frac{10k\log(T^3/\gamma)}{\Delta_*^2} \cdot \frac{1}{k}\sum_{j=2}^{k}\Delta_j}_{(i)} + \underbrace{3E(t)\sum_{j=2}^{k}\Delta_j}_{(ii)}$$

Term (i) is a crude upper bound on the regret incurred in the first $\frac{10k\log(T^3/\gamma)}{\Delta_*^2}$ rounds and (ii) is an upper bound for the regret incurred in the subsequent rounds.

Since $E(t) \leq \frac{20k\log(T^3/\gamma)}{\Delta_*^2}\log(t)$ we conclude that with probability $1 - \frac{2k\gamma}{T}$ for all $t \leq T$ the cumulative regret of epsilon greedy is upper bounded by

$$30K\log(T^3/\gamma)\left(\sum_{j=2}^{k}\frac{\Delta_j}{\Delta_*^2} + \Delta_j\right)\max(\log(t), 1),$$

the result follows by identifying $\delta = \gamma/T^3$.

□

Lemma 2.4.8 gives us an instance dependent upper bound for the $\epsilon-$greedy algorithm. We now show the instance-independent high probability regret bound for ϵ-greedy:

Lemma 2.4.9. *If $c = \frac{10k\log(\frac{1}{\delta})}{\Delta_*^2}$, then $\epsilon-$greedy with $\epsilon_t = \frac{c}{t}$ is $(\delta, U, T)-$bounded for $\delta \leq \frac{\Delta_*^2}{T^3}$ and:*

1. *$U(t,\delta) = 16\sqrt{\log(\frac{1}{\delta})t}$ when $k = 2$.*

2. $U(t,\delta) = 20\left(k\log(\frac{1}{\delta})\left(\sum_{j=2}^{K}\Delta_j\right)\right)^{1/3} t^{2/3}$ *when* $k > 2$.

Proof. Let Δ be some arbitrary gap value. Let $R(t)$ denote the expected regret up to round t. We recycle the notation from the proof of Lemma 2.4.8, recall $\delta = \gamma/T^3$.

$$\begin{aligned}
R(t) &= \sum_{\Delta_j \leq \Delta} \Delta_j \mathbb{E}\left[T_j(t)\right] + \sum_{\Delta_j \geq \Delta} \Delta_j \mathbb{E}\left[T_j(t)\right] \\
&\leq \Delta t + \sum_{\Delta_j \geq \Delta} \Delta_j \mathbb{E}\left[T_j(t)\right] \\
&\leq \Delta t + 30k\log(T^3/\gamma)\left(\sum_{\Delta_j \geq \Delta}^{k} \frac{\Delta_j}{\Delta_*^2} + \Delta_j\right)\log(t) \\
&\leq \Delta t + 30k\log(T^3/\gamma)\left(\sum_{\Delta_j \geq \Delta}^{k} \frac{\Delta_j}{\Delta_*^2}\right) + 30k\log(T^3/\gamma)\log(t)\left(\sum_{\Delta_j \geq \Delta}^{k} \Delta_j\right) \qquad (2.14)
\end{aligned}$$

When $k = 2$, $\Delta_2 = \Delta_*$ and therefore (assuming $\Delta < \Delta_2$):

$$\begin{aligned}
R(t) &\leq \Delta t + \frac{30k\log(T^3/\gamma)}{\Delta_2} + 30k\log(T^3/\gamma)\log(t)\Delta_2 \\
&\leq \Delta t + \frac{30k\log(T^3/\gamma)}{\Delta} + 30k\log(T^3/\gamma)\log(t)\Delta_2 \\
&\overset{\text{A}}{\leq} \sqrt{30k\log(T^3/\gamma)t} + 30k\log(T^3/\gamma)\log(t)\Delta_2 \\
&\overset{\text{B}}{\leq} 8\sqrt{k\log(T^3/\gamma)t} \\
&\leq 16\sqrt{\log(T^3/\gamma)t}
\end{aligned}$$

Inequality A follows from setting Δ to the optimizer, which equals $\Delta = \sqrt{\frac{30k\log(T^3/\gamma)}{t}}$. The second inequality B is satisfied for T large enough. We choose this expression for simplicity of exposition.

When $k > 2$ notice that we can arrive to a bound similar to 2.14:

$$R(t) \leq \Delta t + 30k\log(T^3/\gamma)\left(\sum_{\Delta_j \geq \Delta}^{k} \frac{\Delta_j}{\Delta^2}\right) + 30k\log(T^3/\gamma)\log(t)\left(\sum_{\Delta_j \geq \Delta}^{k} \Delta_j\right)$$

Where Δ_* is substituted by Δ. This can be obtained from Lemma 2.4.8 by simply substituting Δ_* with Δ in the argument for arms $j : \Delta_j \geq \Delta$.

We upper bound $\sum_{\Delta_j \geq \Delta} \Delta_j$ by $\sum_{j=2}^{k} \Delta_j$. Setting Δ to the optimizer of the expression yields $\Delta = \left(\frac{30k\log(T^3/\gamma)\left(\sum_{j=2}^{k}\Delta_j\right)}{t}\right)^{1/3}$, and plugging this back into the equation we obtain:

$$\begin{aligned} R(t) &\leq 2\left(30k\log(T^3/\gamma)\left(\sum_{j=2}^{k}\Delta_j\right)\right)^{1/3} t^{2/3} + 30k\log(T^3/\gamma)\log(t)\left(\sum_{j=2}^{k}\Delta_j\right) \\ &\overset{\xi}{\leq} 20\left(k\log(T^3/\gamma)\left(\sum_{j=2}^{k}\Delta_j\right)\right)^{1/3} t^{2/3} \end{aligned}$$

The inequality ξ is true for T large enough. We choose this expression for simplicity of exposition.

□

Regret Analysis

In this section we go back to sketch the proof of Theorem 2.3.2 by explaining how to bound terms I and II in the regret decomposition of Equation 2.2.

Bounding Term I. Recall that Algorithm 9 only sends the smoothed reward of Step 2 to the master while the base plays and incurs regrets from both Step 1 and Step 2. We show in Section 2.7 that this does not affect the regret of the master significantly. For CORRAL with learning rate η, $\mathbb{E}\left[\mathrm{I}\right] \leq O\left(\sqrt{MT} + \frac{M\ln T}{\eta} + T\eta\right) - \frac{\mathbb{E}\left[\frac{1}{\underline{p}_{i_\star}}\right]}{40\eta\ln T}$. For EXP3.P with exploration rate p, $\mathbb{E}\left[\mathrm{I}\right] < O(\sqrt{MT} + \frac{1}{p} + MTp)$.

Bounding Term II. This quantity represents the regret of the base $i_\star$ when it only updates its state when selected. We assume smoothed base algorithm $\widetilde{\mathcal{B}}_{i_\star}$ satisfies the smoothness and boundedness in Definitions 2.2.1 and 2.4.1. For the purpose of the analysis we declare that when a smoothed base repeats its policy while not played, it repeats its subsequent Step 2 policy (Algorithm 8). This will become clearer in Section 2.7. Since we select $\widetilde{\mathcal{B}}_{i_\star}$ with probability $\underline{p}_{i_\star}$ it will be updated every $1/\underline{p}_i$ time-steps and the regret upper bound will be roughly $\frac{1}{\underline{p}_{i_\star}} U_{i_\star}(T\underline{p}_{i_\star}, \delta)$.

EXP3.P	CORRAL
$\widetilde{O}\left(\sqrt{MT} + MTp + T^{\alpha}p^{\alpha-1}c(\delta)\right)$	$\widetilde{O}\left(\sqrt{MT} + \frac{M}{\eta} + T\eta + T\,c(\delta)^{\frac{1}{\alpha}}\eta^{\frac{1-\alpha}{\alpha}}\right)$
$\widetilde{O}\left(\sqrt{MT} + M^{\frac{1-\alpha}{2-\alpha}}T^{\frac{1}{2-\alpha}}c(\delta)^{\frac{1}{2-\alpha}}\right)$	$\widetilde{O}\left(\sqrt{MT} + M^{\alpha}T^{1-\alpha} + M^{1-\alpha}T^{\alpha}c(\delta)\right)$
$\widetilde{O}\left(\sqrt{MT} + M^{\frac{1-\alpha}{2-\alpha}}T^{\frac{1}{2-\alpha}}c(\delta)\right)$	$\widetilde{O}\left(\sqrt{MT} + M^{\alpha}T^{1-\alpha} + M^{1-\alpha}T^{\alpha}c(\delta)^{\frac{1}{\alpha}}\right)$

Table 2.1: Comparison of model selection guarantees for Stochastic CORRAL between the EXP3.P and CORRAL master. The top row shows the general regret guarantees. The middle row shows the regret guarantees when α and $c(\delta)$ are known. The bottom row shows the regret guarantees when α is known and $c(\delta)$ is unknown.

Theorem 2.4.10. *We have that* $\mathbb{E}\left[\mathrm{II}\right] \leq \mathcal{O}\left(\mathbb{E}\left[\frac{1}{\underline{p}_i} U_i(T\underline{p}_i, \delta)\log T\right] + \delta T(\log T + 1)\right)$. *Here, the expectation is over the random variable* $\underline{p}_i$. *If* $U(t,\delta) = t^{\alpha}c(\delta)$ *for some* $\alpha \in [1/2, 1)$ *then,* $\mathbb{E}\left[\mathrm{II}\right] \leq \widetilde{\mathcal{O}}\left(T^{\alpha}c(\delta)\mathbb{E}\left[\frac{1}{\underline{p}_i^{1-\alpha}}\right] + \delta T(\log T + 1)\right)$.

Total Regret. Adding Term I and Term II gives us the following worst-case model selection regret bound for the CORRAL master (maximized over $\underline{p}_{i_\star}$ and with a chosen η) and the EXP3.P master (with a chosen p):

Theorem 2.4.11. *If a base algorithm is* (U, δ, T)*-bounded for* $U(T,\delta) = T^{\alpha}c(\delta)$ *and some* $\alpha \in [1/2, 1)$ *and the choice of* $\delta = 1/T$, *the regret of the Smooth Stochastic CORRAL (Algorithm 9) where* $b_j(s) = \frac{U_j(s,\delta)}{s}$ *is upper bounded by:*

2.5 Lower Bound

In stochastic environments, algorithms such as UCB can achieve logarithmic regret bounds. Our model selection procedure however has a $O(\sqrt{T})$ overall regret. In this section, we show that in general it is impossible to obtain a regret better than $\Omega(\sqrt{T})$ even when the optimal base algorithm has 0 regret. In order to formalize this statement, let's define a model selection problem formally.

Definition 2.5.1 (Model Selection Problem)**.** *We call a tuple* $(\{\mathcal{B}_i\}_{i=1}^{M}, \mathrm{Env})$ *a model selection problem where* $\{\mathcal{B}_i\}_{i=1}^{M}$ *is a set of* M *base algorithms and* Env *is a bandit environment*[4].

[4]For example if $M = 2$ $(\{\mathrm{UCB}, \mathrm{LinUCB}\}, \mathrm{MAB})$ is a valid Model Selection Problem

Theorem 2.5.2. *Let $T \in \mathbb{N}$. For any model selection algorithm there exists a corresponding model selection problem* $(\{\mathcal{B}_1, \mathcal{B}_2\}, \text{Env})$ *such the regret of this model selection algorithm is lower bounded by* $R(T) = \Omega\left(\frac{\sqrt{T}}{\log(T)}\right)$.

Proof. Consider a stochastic 2-arm bandit problem where the best arm has expected reward 1/2 and the second best arm has expected reward 1/4. We construct base algorithms $\mathcal{B}_1, \mathcal{B}_2$ as follows. $\mathcal{B}_1$ always chooses the optimal arm and its expected instantaneous reward is 1/2. $\mathcal{B}_2$ chooses the second best arm at time step t with probability $\frac{4c}{\sqrt{t+2}\log(t+2)}$ (c will be specified later), and chooses the best arm otherwise. The expected reward at time step t of $\mathcal{B}_2$ is $\frac{1}{2} - \frac{c}{\sqrt{t+2}\log(t+2)}$.

Let A^* be uniformly sampled from $\{1, 2\}$. Consider two environments ν_1 and ν_2 for the master, each made up of two base algorithms $\widetilde{\mathcal{B}}_1, \widetilde{\mathcal{B}}_2$. Under ν_1, $\widetilde{\mathcal{B}}_1$ and $\widetilde{\mathcal{B}}_2$ are both instantiations of $\mathcal{B}_1$. Under ν_2, $\widetilde{\mathcal{B}}_{A^*}$, where A^* is a uniformly sampled index in $\{1, 2\}$, is a copy of $\mathcal{B}_1$ and $\widetilde{\mathcal{B}}_{3-A^*}$ is a copy of $\mathcal{B}_2$.

Let $\mathbb{P}_1, \mathbb{P}_2$ denote the probability measures induced by interaction of the master with ν_1 and ν_2 respectively. Let $\widetilde{\mathcal{B}}_{A_t}$ denote the base algorithm chosen by the master at time t. We have $\mathbb{P}_1(A_t \neq A^*) = \frac{1}{2}$ for all t, since the learner has no information available to identify which algorithm is considered optimal. By Pinskers' inequality we have

$$\mathbb{P}_2(A_t \neq A^*) \geq \mathbb{P}_1(A_t \neq A^*) - \sqrt{\frac{1}{2}\text{KL}(\mathbb{P}_1||\mathbb{P}_2)}$$

By the divergence decomposition [see 41, proof of Lemma 15.1 for the decomposition technique] and using that for $\Delta < \frac{1}{4}$: $\text{KL}(\frac{1}{2}, \frac{1}{2} - \Delta) \leq 3\Delta^2$ (Lemma 2.8.3), we have

$$\begin{aligned}\text{KL}(\mathbb{P}_1||\mathbb{P}_2) &= \sum_{t=2}^{\infty} \frac{1}{2}\text{KL}\left(\frac{1}{2}, \frac{1}{2} - \frac{c}{\sqrt{t+1}\log(t+1)}\right) \\ &\leq \sum_{t=2}^{\infty} \frac{3c^2}{2t\log(t)^2} \leq 3c^2 \,.\end{aligned}$$

Picking $c = \sqrt{\frac{1}{24}}$ leads to $\mathbb{P}_2(A_t \neq A^*) \geq \frac{1}{4}$, and the regret in environment ν_2 is lower bounded by

$$\begin{aligned}R(T) &\geq \sum_{t=1}^{T} \mathbb{P}_2(A_t \neq A^*) \frac{c}{\sqrt{t+1}\log(t+1)} \\ &\geq \frac{c}{4\log(T+1)} \sum_{t=1}^{T} \frac{1}{\sqrt{t+1}} = \Omega(\frac{\sqrt{T}}{\log(T)}) \,.\end{aligned}$$

□

CORRAL needs knowledge of the best base's regret to achieve the same regret. The following lower bound shows that this requirement is unavoidable:

Theorem 2.5.3. *Let* Alg *be a model selection algorithm. There exists a model selection problem with two base algorithms where the best base has regret* $\widetilde{O}(T^x)$ *for some* $0 < x < 1$ *such that if* Alg *has no knowledge of* x *nor of the reward of the best arm, then there exists a potentially different model selection problem where the best base also has regret* $\widetilde{O}(T^x)$ *but the model selection regret guarantee of* Alg *is lower bounded by* $\Omega(T^y)$ *with* $y > x$*.*

Proof. Let the set of arms be $\{a_1, a_2, a_3\}$. Let x and y be such that $0 < x < y \leq 1$. Let $\Delta = T^{x-1+(y-x)/2}$. Define two environment $\mathcal{E}_1$ and $\mathcal{E}_2$ with reward vectors $\{1, 1, 0\}$ and $\{1 + \Delta, 1, 0\}$ for $\{a_1, a_2, a_3\}$, respectively. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two base algorithms defined by the following fixed policies when running alone in $\mathcal{E}_1$ or $\mathcal{E}_2$:

$$\pi_1 = \begin{cases} a_2 & \text{w.p. } 1 - T^{x-1} \\ a_3 & \text{w.p. } T^{x-1} \end{cases}, \qquad \pi_2 = \begin{cases} a_2 & \text{w.p. } 1 - T^{y-1} \\ a_3 & \text{w.p. } T^{y-1} \end{cases}.$$

We also construct base $\mathcal{B}_2'$ defined as follows. Let $c_2 > 0$ and $\epsilon_2 = (y - x)/4$ be two constants. Base $\mathcal{B}_2'$ mimics base $\mathcal{B}_2$ when $t \leq c_2 T^{x-y+1+\epsilon_2}$, and picks arm a_1 when $t > c_2 T^{x-y+1+\epsilon_2}$. The instantaneous rewards of B_1 and B_2 when running alone are $r_t^1 = 1 - T^{x-1}$ and $r_t^2 = 1 - T^{y-1}$ for all $1 \leq t \leq T$. Next, consider model selection with base algorithms $\mathcal{B}_1$ and $\mathcal{B}_2$ in $\mathcal{E}_1$. Let T_1 and T_2 be the number of rounds that $\mathcal{B}_1$ and $\mathcal{B}_2$ are chosen, respectively.

First, assume case (1): There exist constants $c > 0$, $\epsilon > 0$, $p \in (0, 1)$, and $T_0 > 0$ such that with probability at least p, $T_2 \geq c T^{x-y+1+\epsilon}$ for all $T > T_0$.

The regret of base $\mathcal{B}_1$ when running alone for T rounds is $T \cdot T^{x-1} = T^x$. The regret of the model selection method is at least

$$p \cdot T_2 \cdot T^{y-1} \geq p \cdot c T^{x-y+1+\epsilon} \cdot T^{y-1} = p \cdot c \cdot T^{x+\epsilon}.$$

Given that the inequality holds for any $T > T_0$, it proves the statement of the lemma in case (1).

Next, we assume the complement of case (1): For all constants $c > 0$, $\epsilon > 0$, $p \in (0, 1)$, and $T_0 > 0$, with probability at least p, $T_2 < c T^{x-y+1+\epsilon}$ for some $T > T_0$.

Let T be any such time horizon. Consider model selection with base algorithms $\mathcal{B}_1$ and $\mathcal{B}_2'$ in environment $\mathcal{E}_2$ for T rounds. Let T_1' and T_2' be the number of rounds that $\mathcal{B}_1$ and $\mathcal{B}_2'$ are chosen. Note that $\mathcal{B}_2$ and $\mathcal{B}_2'$ behave the same for $c_2 T^{x-y+1+\epsilon}$ time steps, and that $\mathcal{B}_1$ and $\mathcal{B}_2$ never choose action a_1. Therefore for the first $c_2 T^{x-y+1+\epsilon_2}$ time steps, the model selection strategy that selects between $\mathcal{B}_1$ and $\mathcal{B}'_2$ in $\mathcal{E}_2$ behaves the same as when it runs $\mathcal{B}_1$ and $\mathcal{B}_2$ in $\mathcal{E}_1$. Therefore with probability $p > 1/2$, $T'_2 < c_2 T^{x-y+1+\epsilon_2}$, which implies $T'_1 > T/2$.

In environment $\mathcal{E}_2$, the regret of base B'_2 when running alone for T rounds is bounded as

$$(\Delta + T^{y-1}) c_2 T^{x-y+1+\frac{y-x}{4}} = c_2 T^{\frac{5x-y}{4}} + c_2 T^{\frac{3x+y}{4}} < 2c_2 T^{\frac{3x+y}{4}}$$

Given that with probability $p > 1/2$, $T'_1 > T/2$, the regret of the learner is lower bounded as,

$$p(\Delta + T^{x-1}) \cdot \frac{T}{2} > \frac{1}{2}(T^{x-1+\frac{y-x}{2}} + T^{x-1}) \cdot \frac{T}{2} < \frac{1}{2} T^{\frac{x+y}{2}},$$

which is larger than the regret of $\mathcal{B}'_2$ running alone because $\frac{3x+y}{4} < \frac{x+y}{2}$. The statement of the lemma follows given that for any T_0 there exists $T > T_0$ so that the model selection fails. □

2.6 Applications of Stochastic CORRAL

Misspecified Contextual Linear Bandit

We consider model selection in the misspecified linear bandit problem. The learner selects an action $a_t \in \mathcal{A}_t$ and receives a reward r_t such that $|\mathbb{E}[r_t] - a_t^\top \theta| \le \epsilon_*$ where $\theta \in \mathbb{R}^d$ is an unknown parameter vector and ϵ_* is the misspecification error. For this problem, [69] and [42] present variants of LinUCB that achieve a high probability $\widetilde{O}(d\sqrt{T} + \epsilon_* \sqrt{d}T)$ regret bound. Both algorithms require knowledge of ϵ_*, but [42] show a regret bound of the same order without the knowledge of ϵ_* for the version of the problem with a fixed action set $\mathcal{A}_t = \mathcal{A}$. Their method relies on G-optimal design, which does not work for contextual settings. It is an open question whether it is possible to achieve the above regret without knowing ϵ_* for problems with changing action sets.

In this section, we show a $\widetilde{O}(d\sqrt{T} + \epsilon_* \sqrt{d}T)$ regret bound for linear bandit problems with changing action sets without knowing ϵ_*. For problems with fixed action sets, we show an improved regret that matches the lower bound of [41].

Given a constant E so that $|\epsilon_*| \le E$, we divide the interval $[1, E]$ into an exponential grid $\mathcal{G} = [1, 2, 2^2, ..., 2^{\log(E)}]$. We use $\log(E)$ modified LinUCB bases, from either [69] or [42], with each base algorithm instantiated with a value of ϵ in the grid.

Theorem 2.6.1. *For the misspecified linear bandit problem described above, the regret of Stochastic CORRAL with a CORRAL master using learning rate* $\eta = \frac{1}{\sqrt{T}d}$ *and LinUCB base algorithms with target misspecification level* $\epsilon \in \mathcal{G}$*, is upper bounded by* $\widetilde{\mathcal{O}}(d\sqrt{T} + \epsilon_* \sqrt{d}T)$*. In the case of a fixed action linear bandit problem with* k *arms and* $\sqrt{k} > d$*, the regret of Stochastic CORRAL with a CORRAL master using learning rate* $\eta = \frac{1}{\sqrt{T}d}$ *applied to a set of base algorithms consisting of one UCB base and one G-optimal base algorithm [42] is upper bounded by* $\widetilde{\mathcal{O}}\left(\min\left(\frac{k}{d}\sqrt{T}, d\sqrt{T} + \epsilon_*\sqrt{d}T\right)\right)$.

Proof. From Lemma 2.4.7, for UCB, $U(T,\delta) = O(\sqrt{Tk}\log\frac{Tk}{\delta})$. Therefore from Theorem 2.4.11, running CORRAL with smooth UCB results in the following regret bound:

$$\widetilde{O}\left(\sqrt{MT} + \frac{M\ln T}{\eta} + T\eta + T\left(\sqrt{k}\log\frac{Tk}{\delta}\right)^2\eta\right) + \delta T.$$

If we choose $\delta = 1/T$ and hide some log factors, we get $\widetilde{O}\left(\sqrt{T} + \frac{1}{\eta} + Tk\eta\right)$.

For the LinUCB bases in [42] or [69] or the G-optimal algorithm [42], $U(t,\delta) = O(d\sqrt{t}\log(1/\delta) + \epsilon\sqrt{d}t)$. Substituting $\delta = 1/T$ into Theorem 2.4.11 implies:

$$\begin{aligned}
R(T) &\leq O\left(\sqrt{MT\log(\frac{4TM}{\delta})} + \frac{M\ln T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho}{40\eta\ln T} - 2\rho\, U(T/\rho,\delta)\log T\right] + \delta T \\
&\leq \widetilde{O}\left(\sqrt{MT} + \frac{M\ln T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho}{40\eta\ln T} - 2\rho\,(d\sqrt{\frac{T}{\rho}}\log(1/\delta) + \epsilon\sqrt{d}\frac{T}{\rho})\log T\right] \\
&\leq \widetilde{O}\left(\sqrt{MT} + \frac{M\ln T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho}{40\eta\ln T} - 2d\sqrt{T\rho}\log(1/\delta)\log T\right] + 2\epsilon\,\sqrt{d}T\,\log T
\end{aligned}$$

Maximizing over ρ results in a regret guarantee of the form $\widetilde{\mathcal{O}}\left(\sqrt{T} + \frac{1}{\eta} + Td^2\eta + \epsilon\sqrt{d}T\right)$. For the misspecified linear bandit problem we use $M = \mathcal{O}(\log(T))$ LinUCB bases with ϵ defined in the grid, and choose $\eta = \frac{1}{\sqrt{T}d}$. The resulting regret for Stochastic CORRAL is of the form $\widetilde{\mathcal{O}}\left(\sqrt{T}d + \epsilon\sqrt{d}T\right)$.

When the action sets are fixed, by the choice of $\eta = \frac{1}{\sqrt{T}d}$, the regret of Stochastic CORRAL with a CORRAL master over one UCB and one G-optimal base equals:

$$\widetilde{\mathcal{O}}\left(\min\left\{\sqrt{T}\left(d + \frac{k}{d}\right), \sqrt{T}d + \epsilon\sqrt{d}T\right\}\right).$$

If $\sqrt{k} > d$, the above expression becomes $\widetilde{\mathcal{O}}\left(\min\left(\sqrt{T}\frac{k}{d}, \sqrt{T}d + \epsilon\sqrt{d}T\right)\right)$ □

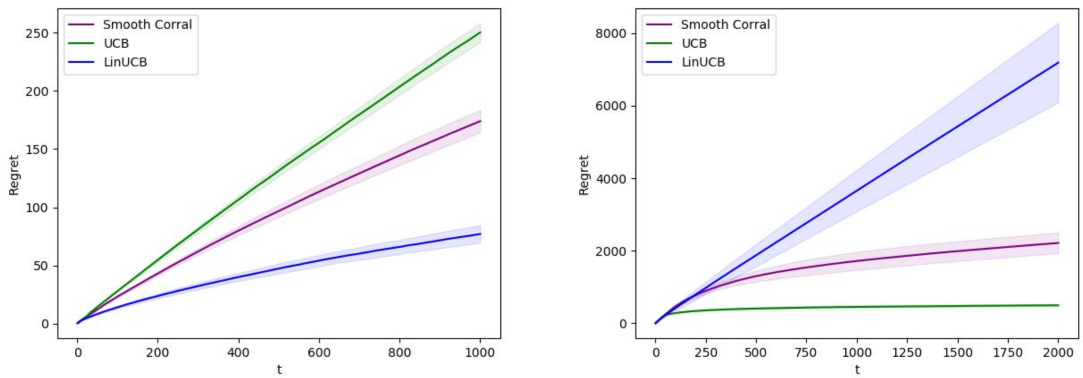
Observe that in the case of a fixed action linear bandit problem, the regret upper bound we achieve for Stochastic CORRAL with a CORRAL master and a learning rate of $\eta = \frac{1}{\sqrt{Td}}$ is of the form $\widetilde{\mathcal{O}}\left(\min\left(\frac{k}{d}\sqrt{T}, d\sqrt{T} + \epsilon_*\sqrt{d}T\right)\right)$. The product of the terms inside the minimum is of order $\widetilde{\mathcal{O}}(kT)$. This result matches the following lower bound that shows that it is impossible to achieve $\widetilde{O}(\min(\sqrt{kT}, d\sqrt{T} + \epsilon_*\sqrt{d}T))$ regret:

Lemma 2.6.2 (Implied by Theorem 24.4 in [41])**.** *Let $R_\nu(T)$ denote the cumulative regret at time T on environment ν. For any algorithm, there exists a 1-dimensional linear bandit environment ν_1 and a k-armed bandit environment ν_2 such that $R_{\nu_1}(T) \cdot R_{\nu_2}(T) \geq T(k-1)e^{-2}$.*

Experiment (Figure 2.1). Let $d = 2$. Consider a contextual bandit problem with $k = 50$ arms, where each arm j has an associated vector $a_j \in \mathbb{R}^d$ sampled uniformly at random from $[0,1]^d$. We consider two cases: (1) For a $\theta \in \mathbb{R}^d$ sampled uniformly at random from $[0,1]^d$, reward of arm j at time t is $a_j^\top \theta + \eta_t$, where $\eta_t \sim N(0,1)$, and (2) There are k parameters μ_j for $j \in [k]$ all sampled uniformly at random from $[0,10]$, so that the reward of arm j at time t is sampled from $N(\mu_j, 1)$. We use CORRAL with learning rate $\eta = \frac{2}{\sqrt{Td}}$ and UCB and LinUCB as base algorithm. In case (1) LinUCB performs better while in case (2) UCB performs better. Each experiment is repeated 500 times.

Contextual Bandits with Unknown Dimension

We consider model selection in the nested contextual linear bandit problem studied by [27]. In this problem the context space $\mathcal{A} \subset \mathbb{R}^D$. Each action is a $D-$dimensional vector and each context $\mathcal{A}_t$ is a subset of $\mathbb{R}^D$. The unknown parameter vector $\theta_* \in \mathbb{R}^D$ but only its first d_* coordinates are nonzero. Here, d_* is unknown and possibly much smaller than D. We assume access to a family of LinUCB algorithms $\{\mathcal{B}_i\}_{i=1}^M$ with increasing dimensionality d_i. Algorithm i is designed to 'believe' the unknown parameter vector θ_* has only nonzero entries in the first d_i entries. In [27] the authors consider the special case when $|\mathcal{A}_t| = k < \infty$ for all t. In order to obtain their model selection guarantees they require a lower bound on the average eigenvalues of the covariance matrices of all actions. In contrast, we do not require any such structural assumptions on the context. We provide the first sublinear regret for this problem when the action set is infinite. Further, we have no eigenvalue assumptions and our regret does not scale with the number of actions k.



Arms with linear rewards. Arms with non-linear rewards.

Figure 2.1: CORRAL with UCB and LinUCB bases. Shaded regions denote the standard deviations.

We use LinUCB with each value of $d \in [1, 2, 2^2, ..., 2^{\log(D)}]$ as a base algorithm for CORRAL and EXP3.P. We also consider the case when both the optimal dimension d_* and the misspecification ϵ_* are unknown: we use $M = \log(E) \cdot \log(D)$ modified LinUCB bases (see the discussion on Misspecified Contextual Linear Bandits above) for each value of (ϵ_*, d_*) in the grid $[1, 2, 2^2, ..., 2^{\log(E)}] \times [1, 2, 2^2, ..., 2^{\log(D)}]$.

From Lemma 2.4.4 and Lemma 2.4.5, for linear contextual bandit, LinUCB is (U, δ, T)-bounded with $U(t, \delta) = O(d\sqrt{t}\log(1/\delta))$ for infinite action sets and $U(t, \delta) = O(\sqrt{dt}\log^3(kT\log(T)/\delta))$ for finite action sets. Choose $\delta = 1/T$ and ignore the log factor, $U(t, \delta) = \widetilde{O}(d\sqrt{t})$ for infinite action sets and $U(t, \delta) = \widetilde{O}(\sqrt{dt})$ for finite action sets. Then $U(t) = c(\delta)t^\alpha$ with $\alpha = 1/2$ and $c(\delta) = \widetilde{O}(d)$ for infinite action sets, and $c(\delta) = \widetilde{O}(\sqrt{d})$ for finite action sets.

Now consider the misspecified linear contextual bandit problem with unknown d_* and ϵ_*. We use the smoothed LinUCB bases [42, 69]. Using the calculation in the proof of Theorem 2.6.1 in Section 2.6, using CORRAL with a smooth LinUCB base with parameters (d, ϵ) in the grids results in $\widetilde{O}\left(\frac{1}{\eta} + Td^2\eta + \epsilon\sqrt{d}T\right)$ regret. Since d is unknown, choosing $\eta = 1/\sqrt{T}$ yields the regret $\widetilde{O}\left(\sqrt{T}d_*^2 + \epsilon\sqrt{d}T\right)$. Using EXP3.P with a smooth LinUCB base with parameters (d, ϵ) in the grids results in:

$$\begin{aligned} R(T) &= \widetilde{O}\left(\sqrt{MT} + MTp + \frac{1}{p} + \frac{1}{p}U_i(Tp, \delta)\right) . \\ &= \widetilde{O}\left(\sqrt{MT} + MTp + \frac{1}{p} + \frac{1}{p}\left(d\sqrt{Tp} + \epsilon\sqrt{d}Tp\right)\right) . \\ &= \widetilde{O}\left(\sqrt{MT} + MTp + \frac{d\sqrt{T}}{p} + \epsilon\sqrt{d}T\right) . \end{aligned}$$

Since d_* is unknown, choosing $p = T^{-1/3}$ yields a $\widetilde{O}(T^{\frac{2}{3}}d_* + \epsilon_*\sqrt{d}T)$ regret bound. We summarize our results in the following table:

	Linear contextual bandit		Misspecified linear contextual bandit
	Unknown d_*		Unknown d_* and ϵ_*
	Finite action sets	Infinite action sets	
[27]	$\widetilde{O}(T^{2/3}k^{1/3}d_*^{1/3})$ or $\widetilde{O}(k^{1/4}T^{3/4} + \sqrt{kTd_*})$	N/A	N/A
EXP3.P	$\widetilde{O}(d_*^{\frac{1}{2}}T^{\frac{2}{3}})$	$\widetilde{O}(d_*T^{\frac{2}{3}})$	$\widetilde{O}(T^{\frac{2}{3}}d_* + \epsilon_*\sqrt{d}T)$
CORRAL	$\widetilde{O}\left(d_*\sqrt{T}\right)$	$\widetilde{O}\left(d_*^2\sqrt{T}\right)$	$\widetilde{O}\left(\sqrt{T}d_*^2 + \epsilon_*\sqrt{d}T\right)$

Non-parametric Contextual Bandit.

We study model selection in the setting of non-parametric contextual bandits.[29] consider non-parametric stochastic contextual bandits. At time t and given a context $x_t \in \mathbb{R}^D$, the learner selects arm $a_t \in [k]$ and observes the reward $f(a_t, x_t) + \xi_t$, where ξ_t is a 1-sub-Gaussian random variable and for all $a \in [k]$, the reward function $f(a, \cdot)$ is $L-$lipschitz in the context $x \in \mathbb{R}^D$. It is assumed that the contexts arrive in an IID fashion. [29] obtain a $\widetilde{O}\left(T^{\frac{1+d}{2+d}}\right)$ regret for this problem. Similar to [27], we assume that only the first d_* context features are relevant for an unknown $d_* < D$. It is important to find d_* because $T^{\frac{1+d_*}{2+d_*}} \ll T^{\frac{1+D}{2+D}}$. Stochastic CORRAL can successfully adapt to this unknown quantity: we can initialize a smoothed copy of Algorithm 2 of [29] for each value of d in the grid $[b^0, b^1, b^2, ..., b^{\log_b(D)}]$ for some $b > 1$ and perform model selection with CORRAL and EXP3.P with these base algorithms.

	EXP3.P	CORRAL
Nonparametric contextual bandit with unknown d_*	$\widetilde{O}\left(T^{\frac{1+bd_*}{2+bd_*} + \frac{1}{3(2+bd_*)}}\right)$	$\widetilde{O}\left(T^{\frac{1+2bd_*}{2+2bd_*}}\right)$

Tuning the Exploration Rate of ϵ-greedy

We study the problem of selecting for the optimal scaling for the exploration probability in the ϵ-greedy algorithm. Recall that for a given positive constant c, the ϵ-greedy algorithm pulls the arm with the largest empirical average reward with probability $1 - c/t$, and otherwise pulls an arm uniformly at random. Let $\epsilon_t = c/t$. It can be shown that the optimal value for ϵ_t is $\min\{1, \frac{5k}{\Delta_*^2 t}\}$ where Δ_* is the smallest gap between the optimal arm and the sub-optimal arms [41]. With this exploration rate, the regret scales as $\widetilde{\mathcal{O}}(\sqrt{T})$ for $k = 2$. We would like to find the optimal value of c without the knowledge of Δ_*. In this discussion we show it is possible to obtain such result by applying CORRAL to a set of ϵ-greedy base algorithms each instantiated with a c in $[1, 2, 2^2, ..., 2^{\log(kT)}]$.

Theorem 2.6.3. *The regret of CORRAL using smoothed ϵ-greedy base algorithms defined on the grid is bounded by $\widetilde{O}(T^{1/2})$ when $k = 2$.*

Proof. From Lemma 2.4.9, we lower bound the smallest gap by $1/T$ (because the gaps smaller than $1/T$ will cause constant regret in T time steps) and choose $\delta = 1/T^5$. From Theorem 2.4.11, the regret is $\widetilde{O}(T^{2/3})$ when $k > 2$ and $\widetilde{O}(T^{1/2})$ when $k = 2$ with the base running alone.

Next we show that the best value of c in the exponential grid gives a regret that is within a constant factor of the regret above where we known the smallest non-zero gap Δ_*. An exploration rates can be at most kT. Since $\frac{5K}{\Delta_*^2} > 1$, we need to search only in the interval $[1, KT]$. Let c_1 be the element in the exponential grid such that $c_1 \leq c^* \leq 2c_1$. Then $2c_1 = \gamma c^*$ where $\gamma < 2$ is a constant, and therefore using $2c_1 = \gamma c^*$ will give a regret up to a constant factor of the optimal regret. □

Experiment (Figure 2.2). Let there be two Bernoulli arms with means 0.5 and 0.45. We use 18 ϵ-greedy base algorithms differing in their choice of c in the exploration rate $\epsilon_t = c/t$. We take $T = 50,000$, $\eta = 20/\sqrt{T}$ and ϵ's to lie on a geometric grid in $[1, 2T]$. Each experiments is repeated 50 times.

[5] The shaded areas around UCB and CORRAL are the std. The shaded areas around the ϵ-greedy bases are 0.1 of std. For small ϵ, ϵ-greedy has a very high variance because it either commits to the optimal arm or the sub-optimal arm at the beginning, so plotting the whole *std* would make the plot unreadable.

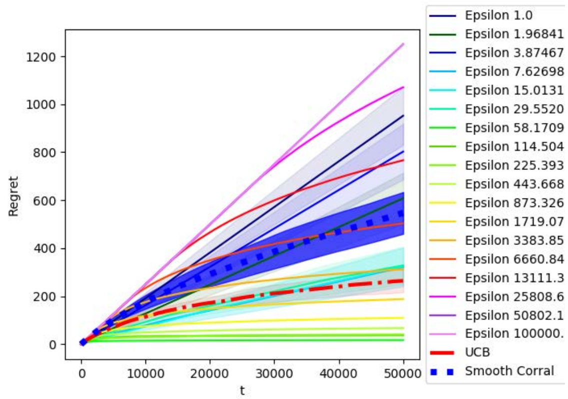


Figure 2.2: CORRAL with ϵ-Greedy bases with different exploration rates. [5]

Reinforcement Learning

We can instantiate Stochastic CORRAL model selection regret guarantees to the episodic linear MDP setting of [38], again with nested feature classes of doubling dimension just as in the case of the Contextual Bandits with Unknown Dimension. Let's formally define a Linear MDP,

Definition 2.6.4 (Linear MDP (Assumption A in [38]))**.** *An episodic MDP (Denoted by the tuple $(\mathcal{S}, A, H, \mathbb{P}, r)$) is a linear MDP with a feature map $\Phi : \mathcal{S} \times A \to \mathbb{R}^d$, if for any $h \in [H]$ there exist d unknown (signed) measures $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that for any $(s, a) \in \mathcal{S} \times A$, we have,*

$$\mathbb{P}_h(\cdot|s, a) = \langle \Phi(s, a), \boldsymbol{\mu}_h(\cdot) \rangle, \qquad r_h(s, a) = \langle \Phi(s, a), \boldsymbol{\theta}_h \rangle.$$

The value function for a linear MDP also satisfies a linear parametrization,

Proposition 2.6.5 (Proposition 2.3 from [38])**.** *For a linear MDP, and for any policy π there exist $d-$dimensional weights $\{\mathbf{w}_h^\pi\}_{h\in[H]}$ such that for any $(s, a, h) \in \mathcal{S} \times A \times [H]$ we have that the value function of policy π satisfies $Q_h^\pi(s, a) = \langle \Phi(s, a), \mathbf{w}_h^\pi \rangle$.*

For the purpose of studying model selection in the setting of linear MDPs we assume access to $D-$dimensional feature maps $\Phi : \mathcal{S} \times A \to \mathbb{R}^D$. For all policies π the unknown parameters $\{\mathbf{w}_h^\pi\}_{h\in[H]}$ are all assumed to have unknown coordinates only in their first d_* dimensions. We assume access to a family of LSVI-UCB (Algorithm 1 of [38]) algorithms $\{\mathcal{B}_i\}_{i=1}^M$ with increasing dimensionality d_i. Algorithm i is designed to 'believe' the unknown parameter vectors $\{\mathbf{w}_h^\pi\}_{h\in[H]}$ has only nonzero entries in the first d_i entries for all policies π.

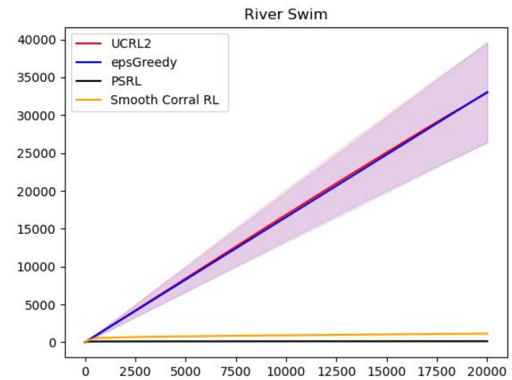


Figure 2.3: ϵ-Greedy vs UCRL2 vs PSRL in the River Swim environment [61].

Theorem 2.6.6. *Let $\mathcal{M} = (\mathcal{S}, A, H, \mathbb{P}, r)$ be a linear MDP parametrized by a feature map $\{\Phi : \mathcal{S} \times A \to \mathbb{R}^D\}$. Let $\{\Phi_i(s,a)\}_{i=1}^M$ be the family of nested feature maps such that $\Phi_i(s,a)$ corresponds to the top d_i entries of $\Phi(s,a)$. Assume that for all policies π the unknown parameters $\{\mathbf{w}_h^\pi\}_{h\in[H]}$ have nonzero coordinates only in their first d_* dimensions and that there exists an index i_* such that $d_* \le d_i \le 2d_*$. Selecting among different smoothed LSVI-UCB base algorithms corresponding to the feature maps $\{\Phi_i\}_{i=1}^M$ using Stochastic CORRAL with a CORRAL master and $\eta = \frac{M^{1/2}}{T^{1/2}d^{3/2}H^{3/2}}$ satisfies a regret guarantee: $R(T) \le \widetilde{\mathcal{O}}\left(\sqrt{Md^3H^3T}\right)$.*

Proof of Theorem 2.6.6. When well specified the LSVI-UCB algorithm [38] satisfies the high probability bound $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$ where H is the length of each episode. The result then follows from Theorem 2.4.11 by setting the CORRAL master learning rate as $\eta = \frac{M^{1/2}}{T^{1/2}d^{3/2}H^{3/2}}$. □

We also observe that in practice, smoothing RL algorithms such as UCRL and PSRL and using a CORRAL master on top of them can lead to improved performance. In Figure 2.3, we present results for the model selection problem among distinct RL algorithms in the River Swim environment [61]. We use three different bases, $\epsilon-$greedy $Q-$learning with $\epsilon = .1$, Posterior Sampling Reinforcement Learning (PSRL), as described in [51] and UCRL2 as described in [34]. The implementation of these algorithms and the environment is taken from TabulaRL (`https://github.com/iosband/TabulaRL`), a popular benchmark suite for tabular reinforcement learning problems. Smooth CORRAL uses a CORRAL master algorithm with a learning rate $\eta = \frac{15}{\sqrt{T}}$, all base algorithms are smoothed using Algorithm 8. The curves for UCRL2, PSRL and $\epsilon-$greedy are all of their un-smoothed versions. Each experiment was repeated 10 times and we have reported the mean cumulative regret and shaded a region around them corresponding to $\pm .3$ the standard deviation across these 10 runs.

Generalized Linear Bandits with Unknown Link Function

[45] study the generalized linear bandit model for the stochastic k-armed contextual bandit problem. In round t and given context $x_t \in \mathbb{R}^{d\times k}$, the learner chooses arm i_t and observes reward $r_t = \mu(x_{t,i_t}^\top \theta^*) + \xi_t$ where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector, ξ_t is a conditionally zero-mean random variable and $\mu : \mathbb{R} \to \mathbb{R}$ is called the link function. [45] obtain the high probability regret bound $\widetilde{O}(\sqrt{dT})$ where the link function is known. Suppose we have a set of link functions $\mathbb{L}$ that contains the true link function μ. Since the target regret $\widetilde{O}(\sqrt{dT})$ is known, we can run CORRAL with the algorithm in [45] with each link function in the set as a base algorithm. From Theorem 2.4.11, CORRAL will achieve regret $\widetilde{O}(\sqrt{|\mathbb{L}|dT})$.

Bandits with Heavy Tail

[59] study the linear stochastic bandit problem with heavy tail. If the reward distribution has finite moment of order $1+\epsilon_*$, [59] obtain the high probability regret bound $\widetilde{O}\left(T^{\frac{1}{1+\epsilon_*}}\right)$. We consider the problem when $\epsilon_* \in (0,1]$ is unknown with a known lower bound L where L is a conservative estimate and ϵ_* could be much larger than L. To the best of our knowledge, we provide the first result when ϵ_* is unknown. We use the algorithms in [59] with value of ϵ_* in the grid $[b^{\log_b(L)}, ..., b^1, b^0]$ for some $0 < b < 1$ as base algorithms with $\eta = T^{-1/2}$ for CORRAL. A direct application of Theorem 2.4.11 yields regret $\widetilde{O}\left(T^{1-0.5b\epsilon_*}\right)$. When $\epsilon_* = 1$ (as in the case of finite variance), $\widetilde{O}\left(T^{1-0.5b\epsilon_*}\right)$ is close to $\widetilde{O}\left(T^{0.5}\right)$ when b is close to 1.

2.7 Omitted proofs of Section 2.3

Bounding term I

When the base algorithms are not chosen, they repeat their step 2's policy to ensure that the conditional instantaneous regret is decreasing. To ensure the decreasing conditional instantaneous regret serves its purpose, when the base algorithms are chosen by the master, we only send step 2's rewards to the master as feedback signals. This is to ensure that the sequence of rewards the master is competing against satisfies the decreasing instantaneous regret condition. However, since the bases play and incur regrets from both step 1 and step 2 when they are chosen, we must account to the difference between the reward of step 1 and step 2 (that the bases incur when they play the arms), and 2 times the reward of step 2 (what the bases send to the master as feedback signals).

Since we assume all base algorithms to be smoothed and satisfy a two step feedback structure, we also denote by $\pi_t^{(j)}$ as the policy used by the master during round t, step j. Term I, the regret of the master with respect to base $i_\star$ can be written as:

$$\mathbb{E}\left[\mathrm{I}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(j)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})\right] \tag{2.15}$$

We cannot emphasize enough that the reader should keep in mind that the master algorithm is updated only using the reward of Step 2 of base algorithms even though the bases play both step 1 and 2. Let $\mathbb{T}_i$ is the random subset of rounds when $\mathcal{M}$ choose base $\widetilde{\mathcal{B}}_i$, $(i_t = i)$ for all $i \in [M]$. Adding and subtracting terms $\{f(\mathcal{A}_t^{(1)}, \pi_t^{(2)}\}_{t=1}^T$ we see that:

$$\begin{aligned}
\mathrm{I} &= \sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(j)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)}) \\
&= \underbrace{\sum_{t\in\mathbb{T}_{i_\star}}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(j)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})}_{\mathrm{I}_0} + \underbrace{\sum_{t\in\mathbb{T}_{i_\star}^c}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(j)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})}_{\mathrm{I}_1} \\
&\overset{(i)}{=} \underbrace{\sum_{t\in\mathbb{T}_{i_\star}}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(2)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(2)})}_{\mathrm{I}_0'} + \underbrace{\sum_{t\in\mathbb{T}_{i_\star}^c}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(2)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})}_{\mathrm{I}_1'} \\
&\overset{(ii)}{=} \underbrace{\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{t,i_\star}^{(2)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(2)})}_{\mathrm{I}_A} + \underbrace{\sum_{t\in\mathbb{T}_{i_\star}^c}^{T} f(\mathcal{A}_t^{(1)}, \pi_t^{(2)}) - f(\mathcal{A}_t^{(1)}, \pi_t^{(1)})}_{\mathrm{I}_B}
\end{aligned}$$

Equality (i) holds because term I_0 equals zero and therefore $\mathrm{I}_0 = \mathrm{I}_0'$ and in all steps $t \in \mathbb{T}_i^c$, base i repeated a policy of step 2 so that $\mathrm{I}_1 = \mathrm{I}_1'$. Equality (ii) follows from adding and subtracting term I_B. Term $\mathbb{E}\left[\mathrm{I}_A\right]$ is the regret of the master with respect to base i. Term $\mathbb{E}\left[\mathrm{I}_B\right]$ accounts for the difference between the rewards of step 1 and step 2 (that the bases incur) and 2 times the rewards of step 2 (that the bases send to the master). We now focus on bounding $\mathbb{E}\left[\mathrm{I}_A\right]$ and $\mathbb{E}\left[\mathrm{I}_B\right]$.

Biased step 2's rewards. We introduce the following small modification to the algorithm's feedback by setting $b_j(s) = \frac{U(s,\delta)}{s}$. This will become useful to control $\mathbb{E}\left[\mathrm{I}_B\right]$. Instead of sending the master the unadulterated $2r_{t,j}^{(2)}$ feedback, at all time step t, all bases will send the following modified feedback:

$$r_{t,j}^{(2)'} = r_{t,j}^{(2)} - \underbrace{\frac{U_j(s_{t,j},\delta)}{s_{t,j}}}_{b_j(s_{t,j})} \tag{2.16}$$

This reward satisfies:

$$\mathbb{E}\left[r_{t,j}^{(2)'}|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[f(\mathcal{A}_t^{(2)},\pi_t^{(2)})|\mathcal{F}_{t-1}\right] - \frac{U_j(s_{t,j},\delta)}{s_{t,j}}$$

In the coming discussion we'll show that this modification allows us to control term I_B. Since this modification is performed internally by all bases, we note that term I_A corresponds to an adversarial master that is always fed biased rewards from all bases and trying to compete against base i also with biased rewards. This means that any worst case bound of term I_A of an adversarial master will not be affected by this modification of the reward sequence of all bases.

Term I_B is the difference between the (modified) rewards of step 2 and step 1 which, due to the introduced modification, should intuitively be small because the cumulative (modified) rewards of step 2 are designed to be smaller than step 1. We will show that $\mathbb{E}\left[\mathrm{I}_B\right] \leq 8\sqrt{MT\log(\frac{4TM}{\delta})}$ and therefore that $\mathbb{E}\left[\mathrm{I}\right] \leq \mathbb{E}\left[\mathrm{I}_A\right] + 8\sqrt{MT\log(\frac{4TM}{\delta})}$.

Since any base j sends the biased reward to the master when it is chosen, when it is not chosen and repeats its step 2's policy, the reward also needs to be modified in the same way as in Equation 2.16. This is to ensure that the rewards of the base at time t do not depend on whether it is selected by the master at time t. We now discuss how the bias modification affects term II. Note that this modification increases term II (which only depends on base i) at each time step t by $b_j(s_{t,j}) = \frac{U_j(s_{t,j},\delta)}{s_{t,j}}$. Since the original instantaneous regret of base i at step 2 is bounded by a term of the same order, the modification increases term II by only a constant factor.

Bounding term $\mathbb{E}\left[\mathrm{I}_A\right]$

As we explain above, since the modification of the bases' rewards in Equation 2.16 is internal within the bases, and the master is a k-armed bandit adversarial algorithm, the worst-case performance of the master against any adversarial sequence of rewards will not be affect when the sequence of rewards of the bases changes.

CORRAL Master

Notice that:

$$\mathbb{E}\left[\mathrm{I}_A\right] = \mathbb{E}\left[\sum_{t=1}^{T} 2f(\mathcal{A}_t^{(2)}, \pi_{t,i_\star}^{(2)}) - 2f(\mathcal{A}_t^{(2)}, \pi_t^{(2)})\right]$$

We can easily bound this term using Lemma 13 from [5]. Indeed, in term I_A, the policy choice for all base algorithms $\{\widetilde{\mathcal{B}}_m\}_{m=1}^{M}$ during any round t is chosen before the value of i_t is revealed. This ensures the estimates $\frac{2r_t^{(2)}}{p_t^{i_t}}$ and 0 for all $i \neq i_t$ are indeed unbiased estimators of the base algorithm's rewards. We conclude:

$$\mathbb{E}\left[\mathrm{I}_A\right] \leq \mathcal{O}\left(\frac{M \ln T}{\eta} + T\eta\right) - \frac{\mathbb{E}\left[\frac{1}{\underline{p}_{i_\star}}\right]}{40\eta \ln T}$$

EXP3.P Master

Since $\mathbb{E}\left[\mathrm{I}_A\right]$ is the regret of base i with respect to the master, it can be upper bounded by the k-armed bandit regret of the master with M arms. Choose $\eta = 1, \gamma = 2k\beta$ in Theorem 3.3 in [14], we have that if $p \leq \frac{1}{2k}$, the regret of EXP3.P:

$$\mathbb{E}\left[\mathrm{I}_A\right] \leq \widetilde{\mathcal{O}}\left(MTp + \frac{\log(k\delta^{-1})}{p}\right)$$

Bounding $\mathbb{E}\left[\mathrm{I}_B\right]$

Notice that:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{I}_B\right] &= \mathbb{E}\left[\sum_{t \in \mathbb{T}_{i_\star}^c} f(\mathcal{A}_t^{(1)}, \pi_t^{(2)}) - f(\mathcal{A}_t^{(1)}, \pi_t^{(1)})\right] \\
&= \mathbb{E}\left[\sum_{t \in \mathbb{T}_{i_\star}^c} f(\mathcal{A}_t^{(2)}, \pi_t^{(2)}) - f(\mathcal{A}_t^{(1)}, \pi_t^{(1)})\right] \\
&= \mathbb{E}\left[\sum_{t \in \mathbb{T}_{i_\star}^c} f(\mathcal{A}_t^{(2)}, \pi_t^{(2)}) - f(\mathcal{A}_t^{(2)}, \pi^*) + f(\mathcal{A}_t^{(2)}, \pi^*) - f(\mathcal{A}_t^{(1)}, \pi_t^{(1)})\right] \\
&= \mathbb{E}\left[\sum_{t \in \mathbb{T}_{i_\star}^c} f(\mathcal{A}_t^{(2)}, \pi_t^{(2)}) - f(\mathcal{A}_t^{(2)}, \pi^*) + f(\mathcal{A}_t^{(1)}, \pi^*) - f(\mathcal{A}_t^{(1)}, \pi_t^{(1)})\right]
\end{aligned}$$

Substituting the biased Step 2 rewards in Equation 2.16 back into the expectation for $\mathbb{E}\left[\mathrm{I}_B\right]$ becomes:

$$\begin{aligned}
\mathbb{E}[\mathrm{I}_B] &= \mathbb{E}\Big[\sum_{t\in\mathbb{T}^c_{i_\star}} f(\mathcal{A}^{(2)}_t, \pi^{(2)}_t) - f(\mathcal{A}^{(2)}_t, \pi^*) - \frac{U_{j_t}(s_{t,j_t}(t),\delta)}{s_{t,j_t}} + \\
&\quad f(\mathcal{A}^{(1)}_t, \pi^*) - f(\mathcal{A}^{(1)}_t, \pi^{(1)}_t)\Big] \\
&= \sum_{j\neq i_\star} \mathbb{E}\Big[\sum_{t\in\mathbb{T}_j} f(\mathcal{A}^{(2)}_t, \pi^{(2)}_{t,j}) - f(\mathcal{A}^{(2)}_t, \pi^*) - \frac{U_j(s_{t,j},\delta)}{s_{t,j}} + \\
&\quad f(\mathcal{A}^{(1)}_t, \pi^*) - f(\mathcal{A}^{(1)}_t, \pi^{(1)}_{t,j})\Big] \\
&\overset{(1)}{\leq} \sum_{j\neq i} \mathbb{E}\Big[\sum_{t\in\mathbb{T}_j} f(\mathcal{A}^{(2)}_t, \pi^{(2)}_{t,j}) - f(\mathcal{A}^{(2)}_t, \pi^*) + f(\mathcal{A}^{(1)}_t, \pi^*) - \\
&\quad f(\mathcal{A}^{(1)}_t, \pi^{(1)}_{t,j})\Big] - U_j(s_{T,j},\delta)
\end{aligned} \tag{2.17}$$

Inequality (1) follows because by Lemma 2.8.1 applied to $U_j(t,\delta)$.

Observe that if the $j-$th algorithm was in its U_j-compatible environment (also referred to as its "natural environment"), then for any instantiation of $\mathbb{T}_j$ and with high probability:

$$\begin{aligned}
&\left(\sum_{t\in\mathbb{T}_j} f(\mathcal{A}^{(2)}_t, \pi^{(2)}_{t,j}) - f(\mathcal{A}^{(2)}_t, \pi^*) + f(\mathcal{A}^{(1)}_t, \pi^*) - f(\mathcal{A}^{(1)}_t, \pi^{(1)}_{t,j})\right) - U_j(T_j(T),\delta) \leq \\
&\qquad \left(\sum_{t\in\mathbb{T}_j} f(\mathcal{A}^{(1)}_t, \pi^*) - f(\mathcal{A}^{(1)}_t, \pi^{(1)}_{t,j})\right) - U_j(T_j(T),\delta) \leq 0
\end{aligned} \tag{2.18}$$

The first inequality follows because by definition $f(\mathcal{A}^{(2)}_t, \pi^*) \geq f(\mathcal{A}^{(2)}_t, \pi^{(2)}_t)$ and the last because of the high probability regret bound satisfied by $\widetilde{\mathcal{B}}_j$.

When $\widetilde{\mathcal{B}}_j$ is not in its U_j-compatible environment, this condition may or may not be violated. If this condition is violated, we need to make sure $\widetilde{\mathcal{B}}_j$ is dropped by the master. Since it is impossible to compute the terms $f(\mathcal{A}^{(2)}_t, \pi^{(2)}_t) - f(\mathcal{A}^{(2)}_t, \pi^*)$ and $f(\mathcal{A}^{(1)}_t, \pi^*) - f(\mathcal{A}^{(1)}_t, \pi^{(1)}_t)$ directly, we instead rely on the following test:

Base Test. Let $\mathbb{T}_j(l)$ be the set of time indices in $[l]$ when the master chose to play base j. We drop base $\widetilde{\mathcal{B}}_j$ if at any point during the history of the algorithm,

$$\sum_{t \in \mathbb{T}_j(l)} r_{t,j}^{(2)} - r_{t,j}^{(1)} > U_j(T_j(T), \delta) + 2\sqrt{2l \log\left(\frac{4TM}{\delta}\right)} \tag{2.19}$$

The logic of this step comes because as a simple consequence of the Azuma-Hoeffding martingale bound and Assumption 2.2.1, with probability at least $1 - \delta/M$ and for all $l \in [T]$:

$$\left| \sum_{\ell=1}^{l} f(\mathcal{A}_\ell^{(2)}, \pi^*) - f(\mathcal{A}_\ell^{(1)}, \pi^*) \right| \leq \sqrt{2l \log\left(\frac{4TM}{\delta}\right)} \tag{2.20}$$

$$\left| \sum_{\ell=1}^{l} r_{\ell,j}^{(2)} - r_{\ell,j}^{(1)} - f(\mathcal{A}_\ell^{(2)}, \pi_{\ell,j}^{(2)}) - f(\mathcal{A}_\ell^{(1)}, \pi_{\ell,j}^{(1)}) \right| \leq \sqrt{2l \log\left(\frac{4TM}{\delta}\right)} \tag{2.21}$$

This means that whenever $\widetilde{\mathcal{B}}_j$ is in its U_j-compatible environment, combining Equation 2.17, with Equation 2.20 and Equation 2.21 we get, with probability at least $1 - \delta$ for all $l \in [T]$:

$$\left| \left(\sum_{t \in \mathbb{T}_j(l)} r_{t,j}^{(2)} - r_{t,j}^{(1)} \right) - \left(\sum_{t \in \mathbb{T}_j(l)} f(\mathcal{A}_t^{(2)}, \pi_{t,j}^{(2)}) - f(\mathcal{A}_t^{(2)}, \pi^*) + f(\mathcal{A}_t^{(1)}, \pi^*) - f(\mathcal{A}_t^{(1)}, \pi_{t,j}^{(1)}) \right) \right| \leq 2\sqrt{2l \log\left(\frac{4TM}{\delta}\right)}$$

Plugging in inequality 2.18, we conclude that if $\widetilde{\mathcal{B}}_j$ is in its U_j-compatible environment with probability at least $1 - \delta$ for all $l \in [T]$:

$$\sum_{t \in \mathbb{T}_j(l)} r_{t,j}^{(2)} - r_{t,j}^{(1)} \leq U_j(s_{l,j}, \delta) + 2\sqrt{2l \log\left(\frac{4TM}{\delta}\right)}$$

Therefore the violation of condition in Equation 2.19, means $\widetilde{\mathcal{B}}_j$ couldn't have possibly been in its U_j-compatible environment. Furthermore, notice that in case Equation 2.19 holds (even if $\widetilde{\mathcal{B}}_j$ is not in its U_j-compatible environment), then with probability at least $1-\delta/M$:

$$\sum_{t\in\mathbb{T}_j(l)} f(\mathcal{A}_t^{(2)},\pi_{t,j}^{(2)}) - f(\mathcal{A}_t^{(2)},\pi^*) + f(\mathcal{A}_t^{(1)},\pi^*) - f(\mathcal{A}_t^{(1)},\pi_{t,j}^{(1)}) \leq U_j(s_{l,j},\delta) + 4\sqrt{2|\mathbb{T}_j(l)|\log(4TM)} \tag{2.22}$$

This test guarantees condition 2.22 is satisfied for all $j\in[M]$ and with probability at least $1-\delta$, thus implying:

$$\mathbb{E}\left[\mathrm{I}_B\right] \leq \sum_{j\neq i_\star} 4\sqrt{2|\mathbb{T}_j|\log(4TM)} \leq 8\sqrt{MT\log\left(\frac{4TM}{\delta}\right)}$$

The last inequality holds because $\sum_{j\neq i_\star}\sqrt{|\mathbb{T}_j|} \leq \sqrt{TM}$.

Bounding term II

Recall term II equals:

$$\mathbb{E}\left[\mathrm{II}\right] = \mathbb{E}\left[\sum_{t=1}^{T} f(\mathcal{A}_t,\pi^*) - f(\mathcal{A}_t,\pi_{s_{t,i},i})\right] \tag{2.23}$$

We use n_t^i to denote the number of rounds base i is chosen up to time t for all $i\in[M]$. Let $t_{l,i}$ be the round index of the $l-$th time the master chooses algorithm $\mathcal{B}_i$ and let $b_{l,i} = t_{l,i} - t_{l-1,i}$ with $t_{0,i}=0$ and $t_{n_T^i+1,i} = T+1$. Let $\mathbb{T}_i\subset[T]$ be the set of rounds where base i is chosen and $\mathbb{T}_i^c = [T]\backslash\mathbb{T}_i$. For $S\subset[T]$ and $j\in\{1,2\}$, we define the regret of the $i-$th base algorithm during Step j of rounds S as $R_i^{(j)}(S) = \sum_{t\in S} f(\mathcal{A}_t^{(j)},\pi^*) - f(\mathcal{A}_t^{(j)},\pi_{t,i}^{(j)})$. The following decomposition of $\mathbb{E}\left[\mathrm{II}\right]$ holds:

$$\mathbb{E}\left[\mathrm{II}\right] = \mathbb{E}\left[R_{i_\star}^{(1)}(\mathbb{T}_{i_\star}) + \underbrace{R_{i_\star}^{(2)}(\mathbb{T}_{i_\star}) + R_{i_\star}^{(1)}(\mathbb{T}_{i_\star}^c) + R_{i_\star}^{(2)}(\mathbb{T}_{i_\star}^c)}_{\mathrm{II}_0}\right]. \tag{2.24}$$

$R_{i_\star}^{(1)}(\mathbb{T}_{i_\star})$ consists of the regret when base $i_\star$ was updated in step 1 while the remaining 3 terms consists of the regret when the policies are reused by step 2.

Biased step 2's rewards

Note that we modified the rewards of step 2 as defined in Equation 2.16, both when the base is chosen and not chosen. We now analyze the effect of this modification:

$$
\begin{aligned}
&R(T)\\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi^*) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})\right]\\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)}) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})\right]}_{\text{I}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi^*) - f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)})\right]}_{\text{II}}\\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} \left(f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)}) - \mathbf{1}(t \in \mathbb{T}_{i_\star}^c \text{ or } j=2)\frac{U_i(s_{t,i_\star}, \delta)}{s_{t,i_\star}}\right) - f(\mathcal{A}_t^{(j)}, \pi_t^{(j)})\right]\\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi^*) - \left(f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)}) - \mathbf{1}(t \in \mathbb{T}_{i_\star}^c \text{ or } j=2)\frac{U_{i_\star}(s_{t,i_\star}, \delta)}{s_{t,i_\star}}\right)\right]\\
&\leq \text{I} - \text{modified} + \text{II} - \text{modified}
\end{aligned}
$$

Where I − modified and II − modified are defined as,

$$
\begin{aligned}
\text{I} - \text{modified} &= \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{j=1}^{2}\left(f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)}) - \mathbf{1}(t \in \mathbb{T}_{i_\star}^c \text{ or } j=2)\frac{U_{i_\star}(s_{t,i_\star}, \delta)}{s_{t,i_\star}}\right) -\\
&\qquad \left(f(\mathcal{A}_t^{(j)}, \pi_t^{(j)}) - \frac{U_{j_t}(s_{t,j_t}, \delta)}{s_{t,j_t}}\right)\Big]\\
\text{II} - \text{modified} &= \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{j=1}^{2} f(\mathcal{A}_t^{(j)}, \pi^*) -\\
&\qquad \left(f(\mathcal{A}_t^{(j)}, \pi_{s_{t,i_\star},i_\star}^{(j)}) - \mathbf{1}(t \in \mathbb{T}_{i_\star}^c \text{ or } j=2)\frac{U_i(s_{t,i_\star}, \delta)}{s_{t,i_\star}}\right)\Big]
\end{aligned}
$$

We provided a bound for term I-modified at the beginning of Section 2.7. In this section we concern ourselves with II−modified. Notice its expectation can be written as:

$$
\mathbb{E}\left[\text{II} - \text{modified}\right] = \mathbb{E}\left[\text{II}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2}\mathbf{1}(t \in \mathbb{T}_{i_\star}^c \text{ or } j=2)\frac{U_{i_\star}(s_{t,i_\star}, \delta)}{s_{t,i_\star}}\right]
$$

Now the second part of this sum is easy to deal with as it can be incorporated into the bound of $\mathbb{E}[\text{II}]$ by slightly modifying the bound given by Equation 2.25 below and changing $2b_l - 1$ to $2b_l + 1$. The rest of the argument remains the same.

Bounding $\mathbb{E}[\text{II}]$ when $\underline{p}_{i_\star}$ is fixed

From this section onward we drop the subscript $i_\star$ whenever clear to simplify the notations. In this section we show an upper bound for Term II when there is a value $\underline{p}_{i_\star} \in (0,1)$ that lower bounds $p_1^i, \cdots, p_T^{i_\star}$ with probability 1. We then use the restarting trick to extend the proof to the case when $\underline{p}_i$ is random in Theorem 2.4.10

Lemma 2.7.1 (Fixed $\underline{p}_{i_\star}$)**.** *Let $\underline{p}_{i_\star} \in (0,1)$ be such that $\frac{1}{\rho_{i_\star}} = \underline{p}_{i_\star} \le p_1^{i_\star}, \cdots, p_T^{i_\star}$ with probability one, then, $\mathbb{E}[\text{II}] \le 4\rho_{i_\star} U_i(T/\rho_{i_\star}, \delta)\log T + \delta T$.*

Proof of Lemma 2.7.1. Since $\mathbb{E}[\text{II}] \le \mathbb{E}[\mathbf{1}\{\mathcal{E}\}\text{II}] + \delta T$, we focus on bounding $\mathbb{E}[\mathbf{1}\{\mathcal{E}\}\text{II}]$. since base i is $(U, T, \delta)-$bounded, $\mathbb{E}\left[R^{(1)}_{i_\star}(\mathbb{T}_i)\mathbf{1}(\mathcal{E})\right] \le \mathbb{E}\left[U_{i_\star}(\delta, n_T^{i_\star})\mathbf{1}(\mathcal{E})\right]$. We proceed to bound the regret corresponding to the remaining terms in II_0:

$$\begin{aligned}
\mathbb{E}[\text{II}_0\mathbf{1}(\mathcal{E})] &= \mathbb{E}\left[\sum_{l=1}^{n_T^{i_\star}+1} \mathbf{1}\{\mathcal{E}\}(2b_l - 1)\mathbb{E}\left[r^{(2)}_{t_l, i_\star}|\mathcal{F}_{t_{l-1}}\right]\right] \\
&\le \mathbb{E}\left[\sum_{l=1}^{n_T^{i_\star}+1} \mathbf{1}\{\mathcal{E}\}(2b_l - 1)\frac{U_{i_\star}(l, \delta/2M)}{l}\right] \qquad (2.25)
\end{aligned}$$

The multiplier $2b_l - 1$ arises because the policies proposed by the base algorithm during the rounds it is not selected by $\mathcal{M}$ satisfy $\pi^{(1)}_{t,i_\star} = \pi^{(2)}_{t,i_\star} = \pi^{(2)}_{t_l,i}$ for all $l \le n^T_{i_\star} + 1$ and $t = t_{l-1} + 1, \cdots, t_l - 1$. The factorization is a result of conditional independence between $\mathbb{E}\left[r^{(2)}_{t_l,i_\star}|\mathcal{F}_{t_{l-1}}\right]$ and $\mathbb{E}\left[b_l|\mathcal{F}_{t_{l-1}}\right]$ where $\mathcal{F}_{t_{l-1}}$ already includes algorithm $\widetilde{B}_{i_\star}$ update right after round t_{l-1}. The inequality holds because $\widetilde{\mathcal{B}}_{i_\star}$ is $(U_{i_\star}, \frac{\delta}{2M}, \mathcal{T}^{(2)})-$smooth and therefore satisfies Equation 2.3 on event $\mathcal{E}$. Recall that as a consequence of Equation 2.24 we have

$$\mathbb{E}[\text{II}] \le \mathbb{E}\left[R^{(1)}_{i_\star}(\mathbb{T}_i)\mathbf{1}(\mathcal{E}) + \text{II}_0\mathbf{1}\{\mathcal{E}\}\right] + \delta T.$$

The first term is bounded by $\mathbb{E}\left[U_{i_\star}(n_T^{i_\star}, \delta)\mathbf{1}(\mathcal{E})\right]$ while the second term satisfies the bound in (2.25). Let $u_l = \frac{U_{i_\star}(l, \delta/2M)}{l}$. By Lemma 2.8.1, $\sum_{l=1}^t u_l \ge U_{i_\star}(t, \delta/M)$ for all

t, and so,

$$\mathbb{E}\left[\mathbf{1}\mathcal{E}U_{i_\star}(n_T^{i_\star},\delta)\right] \leq \mathbb{E}\left[\sum_{l=1}^{n_T^{i_\star}+1} \mathbf{1}\mathcal{E}u_l\right] . \tag{2.26}$$

By (2.25) and (2.26),

$$\mathbb{E}\left[R_{i_\star}^{(1)}(\mathbb{T}_{i_\star})\mathbf{1}(\mathcal{E}) + \mathrm{II}_0\mathbf{1}\{\mathcal{E}\}\right] \leq \mathbb{E}\left[\sum_{l=1}^{n_T^{i_\star}+1} \mathbf{1}\{\mathcal{E}\}2b_l u_l\right] .$$

Let $a_l = \mathbb{E}[b_l]$ for all l. Consider a master algorithm that uses $\underline{p}_{i_\star}$ instead of $p_t^{i_\star}$. In this new process let t_l' be the corresponding rounds when the base is selected, $\bar{n}_T^{i_\star}$ be the total number of rounds the base is selected, and $c_l = \mathbb{E}\left[t_l' - t_{l-1}'\right]$. Since $\underline{p}_{i_\star} \leq p_t^{i_\star}$ for all t it holds that $\sum_{l=1}^{j} a_l \leq \sum_{l=1}^{j} c_l$ for all j. If we use the same coin flips used to generate t_l to generate t_l', we observe that $t_l' \subset t_l$ and $\bar{n}_T^{i_\star} \leq n_T^{i_\star}$. Let $f : \mathbb{R} \to [0,1]$ be a decreasing function such that for integer $i_\star$, $f(i_\star) = u_{i_\star}$. Then $\sum_{l=1}^{n_T^{i_\star}+1} a_l u_l$ and $\sum_{l=1}^{\bar{n}_T^{i_\star}+1} c_l u_l$ are two estimates of integral $\int_0^T f(x)dx$. Given that $t_l' \subset t_l$ and u_l is a decreasing sequence in l,

$$\sum_{l=1}^{n_T^{i_\star}+1} \mathbb{E}\left[t_l - t_{l-1}\right] u_l \leq \sum_{l=1}^{\bar{n}_T^{i_\star}+1} \mathbb{E}\left[t_l' - t_{l-1}'\right] u_l \,,$$

and thus

$$\mathbb{E}\left[R_{i_\star}^{(1)}(\mathbb{T}_{i_\star})\mathbf{1}(\mathcal{E}) + \mathrm{II}_0\mathbf{1}\{\mathcal{E}\}\right] \leq \mathbb{E}\sum_{l=1}^{\bar{n}_T^{i_\star}+1} 2\mathbb{E}\left[t_l' - t_{l-1}'\right] u_l \,.$$

We proceed to upper bound the right hand side of this inequality:

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{\bar{n}_T^{i_\star}+1} u_l \mathbb{E}\left[t_l' - t_{l-1}'\right]\right] &\leq \mathbb{E}\left[\sum_{l=1}^{\bar{n}_T^{i_\star}+1} \frac{u_l}{\underline{p}_i}\right] \\ &\leq 2\rho_{i_\star} U_{i_\star}(T/\rho_{i_\star},\delta)\log(T).\end{aligned}$$

The first inequality holds because $\mathbb{E}\left[t_l' - t_{l-1}'\right] \leq \frac{1}{\underline{p}_{i_\star}}$ and the second inequality follows by concavity of $U_{i_\star}(t,\delta)$ as a function of t. The proof follows. □

Proof of Theorem 2.4.10

We use the restarting trick to extend Lemma 2.7.1 to the case when the lower bound $\underline{p}_{i_\star}$ is random (more specifically the algorithm (CORRAL) will maintain a lower bound that in the end will satisfy $\underline{p}_{i_\star} \approx \min_t p_t^{i_\star}$) in Theorem 2.4.10. We restate the theorem statement here for convenience.

Theorem 2.7.2 (Theorem 2.4.10)**.**

$$\mathbb{E}\left[\text{II}\right] \leq \mathcal{O}(\mathbb{E}\left[\rho_{i_\star}, U_{i_\star}(T/\rho_{i_\star}, \delta) \log T\right] + \delta T(\log T + 1)).$$

Here, the expectation is over the random variable $\rho_{i_\star} = \max_t \frac{1}{p_t^{i_\star}}$*. If* $U(t,\delta) = t^\alpha c(\delta)$ *for some* $\alpha \in [1/2, 1)$ *then,* $\mathbb{E}\left[\text{II}\right] \leq 4\frac{2^{1-\alpha}}{2^{1-\alpha}-1} T^\alpha c(\delta) \mathbb{E}\left[\rho_i^{1-\alpha}\right] + \delta T(\log T + 1)$*.*

Restarting trick: Initialize $\underline{p}_{i_\star} = \frac{1}{2M}$. If $p_t^{i_\star} < \underline{p}_{i_\star}$, set $\underline{p}_{i_\star} = \frac{p_t^{i_\star}}{2}$ and restart the base.

Proof of Theorem 2.4.10. The proof follows that of Theorem 15 in [5]. Let $\ell_1, \cdots, \ell_{d_i} < T$ be the rounds where Line 10 of the CORRAL is executed. Let $\ell_0 = 0$ and $\ell_{d_{i_\star}+1} = T$ for notational convenience. Let $e_l = [\ell_{l-1} + 1, \cdots, \ell_l]$. Denote by $\underline{p}_{i_\star, \ell_l}$ the probability lower bound maintained by CORRAL during time-steps $t \in [\ell_{l-1}, \cdots, \ell_l]$ and $\rho_{i_\star, \ell_l} = 1/\underline{p}_{i_\star, \ell_l}$. In the proof of Lemma 13 in [5], the authors prove $d_{i_\star} \leq \log(T)$ with probability one. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\text{II}\right] &= \sum_{l=1}^{\lceil \log(T) \rceil} \mathbb{P}(\underbrace{d_{i_\star} + 1 \geq l}_{I(l)}) \mathbb{E}\left[R_{i_\star}^{(1)}(e_l) + R_{i_\star}^{(2)}(e_l) | d_{i_\star} + 1 \geq l\right] \\
&\leq \log T \sum_{l=1}^{\lceil \log(T) \rceil} \mathbb{P}(I(l)) \mathbb{E}\left[4\rho_{i_\star, \ell_l} U_i(T/\rho_{i_\star, \ell_l}, \delta) | I(l)\right] + \delta T(\log T + 1) \\
&= \log T \mathbb{E}\left[\sum_{l=1}^{b_i+1} 4\rho_{i_\star, \ell_l} U_{i_\star}(T/\rho_{i_\star, \ell_l}, \delta)\right] + \delta T(\log T + 1).
\end{aligned}$$

The inequality is a consequence of Lemma 2.7.1 applied to the restarted segment $[\ell_{l-1}, \cdots, \ell_l]$. This step is valid because by assumption $\frac{1}{\rho_{i_\star, \ell_l}} \leq \min_{t \in [\ell_{l-1}, \cdots, \ell_l]} p_t$.

If $U_{i_\star}(t,\delta) = t^\alpha c(\delta)$ for some function $c : \mathbb{R} \to \mathbb{R}^+$, then $\rho_{i_\star} U(T/\rho_{i_\star},\delta) = \rho_{i_\star}^{1-\alpha} T^\alpha c(\delta)$. And therefore:

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{b_{i_\star}+1} \rho_{i_\star,\ell_l} U_{i_\star}(T/\rho_{i_\star,\ell_l},\delta)\right] &\le T^\alpha g(\delta)\mathbb{E}\left[\sum_{l=1}^{b_i+1} \rho_{i_\star,\ell_l}^{1-\alpha}\right] \\ &\le \frac{2^{\bar\alpha}}{2^{\bar\alpha}-1} T^\alpha c(\delta)\mathbb{E}\left[\rho_{i_\star}^{1-\alpha}\right]\end{aligned}$$

Where $\bar\alpha = 1-\alpha$. The last inequality follows from the same argument as in Theorem 15 in [5]. □

Proof of Theorem 2.4.11

Proof. For the CORRAL master,

$$\mathbb{E}[\mathrm{I}] \le \mathbb{E}[\mathrm{I}_A] + \mathbb{E}[\mathrm{I}_B] \le O\left(\frac{M\ln T}{\eta} + T\eta\right) - \frac{\mathbb{E}[\rho]}{40\eta\ln T} + 8\sqrt{MT\log(\frac{4TM}{\delta})}$$

Using Theorem 2.4.10 to control term II, the total regret of CORRAL is:

$$\begin{aligned}R(T) &\le \mathcal{O}\left(\frac{M\ln T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho}{40\eta\ln T} - 2\rho U(T/\rho,\delta)\log T\right] + \delta T + \\ &\quad 8\sqrt{MT\log(\frac{4TM}{\delta})} \\ &\le \mathcal{O}\left(\frac{M\ln T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho}{40\eta\ln T} - 2\rho^{1-\alpha}T^\alpha c(\delta)\log T\right] + \delta T + \\ &\quad 8\sqrt{MT\log(\frac{4TM}{\delta})} \\ &\le \widetilde{\mathcal{O}}\left(\sqrt{MT} + \frac{M}{\eta} + T\eta + Tc(\delta)^{\frac{1}{\alpha}}\eta^{\frac{1-\alpha}{\alpha}}\right) + \delta T,\end{aligned}$$

where the last step is by maximizing the function over ρ. Choose $\delta = 1/T$. When both α and $c(\delta)$ are known, choose $\eta = \frac{M^\alpha}{c(\delta)T^\alpha}$. When only α is known, choose $\eta = \frac{M^\alpha}{T^\alpha}$.

For the EXP3.P master, if $p \le \frac{1}{2k}$:

$$\mathbb{E}[I] \le \mathbb{E}[\mathrm{I}_A] + \mathbb{E}[\mathrm{I}_B] \le \widetilde{O}\left(MTp + \frac{\log(k\delta^{-1})}{p} + \sqrt{MT\log(\frac{4TM}{\delta})}\right)$$

EXP3.P	CORRAL
$\widetilde{\mathcal{O}}\left(\sqrt{MT} + MTp + T^{\alpha}p^{\alpha-1}c(\delta)\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{MT} + \frac{M}{\eta} + T\eta + T\, c(\delta)^{\frac{1}{\alpha}}\eta^{\frac{1-\alpha}{\alpha}}\right)$
$\widetilde{\mathcal{O}}\left(\sqrt{MT} + M^{\frac{1-\alpha}{2-\alpha}}T^{\frac{1}{2-\alpha}}c(\delta)^{\frac{1}{2-\alpha}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{MT} + M^{\alpha}T^{1-\alpha} + M^{1-\alpha}T^{\alpha}c(\delta)\right)$
$\widetilde{\mathcal{O}}\left(\sqrt{MT} + M^{\frac{1-\alpha}{2-\alpha}}T^{\frac{1}{2-\alpha}}c(\delta)\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{MT} + M^{\alpha}T^{1-\alpha} + M^{1-\alpha}T^{\alpha}c(\delta)^{\frac{1}{\alpha}}\right)$

Table 2.2: The top row shows the general regret guarantees. The middle row shows the regret guarantees when α and $c(\delta)$ are known. The bottom row shows the regret guarantees when α is known and $c(\delta)$ is unknown.

Using Lemma 2.7.1 to control term II, we have the total regret of EXP3.P when $\delta = 1/T$:

$$\begin{aligned} R(T) &= \widetilde{\mathcal{O}}(\sqrt{MT} + MTp + \frac{1}{p} + \frac{1}{p}U_i(Tp, \delta)) \ . \\ &= \widetilde{\mathcal{O}}(\sqrt{MT} + MTp + T^{\alpha}p^{\alpha-1}c(\delta)) \end{aligned}$$

When both α and $c(\delta)$ are known, choose $p = T^{-\frac{1-\alpha}{2-\alpha}}M^{-\frac{1}{2-\alpha}}c(\delta)^{\frac{1}{2-\alpha}}$. When only α is known, choose $p = T^{-\frac{1-\alpha}{2-\alpha}}M^{-\frac{1}{2-\alpha}}$. We then have the following regret: □

2.8 Ancillary Technical Results

Lemma 2.8.1. *If $U(t,\delta) = t^{\beta}c(\delta)$, for $0 \le \beta \le 1$ then:*

$$U(l,\delta) \le \sum_{t=1}^{l} \frac{U(t,\delta)}{t} \le \frac{1}{\beta}U(l,\delta)$$

Proof. The LHS follows immediately from observing $\frac{U(t,\delta)}{t}$ is decreasing as a function of t and therefore $\sum_{t=1}^{l}\frac{U(t,\delta)}{t} \ge l\frac{U(l,\delta)}{l} = U(l,\delta)$. The RHS is a consequence of bounding the sum by the integral $\int_0^l \frac{U(t,\delta)}{t}dt$, substituting the definition $U(t,\delta) = t^{\beta}c(\delta)$ and solving it. □

Lemma 2.8.2. *If $f(x)$ is a concave and doubly differentiable function on $x > 0$ and $f(0) \ge 0$ then $f(x)/x$ is decreasing on $x > 0$*

Proof. In order to show that $f(x)/x$ is decreasing when $x > 0$, we want to show that $\left(\frac{f(x)}{x}\right)' = \frac{xf'(x)-f(x)}{x^2} < 0$ when $x > 0$. Since $0f'(0) - f(0) \le 0$, we will show that

$g(x) = xf'(x) - f(x)$ is a non-increasing function on $x > 0$. We have $g'(x) = xf''(x) \leq 0$ when $x \geq 0$ because $f(x)$ is concave. Therefore $xf'(x) - f(x) \leq 0f'(0) - f(0) \leq 0$ for all $x \geq 0$, which completes the proof.

□

Lemma 2.8.3. *For any* $\Delta \leq \frac{1}{4} : \mathrm{KL}(\frac{1}{2}, \frac{1}{2} - \Delta) \leq 3\Delta^2$.

Proof. By definition $kl(p,q) = p\log(p/q) + (1-p)\log(\frac{1-p}{1-q})$, so

$$\begin{aligned}
\mathrm{KL}\left(\frac{1}{2}, \frac{1}{2} - \Delta\right) &= \frac{1}{2}\left(\log(\frac{1}{1-2\Delta}) + \log(\frac{1}{1+2\Delta})\right) \\
&= \frac{1}{2}\log\left(\frac{1}{1-4\Delta^2}\right) = \frac{1}{2}\log\left(1 + \frac{4\Delta^2}{1-4\Delta^2}\right) \\
&\leq \frac{2\Delta^2}{1-4\Delta^2} \leq \frac{2\Delta^2}{\frac{3}{4}} \leq 3\Delta^2
\end{aligned}$$

□

Chapter 3

The Explore-Commit-Eliminate Algorithm (ECE)

3.1 Introduction

Deep reinforcement learning has achieved impressive successes, yet often requires a very large amount of interaction data. This result is perhaps unsurprising, as more complicated function approximations often require more data to fit. Recent work on theoretical reinforcement learning for some structured function approximation settings has shown regret bounds that scale with a parameter characterizing the complexity of a particular function class. For example, for a type of function approximation by a d-dimensional linear model in Markov decision processes (MDPs), prior work has provided bounds that scale as $O(d^{3/2})$ regret [38], which have been improved to $O(d)$ even given small inherent Bellman error [69]. When the dynamics can be expressed using a matrix, $O(d^{3/2})$ regret bounds have also been provided [68]. The choice of dimension d is important: on one hand, if d is too small, such regret bounds typically either fail to hold or incur linear regret. On the other hand, if d is too large, the above regret bounds are unnecessarily large. Thus, a natural goal is to use the most compact representation suitable to encode the optimal policy for a domain (which we denote as d_*). This optimal representation is typically unknown a priori.

In this chapter we frame this as a model selection question among a set of algorithms with model classes, parameterized by dimensions $\{d \geq 1\}$, that are nested in their regret bound guarantees. We assume that at least one class can realize the true underlying domain. We ask if there is an algorithm that can achieve regret bounds that scale with the minimal realizable model class, given by d_*. Doing so seems subtle: provably efficient reinforcement learning algorithms typically rely heavily on strategic exploration, and using the wrong model class during learning may alias states, resulting in performance that appears strong under the current (incorrect) model class but is actually suboptimal. Conversely, forced exploration under more complex classes mitigates this problem, but could introduce regret that scales with the more complex model class dependence, even when a simpler model suffices.

Most prior work on model selection for online decision making has focused on contextual bandit settings. Here, minimax-optimal guarantees were recently shown under eigenvalue assumptions on the features by leveraging the special structure of the stochastic linear contextual bandit setting [18, 27]. These results also assume the *knowledge* of a good exploration policy, but such knowledge cannot be relied on in the reinforcement learning setting, where some "high-reward" states may only be observed under specific, initially unknown sequences of actions. Slightly weaker model selection guarantees can also be obtained under far more general assumptions by using a *corralling framework* that assumes access to a set of base algorithms, and provides a meta-algorithm that aims to realize the best regret of the (unknown) best algorithm [5, 7, 54] and Chapter 2.

Our contributions We tackle the challenge of model selection in RL under minimal assumptions. Our main insight is to leverage the knowledge of expected regret that is achievable under a particular model *when it realizes the data*. Thus, we propose an algorithm in Section 3.3 that maintains a candidate set of model classes at every round, and statistically tests whether each of them is well-specified, or not, by comparing the observed returns under that model class to the regret we should expect from a well-specified model. Model classes detected as misspecified at any round are permanently eliminated there-after in a manner reminiscent of *active-arm elimination* in the multi-armed bandit problem [23]; this is a significant simplification over previous meta-algorithms for model selection that were based on adversarial bandit algorithms. Our choice of action at every round carefully interleaves executing the candidate model class of minimal complexity with executing algorithms using higher-order models. This procedure is shown to automatically satisfy the needed exploration-exploitation trade-off for model selection. In Section 3.4, we show the regret bounds exactly match the model complexity of the unknown best model in d_* (and the finite episode length H in RL), and achieve a $T^{2/3}$ rate when the underlying algorithms have a $T^{1/2}$ rate under minimal assumptions about the underlying dynamics process. This is similar to recent model selection algorithms under general assumptions [54] which sacrifice either a tight dependence on T or d_*. We also demonstrate how our approach is compatible with multiple recently introduced RL results, and provide specific bounds for model selection in such settings. In addition to our algorithm being simpler than a recent model-selection approach [54], we provide new, significantly improved bounds for instances in which there is a constant gap in performance between model classes in Section 3.5. These guarantees are in part *instance-dependent*, as they scale inversely with this performance gap. From a practical perspective, our wrapper algorithm can be used given any input algorithms with regret guarantees that are nested, which will allow it to directly inherit future advances in provably efficient reinforcement learning. Finally, the computational complexity of our meta-algorithm only adds an extra factor on the order of the total number of model classes over and above the computational complexity of a single base algorithm.

3.2 Problem Statement

We consider the setting of an episodic Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{U}, H, r, P, \rho)$, where $\mathcal{S}$ and $\mathcal{U}$ are state and action spaces, $H \in \mathbb{N}$ is the length of an episode, $r = \{r_h(s_h, u_h)\}$ is the per-step reward function with $r_h(s_h, u_h) \in [0, 1]$, $P = \{P_h(s_{h+1}|s_h, u_h)\}$ is the transition dynamics, and $\rho(s)$ is a fixed initial state distribution. A policy maps times and states to actions, $\pi : [H] \times \mathcal{S} \to \mathcal{U}$.

For a given $h \in [H]$ and $s \in \mathcal{S}$, the value function is the expected cumulative reward following policy π:

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, u_{h'}) | s_h = s \right]$$

where $\mathbb{E}_\pi$ corresponds to the expectation over trajectories sampled according to policy π. We suppress dependence on $\mathcal{M}$ from V_h^π to avoid notational clutter. Similarly the action-value function is defined as the expected return from first taking action u and then following policy π: $Q_h^\pi(s, u) = r_h(s, u) + \mathbb{E}_{s' \sim P_h(\cdot|s,u)} V_{h+1}^\pi(s')$. The optimal value function is denoted $V_h^*(s) = \sup_\pi V_h^\pi(s)$. We write $V^\pi := \mathbb{E}_{s\sim\rho} V_1^\pi(s)$ and denote the optimal value under ρ as $V^* = \sup_\pi V^\pi$. In this work we primarily evaluate the quality of an algorithm $\mathcal{A}$ in an MDP $\mathcal{M}$ by its regret[1] with respect to the (unknown) optimal policy value V^* over T episodes:

$$\mathsf{Reg}_T(\mathcal{A}; \mathcal{M}) := \sum_{t=1}^{T} V^* - V^{\pi_t}. \tag{3.1}$$

We are interested in settings where the size of the state space $\mathcal{S}$ and/or action space $\mathcal{U}$ could be very large. Hence, we focus on function approximation methods

[1]Note that regret is here defined with respect to the optimal value. We will also consider algorithms satisfying "best-in-class" regret guarantees in Section 3.5.

for minimizing regret. A function approximation algorithm takes as input a model class $\mathcal{F}$ to generalize across states and actions [6]. Several natural examples include value-based classes where $\mathcal{F} : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ is used to predict action-value functions Q^π and model-based classes where $\mathcal{F} : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to \mathbb{R}$ is used to predict the transition dynamics P and reward r. Concretely, linear MDPs [38, 68] model the transition dynamics as $\langle \phi(s, a), \mu(s') \rangle$, where $\phi \in \mathbb{R}^d$ and μ is a d-dimensional vector of measures.

We let $(\mathcal{A}, \mathcal{F})$ denote the pair of algorithm $\mathcal{A}$ equipped with model class $\mathcal{F}$. Recent high probability regret (upper) bounds in this setting are sublinear in T and typically depend polynomially on $d_\mathcal{F}$, H, and $\log(T/\delta)$, where $d_\mathcal{F}$ is a measure of statistical complexity of $\mathcal{F}$ and $\delta \in (0, 1)$ is a failure probability. For example, if $\mathcal{F}$ is finite, we often have $d_\mathcal{F} = \log |\mathcal{F}|$ and if $\mathcal{F}$ is a class of linear functions of dimension d, we have $d_\mathcal{F} = d$. However, provably sublinear regret bounds in T are generally only known for algorithms under problem-specific assumptions for $\mathcal{F}$—for example, there exists $f^* \in \mathcal{F}$ such that the function approximation error is 0. If this condition holds, we say that $\mathcal{F}$ *realizes* the MDP $\mathcal{M}$. Conversely, if $\mathcal{F}$ does not realize $\mathcal{M}$, then it is *misspecified*. Since we consider settings where $\mathcal{F}$ may or may not realize $\mathcal{M}$ and realizability is almost universally assumed among modern RL algorithms with function approximation, we define a general notion of the regret of $\mathcal{A}$ using $\mathcal{F}$ under realizability, following [54].

Definition 3.2.1. *For an MDP $\mathcal{M}$, let algorithm $\mathcal{A}$ be equipped with a model class $\mathcal{F}$. Let $\mathcal{R}$ be a known function that is* $\mathrm{poly}(d_\mathcal{F}, H, \log(T/\delta))$*. The pair $(\mathcal{A}, \mathcal{F})$ is said to be $\mathcal{R}$-compatible if $\mathcal{F}$ realizes $\mathcal{M}$ and we have*

$$\mathsf{Reg}_t(\mathcal{A}; \mathcal{M}) \le \mathcal{R}(d_\mathcal{F}, H, \log(T/\delta)) \cdot \sqrt{t}.$$

for all t with probability at least $1 - \delta$. $\mathcal{R}$ is called a nominal regret coefficient[2] *for $(\mathcal{A}, \mathcal{F})$.*

The rationale behind $\mathcal{R}$-compatible algorithms is the following. For any $(\mathcal{A}, \mathcal{F})$, we may have a regret coefficient $\mathcal{R}$ in mind (from a provable guarantee) that holds if $\mathcal{F}$ realizes $\mathcal{M}$. The regret $\mathcal{R} \cdot \sqrt{t}$ reflects what we hope to achieve if $\mathcal{F}$ does actually realize $\mathcal{M}$, and $(\mathcal{A}, \mathcal{F})$ is only defined to be compatible if this happens. We remark that realizability is not necessary for a sublinear regret guarantee to hold, but most RL algorithms using function approximation assume it holds, so it is convenient to view both conditions together.

Note that Definition 3.2.1 requires that $\mathcal{A}$ be anytime, meaning the bound holds at any arbitrary round index $t \in [T]$ even though only the maximal number of rounds,

[2]It is not necessary that $\mathcal{R}$ depend only on these arguments; but these arguments are typically of interest in RL regret bounds.

T, may be specified. For algorithms without automatic anytime guarantees, this can be remedied up to constant factors via the doubling trick [16]. We will later give examples of how our model selection algorithm can be used with some recent single task RL algorithms with formal bounds in the function approximation setting.

Problem Statement Here, our goal in model selection is to obtain a regret guarantee that adapts on-the-fly to the model class of minimal complexity that remains competitive with the optimal value. That is, we wish to find the combination of algorithm $\mathcal{A}$ and model class $\mathcal{F}$, that is compatible in the sense of Definition 3.2.1, with the smallest possible leading coefficient $\mathcal{R}(d_{\mathcal{F}}, \cdot, \cdot)$. We consider a setting where we are choosing among a set of candidate algorithms $\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_L$ with model classes $\{\mathcal{F}_i\}_{i\in[L]}$, *known* nominal regret coefficients $\{\mathcal{R}_i\}_{i\in[L]}$, and complexities $\{d_i\}_{i\in[L]}$ where $d_i := d_{\mathcal{F}_i}$ and $\mathcal{F}_i$ is the model class of $\mathcal{A}_i$. Without loss of generality, we assume the algorithm-model class pairs can be ordered by their regret such that we have

$$\mathcal{R}_i(d_i, H, \log(T/\delta)) \leq \mathcal{R}_{i+1}(d_{i+1}, H, \log(T/\delta)) \tag{3.2}$$

for all $i \in [L-1]$, $T, H \in \mathbb{N}$, and $\delta \in (0,1)$. For example, if $\{\mathcal{A}_i\}$ are all instances of the same algorithm that use as input nested model classes $\{\mathcal{F}_i\}$, then (3.2) is satisfied by ordering $d_1 \leq \ldots \leq d_L$. This naturally captures, among other cases, linear models with nested features [27]. We also assume[3] that at least one algorithm is $\mathcal{R}_i$-compatible for its respective regret coefficient $\mathcal{R}_i$. Define $i_* = \min\{i \in [L] : (\mathcal{A}_i, \mathcal{F}_i) \text{ is } \mathcal{R}_i\text{-compatible}\}$.

We aim to design a meta-algorithm $\mathcal{A}$ that selects among $\{\mathcal{A}_i\}_{i=1}^L$ without knowing i_* *a priori* and, for some $\alpha \geq 0$ and $\beta \in [1/2, 1)$, achieves a guarantee of

$$\mathsf{Reg}_T(\mathcal{A}) = O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot L^\alpha T^\beta\right).$$

3.3 The Explore-Commit-Eliminate Algorithm

In this section, we present our model selection meta-algorithm, Explore-Commit-Eliminate (ECE) and detail the simple statistical test underlying our approach. Our meta-algorithm for model selection is described in Algorithm 10. At a high level, the algorithm proceeds in the following way. It takes as input the base algorithms and model classes, their nominal regret coefficients, and their model complexities;

[3] Note that for all other misspecified algorithms, their nominal regret bounds will, in general, not hold. As regret is being measured with respect to V^*, it will include the misspecification error terms.

mathematically, the input is given by $\{\mathcal{A}_i, \mathcal{F}_i, \mathcal{R}_i, d_i\}_{i\in[L]}$. The number of algorithms L, episodes $T \in \mathbb{N}$ and failure probability $\delta' \in (0, 1/e)$ are also specified. First, we set $\delta = \frac{\delta'}{10LT^2 \log_2 T}$. The meta-algorithm tracks a candidate algorithm index $\widehat{\imath}_t$, corresponding to pair $(\mathcal{A}_{\widehat{\imath}_t}, \mathcal{F}_{\widehat{\imath}_t})$ that is believed to be $\mathcal{R}_{\widehat{\imath}_t}$-compatible at any given time — as well as a set B_t of indices of algorithms with more complex models. At the start of each episode, the meta-algorithm determines whether to use the algorithm $\mathcal{A}_{\widehat{\imath}_t}$ or explore using a randomly selected algorithm from the indices B_t, based on the outcome of a Bernoulli variable U_t with success probability $1/t^\kappa$ where $\kappa \in (0, 1/2]$. This random variable U_t represents an indicator that model exploration will occur. After executing the policy from the chosen algorithm, the data is fed back to the algorithm to update, and a test is run to determine whether the algorithm should reject $\mathcal{A}_{\widehat{\imath}_t}$. The test checks the following condition for each $j \in B_t$:

$$\mathcal{G}_t(\widehat{\imath}_t, j) > \mathcal{W}(|\mathcal{T}_t^{\widehat{\imath}_t}|, \mathcal{R}_{\widehat{\imath}_t}, d_{\widehat{\imath}_t}, \delta)$$

where for all $i < j \in [L]$, $t \in [T]$, $\mathcal{T}_t^i$ is the set of times when $\mathcal{A}_i$ is chosen up to t, and $\mathcal{G}$ is a scaled estimate of the excess gap between models i and j, given by

$$\mathcal{G}_t(i, j) := \frac{|\mathcal{T}_t^i|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} g_{t'} - \sum_{t' \in \mathcal{T}_t^i} g_{t'}$$

and $\mathcal{W}$ is defined as

$$\begin{aligned}\mathcal{W}(t, \mathcal{R}, d, \delta) := &\ C_{\mathcal{W}} \cdot \mathcal{R}(d, H, \log(T/\delta)) \cdot \sqrt{t} \\ &+ C_{\mathcal{W}} \cdot H\sqrt{Lt^{1+\kappa} \cdot \log(1/\delta)} \\ &+ C_{\mathcal{W}} \cdot H\sqrt{t \cdot \log(1/\delta)}\end{aligned}$$

for a sufficiently large constant $C_{\mathcal{W}} > 0$. The test is only valid after a minimal burn-in period, $t \geq \tau_{\min}(\delta) = C_{\min} \cdot L^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)$ for a sufficiently large $C_{\min} > 0$, so this condition is also checked. If these conditions are true for some $j \in B_t$, meaning that the test fails, then ECE rejects $\mathcal{A}_{\widehat{\imath}_t}$ and switches to $\mathcal{A}_{\widehat{\imath}_t+1}$. This process repeats until episode T.

Note that although the algorithm uniformly explores among the algorithms in B_t, it does not require any explicit uniform or directed exploration within episodes that may be a tougher problem in RL settings than regret-minimization—one can simply run the algorithms as they were prescribed. In fact, we can interpret our meta-algorithm as automatically leveraging the exploration already in-built in the regret-minimizing base algorithms.

Algorithm 10: Explore-Commit-Eliminate (ECE)

Input: $\{\mathcal{A}_i, \mathcal{F}_i, \mathcal{R}_i, d_i\}_{i\in[L]}, L, T, \delta', \tau_{\min}(\cdot), \kappa$

$\delta \leftarrow \frac{\delta'}{10LT^2 \log_2 T}$, $\widehat{\imath}_t \leftarrow 1$, $\mathcal{T}_0^i = \emptyset$ for all $i \in [L]$, $B_1 = [2, L]$.

$U_t = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{t^\kappa} \\ 1 & \text{w.p. } \frac{1}{t^\kappa} \end{cases}$ for all $t \in [T]$.

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

  Set $j = \begin{cases} \widehat{\imath}_t & U_t = 0 \\ J_t \sim \text{Unif}\{B_t\} & U_t = 1 \end{cases}$

  $\mathcal{T}_t^j \leftarrow \mathcal{T}_{t-1}^j \cup \{t\}$ and $\mathcal{T}_t^k \leftarrow \mathcal{T}_{t-1}^k$ for all $k \neq j$.

  Rollout policy π_t from $\mathcal{A}_j$

  Observe $z_t := (s_{t,1}, u_{t,1}, \ldots, u_{t,H}, s_{t,H+1})$ and $g_t := \sum_{h\in[H]} r_{t,h}$

  Update $\mathcal{A}_j$ with t, z_t, g_t

  if $t \geq \tau_{\min}(\delta)$ *and there exists* $j \in B_t$ *such that* $\mathcal{G}_t(\widehat{\imath}_t, j) > \mathcal{W}(|\mathcal{T}_t^{\widehat{\imath}_t}|, \mathcal{R}_{\widehat{\imath}_t}, d_{\widehat{\imath}_t}, \delta)$ **then**

    $\widehat{\imath}_{t+1} \leftarrow \widehat{\imath}_t + 1$

    $B_{t+1} \leftarrow B_t \setminus \{\widehat{\imath}_{t+1}\}$

    If $\widehat{\imath}_{t+1} = L$, break and run $\mathcal{A}_L$ to end of time

  else

    $B_{t+1} \leftarrow B_t$

Statistical Test on Excess Gap

The ability of ECE to judiciously accept or reject base algorithms lies in the simple statistical test at the end of each episode. The test can be viewed as a comparison between the scaled expected return obtained by a "higher-order" algorithm, $\mathcal{A}_j$, corresponding to index $j \in B_t$ during exploration rounds; and that of the active candidate algorithm $\mathcal{A}_{\widehat{\imath}_t}$ during all rounds of its usage. If we find that the return of $\mathcal{A}_j$ is significantly higher than that of $\mathcal{A}_{\widehat{\imath}_t}$, it suggests that switching to the more complex algorithm $\mathcal{A}_j$ would yield significantly higher return, despite the fact that $\mathcal{A}_j$ has a larger nominal regret bound and might have received much less data than $\mathcal{A}_{\widehat{\imath}_t}$ (as it is also competing for data with the other algorithms in B_t). The requirement that $t \geq \tau_{\min}(\delta)$ and our special choice of exploration schedule ensures that the algorithms in B_t will have sufficient data to be useful in the test with high probability, while still exploiting the candidate model $\mathcal{A}_{\widehat{\imath}_t}$ whenever possible.

While we want ECE to reject lower-order models when they perform poorly, the test cannot be too sensitive. Otherwise, it could reject the optimal i_* and choose some unnecessarily large $j > i_*$, leading to highly suboptimal model complexity dependence in the regret bound. Our statistical test is designed to avoid this situation, as we prove in Section 3.4.

To give some additional intuition behind the test, it is useful to view the expected returns $\frac{1}{|\mathcal{T}_t^j|}\sum_{s\in\mathcal{T}_t^j} g_s$ as a noisy *lower bound* of the optimal value V^*; meanwhile the expected returns of $\frac{1}{|\mathcal{T}_t^{i_*}|}\sum_{s\in\mathcal{T}_t^{i_*}} g_s$ plus the regret incurred, $\mathsf{Reg}(\mathcal{A}_{i_*})$, should be an *upper bound* of the optimal value V^* up to some noise as well, if $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}_{i_*}$-compatible. Thus, as long as these intervals intersect, the test should succeed and i_* continues to be accepted. If the intervals separate, the current candidate is rejected. This intuition is reflected in the three terms comprising the definition of $\mathcal{W}$. The first is the nominal regret one expects to see from $\mathcal{A}_{\widehat{i}_t}$ if it is compatible. The last two follow from concentration of the averaging over returns of the algorithms.

3.4 Regret Guarantees for ECE

Our main result shows that the meta-algorithm automatically adapts to the regret of the optimal pair $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ that is $\mathcal{R}_{i_*}$-compatible. One of the main mechanisms behind this result is ensuring the validity of the test. The following lemma shows that ECE will never reject $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ with high probability.

Lemma 3.4.1. *We have $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$ with probability at least $1-\delta'$ for all $j \in [i_* + 1, L]$ and $t \geq \tau_{\min}(\delta'/10LT^2 \log_2 T)$.*

Thus, since the meta-algorithm steps through the base-algorithms incrementally, Lemma 3.4.1 shows that once it reaches $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$, the first $\mathcal{R}_{i_*}$-compatible pair, an algorithm with a more complex model class will never be selected. Our main theorem combines this result with the fact that, if the ECE has not rejected a misspecified algorithm $(\mathcal{A}_j, \mathcal{F}_j)$ with $j < i_*$, then the suboptimality of $\mathcal{A}_j$ must not be significant.

Theorem 3.4.2. *Let the model exploration parameter $\kappa = 1/3$. Then, the model selection algorithm ECE satisfies the regret bound*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}(d_{i_*}, H, \log(LT/\delta')) \cdot i_*^{1/3} L^{1/2} T^{2/3}\right).$$

with probability at least $1 - \delta'$, where $\widetilde{O}$ hides logs and terms independent of T and $\mathcal{R}$.

The regret bound of the meta-algorithm matches that of the optimal algorithm in dependence on the complexity of its model class d_{i_*} and horizon H, i.e., the best

Alg.	**Env.**	**Regret**
ModCB	CB	$\widetilde{O}\left(L^{2/3}d_{i_*}^{1/3}T^{2/3}\right)$
OSOM	CB	$\widetilde{O}\left(d_{i_*}^{1/2}T^{1/2}\right)$
CORRAL	RL	$\widetilde{O}\left(L^{1/2}\mathcal{R}_{i_*}^2T^{1/2}\right)$
Exp3.P	RL	$\widetilde{O}\left(L^{1/3}\mathcal{R}_{i_*}T^{2/3}\right)$
Ours	RL	MM: $\widetilde{O}\left(L^{5/6}\mathcal{R}_{i_*}T^{2/3}\right)$ ID: $\widetilde{O}\left(\frac{L^{5/2}\mathcal{R}_{i_*}^3}{\Delta_{\min}^2}+\mathcal{R}_{i_*}T^{1/2}+LT^{2/3}\right)$

Table 3.1: We compare the theoretical guarantees of our algorithm to recent model selection work: ModCB [27], OSOM [18], CORRAL [5, 54], and Exp3.P [54]. The first two apply to the contextual bandit (CB) setting and leverage distribution assumptions on the contexts to get nearly optimal regret. CORRAL and Exp3.P apply generally, but are suboptimal and require modifying the base algorithms in non-trivial ways. Our rate matches that of Exp3.P in the minimax (MM) setting without significant assumptions or modifications to the algorithms. We also achieve an improved instance-dependent (ID) rate when the gaps in performance between base algorithms are constant with minimal gap $\Delta_{\min}$.

dependence if the optimal algorithm were provided *a priori*. We do incur a worse dependence on T, which is now $T^{2/3}$, compared to the nominal $\sqrt{T}$ rate, and a dependence of $L^{1/2}$, the total number of algorithms, and i_*, the index of the optimal algorithm. Note that this type of trade-off in the parameter optimality for model selection is typical in recent results focused on contextual bandits, where methods making less strong assumptions typically incur sub-optimality in either the dependence on d_{i_*} or T. In particular, Theorem 3.4.2 matches the rate of Exp3.P [54] and does so without non-trivially modifying the base algorithms. In addition to the minimax guarantee of Theorem 3.4.2, we show in Section 3.5 that this can be improved to instance-dependent bounds, in contrast to Exp3.P and CORRAL.

Proofs

All proofs of Theorem 3.4.2, when not provided here, are available in Section 3.8. In this section, we prove Lemma 3.4.1 and provide a proof sketch for Theorem 3.4.2 to illustrate the main idea behind handling pairs $(\mathcal{A}_j, \mathcal{F}_j)$ that are not $\mathcal{R}_j$-compatible. In both cases, we require that three events hold and will show that they do with high probability. Define $\epsilon_t = g_t - V^{\pi_t}$ and let τ_i denote the first episode in which $\mathcal{A}_i$ is chosen as the candidate $\widehat{\imath}_t$. If $\mathcal{A}_i$ is never chosen then default to $\tau_i = T$. Recall that $\delta = \frac{\delta'}{10LT^2 \log_2 T}$.

1. Event E_1: For all $j \in [L]$ and all $t \in [T]$ such that $t \geq \tau_{\min}(\delta)$, if $t \leq \tau_i$, then $\frac{t^{1-\kappa}}{8L} \leq |\mathcal{T}_t^i| \leq 4t^{1-\kappa}$. If $t > \tau_i$, then $|\mathcal{T}_t^i| \leq t - \tau_i + 4t^{1-\kappa}$
2. Event E_2: For all $t \in [T]$,
$$\textstyle\sum_{t' \in \mathcal{T}_t^{i_*}} V^* - V^{\pi_{t'}} \leq \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta))\sqrt{|\mathcal{T}_t^{i_*}|}$$
3. Event E_3: For all $j \in [L]$ and all $t \in [T]$, $|\sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'}| \leq H\sqrt{2|\mathcal{T}_t^j| \log(2/\delta)}$

The first event ensures that the exploration schedule yields sufficient data to all the algorithms before they are chosen. The second states that the nominal anytime regret guarantee holds for $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$. The third handles concentration of the noisy returns that the algorithm observes from deploying policies. The following lemma shows that all three events happen with high probability.

Lemma 3.4.3. *The event* $E = \bigcap_{i \in \{1,2,3\}} E_i$ *holds with probability at least* $1 - 10LT^2\delta \log_2 T$.

Lemma 3.4.3 is proved in Section 3.8. The proof for the first event uses a Freedman inequality (details in Section 3.9) to bound the sizes of all sets given that enough time has passed. The second event holds with high probability under the assumption that $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}_{i_*}$-compatible. The third event can be shown to hold with high probability using the Azuma-Hoeffding inequality with appropriate union bounds.

Proof of Lemma 3.4.1

We now prove the statement of Lemma 3.4.1 under the event E. Adding and subtracting the sum of appropriately scaled value functions $\sum_{t' \in \mathcal{T}_t^j} V^{\pi_{t'}}$ and $\sum_{t' \in \mathcal{T}_t^{i_*}} V^{\pi_{t'}}$, we can write $\mathcal{G}_t(i_*, j)$ in terms of value functions and conditionally zero-mean errors:

$$\begin{aligned}
\mathcal{G}_t(i_*, j) &= \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} g_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} g_{t'} \\
&= \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} (V^{\pi_{t'}} + \epsilon_{t'}) - \sum_{t' \in \mathcal{T}_t^{i_*}} (V^{\pi_{t'}} + \epsilon_{t'}) \\
&\leq \sum_{t' \in \mathcal{T}_t^{i_*}} (V^* - V^{\pi_{t'}}) + \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} \epsilon_{t'}
\end{aligned}$$

The last inequality follows as $V^* \geq V^{\pi_{t'}}$ for all $t' \in [T]$. If events E_2 and E_3 hold then

$$\mathcal{G}_t(i_*, j) \leq \mathcal{R}_{i_*}(d_{i_*}, H, \log(1/\delta)) \cdot \sqrt{|\mathcal{T}_t^{i_*}|} + H\sqrt{2|\mathcal{T}_t^{i_*}|\log(2/\delta)} + H\sqrt{\frac{2|\mathcal{T}_t^{i_*}|^2}{|\mathcal{T}_t^{j}|}\log(2/\delta)}$$

By event E_1 and the fact that $j > i_*$ and $t \geq \tau_{\min}(\delta)$, $|\mathcal{T}_t^j| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^{i_*}|^{1-\kappa}}{8L}$. Therefore, for the third term,

$$H\sqrt{\frac{2|\mathcal{T}_t^{i_*}|^2}{|\mathcal{T}_t^{j}|}\log(2/\delta)} \leq H\sqrt{16L|\mathcal{T}_t^{i_*}|^{1+\kappa}\log(2/\delta)}$$

Applying this bound to the result in the previous display and given the definition of $\mathcal{W}$, it follows that $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$ for a sufficiently large constant $C_{\mathcal{W}} > 0$, independent of t, d_{i_*}, H, and δ.

Proof Sketch of Theorem 3.4.2

In bounding the regret of the meta-algorithm, there are three cases to handle: (1) before the test becomes valid, (2) once the test is valid but i_* has not been chosen yet, and finally (3) once i_* is chosen. We address the first and third cases before addressing the second, which is more involved. We define $\tau_* = \tau_{i_*}$ for shorthand.

Case (1): When $t < \tau_{\min}(\delta)$, the test to determine switching among any of the model classes is not yet valid. Here we simply pay the burn-in period giving $\mathsf{Reg}_{1:\tau_{\min}(\delta)-1} \leq O(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta))$.

Case (3): If $t > \tau_*$, then the meta-algorithm has switched to $\mathcal{A}_{i_*}$. Under event E, the condition in Lemma 3.4.1 is met and so the test no longer fails. Therefore $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ which is $\mathcal{R}_{i_*}$-compatible is not rejected in the remaining episodes. The regret during this phase scales as $\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{T}$ plus additional $O(HLT^{1-\kappa})$ regret due to exploration of the remaining base algorithms in B_t.

Case (2) is when $\tau_{\min} < t \leq \tau_*$—the test is eligible but the meta-algorithm is either switching among misspecified models or unable to detect that they are misspecified. Since the misspecification is not detected for any of the algorithms in B_t, we know $\mathcal{G}_t(\widehat{\imath}_t, i_*) \leq \mathcal{W}(|\mathcal{T}_t^{\widehat{\imath}_t}|, \mathcal{R}_{\widehat{\imath}_t}, d_{\widehat{\imath}_t}, \delta)$. That is, the average reward for $\mathcal{A}_{\widehat{\imath}_t}$ is not significantly different from that of $\mathcal{A}_{i_*}$. Since $\mathcal{A}_{i_*}$ is only played during exploration and $t \geq \tau_{\min}(\delta)$, its number of rounds played can be lower bounded by $t^{1-\kappa}/8L$ and thus its average regret is at most roughly

$$\widetilde{O}\left(\frac{L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta))}{t^{\frac{1-\kappa}{2}}}\right).$$

The success of the test suggests that the average reward of $\mathcal{A}_{\widehat{\imath}_t}$ should be close to this. Extrapolating over the rounds played by $\mathcal{A}_{\widehat{\imath}_t}$, the regret for $\widehat{\imath}_t$ will be

$$\widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot L^{1/2}|\mathcal{T}_t^{\widehat{\imath}_t}|^{\frac{1+\kappa}{2}}\right)$$

up to a constant shift by $\mathcal{W}(|\mathcal{T}_t^{\widehat{\imath}}|, \mathcal{R}_{\widehat{\imath}_t}, d_{\widehat{\imath}_t}, \delta)$. The shift is dominated by the above display because $\mathcal{R}_{\widehat{\imath}_t} \leq \mathcal{R}_{i_*}$ and $\kappa \in (0, 1/2]$. Finally, since we must account for the cumulative effect for all $i < i_*$, Jensen's inequality shows the sum of these terms is bounded above by

$$\widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} T^{\frac{1+\kappa}{2}}\right).$$

This becomes the dominant term in the regret. Additional regret of $O(HLT^{1-\kappa} + Hi_* + HT^{\frac{1+\kappa}{2}} \log^{1/2}(1/\delta))$ is also paid for exploration, switching costs, and estimation error of the averages. Summing these three cases and taking $\kappa = 1/3$ proves Theorem 3.4.2.

3.5 Instance Dependent Bounds

We now prove a stronger "instance-dependent" guarantee on online selection over more specialized base algorithms which have provable regret guarantees that are sublinear in T, but compared to the best policy within its respective policy class. For example, for an algorithm and model class $(\mathcal{A}, \mathcal{F})$ using value-based function approximation we might consider the greedy policy class:

$$\Pi_{\mathcal{F}} = \left\{(s, h) \mapsto \underset{u \in \mathcal{U}}{\operatorname{argmax}} \ f(s, u, h) \ : \ f \in \mathcal{F}\right\}.$$

The regret with respect to the best-in-class is

$$\mathsf{Reg}_T(\mathcal{A}, \Pi_{\mathcal{F}}; \mathcal{M}) = \max\nolimits_{\pi \in \Pi_{\mathcal{F}}} \sum\nolimits_{t \in [T]} V^{\pi} - V^{\pi_t}$$

To consider algorithms that may obtain sublinear regret with respect to this weaker benchmark but not with respect to V^*, we give a refined definition of $\mathcal{R}$-compatible algorithms.

Definition 3.5.1. *The pair $(\mathcal{A}, \mathcal{F})$ is said to be $\mathcal{R}^{\Pi_{\mathcal{F}}}$-compatible with respect to $\Pi_{\mathcal{F}}$ on the MDP $\mathcal{M}$ if we have*

$$\mathsf{Reg}_T(\mathcal{A}, \Pi_{\mathcal{F}}; \mathcal{M}) \leq \mathcal{R}^{\Pi_{\mathcal{F}}}(d_{\mathcal{F}}, H, \log(T/\delta)) \cdot \sqrt{t}$$

for all t with probability at least $1 - \delta$.

The value of $\max_{\pi \in \Pi_{\mathcal{F}}} V^\pi$ is typically unknown because of the complex dependence between $\Pi_{\mathcal{F}}$ and $\mathcal{M}$, and because $\Pi_{\mathcal{F}}$ is often determined by $\mathcal{F}$. Given a set of algorithms with different policy classes, we would like to select the one with the smallest regret compared to the optimal best-in-class value. Formally, we assume there are given algorithms $\{(\mathcal{A}_i, \mathcal{F}_i)\}$ with policy classes $\{\Pi_i\}$ each having optimal values $V_i^* := \max_{\pi \in \Pi_i} V^\pi$ and regret coefficients $\{\mathcal{R}_i^{\Pi_i}\}$ such that *for all* i the pair $(\mathcal{A}_i, \mathcal{F}_i)$ is $\mathcal{R}_i^{\Pi_i}$-compatible and $\mathcal{R}_i(d_i, \cdot, \cdot) \leq \mathcal{R}_{i+1}(d_{i+1}, \cdot, \cdot)$. Our goal is to select $i_* \in B_* := \operatorname{argmax}_{j \in [L]} V_j^*$ that has the smallest complexity dependence i.e. $i_* = \operatorname{argmin}_{i \in B_*} \mathcal{R}_i^{\Pi_i}(d_i, \cdot, \cdot)$. We emphasize that even if no algorithm is compatible in the sense of Definition 3.2.1, we want the optimal best-in-class guarantee[4] in the sense of Definition 3.5.1.

The difference between this setting and the last is that all algorithms are assumed to be compatible with respect to their own policy classes now, but the differing Π_i mean that some can have lower V_i^*, which we want to eliminate. Note that although the regret coefficients are ordered as in (3.2), the values $\{V_i^*\}$ are unknown and not necessarily ordered. Observe that $i_* = \min B_*$, so that $V_{i_*}^* > V_i^*$ for all $i < i_*$ and $V_{i_*}^* \geq V_i^*$ for all $i > i_*$. Thus i_* has the lowest regret for the best policy class. We would like an algorithm $\mathcal{A}$ that bounds $\mathsf{Reg}_T(\mathcal{A}, \Pi_{i_*}; \mathcal{M})$ with dependence on only the complexity of $\mathcal{F}_{i_*}$. The following result shows that Algorithm 10, without any modifications, can obtain an *instance-dependent* regret guarantee based on the size of the gaps $\Delta_{j,i_*} := V_{i_*}^* - V_j^*$ for $j < i_*$.

Theorem 3.5.2. *For a given $\mathcal{M}$, let $(\mathcal{A}_i, \mathcal{F}_i)$ be $\mathcal{R}_i^{\Pi_i}$-compatible with respect to Π_i for all $i \in [L]$. Then, with probability at least $1 - \delta'$, ECE with $\kappa = 1/3$ satisfies the regret bound with respect to policy class Π_{i_*}:*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^{3/2}(\mathcal{R}_{i_*}^{\Pi_{i_*}})^3 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-2}\right)$$

If $\kappa = 1/2$, then it satisfies

$$\widetilde{O}\left(HL\sqrt{T} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^{2}(\mathcal{R}_{i_*}^{\Pi_{i_*}})^4 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-3}\right)$$

Comparing this result to Theorem 3.4.2, if ECE is run with the same $\kappa = 1/3$ and the gaps are constant, a significantly better rate is possible since the third term has no dependence on T. With a more aggressive exploration choice of $\kappa = 1/2$, an even stronger instance-dependent guarantee is possible, matching the optimal $\mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T}$

[4] In essence, the best-in-class guarantee needs to hold even under model misspecification. A good example of a base algorithm satisfying this condition would be Exp4 in the contextual bandits setting.

rate of the best algorithm. However, this comes at the price of worse dependence on the gaps and $\mathcal{R}_{i_*}^{\Pi_{i_*}}$ factors, in the term that does not increase polynomially with T. In either case, Theorem 3.5.2 shows that we can obtain optimal or near-optimal dependence in T and only suboptimal $\mathcal{R}_{i_*}^{\Pi_{i_*}}$-dependence on terms that do not grow with T, as long as the gaps are constant. In Section 3.6, we show that these rates can be even further improved with only minimal modifications to ECE if given access to fast estimators of the gaps or V^*.

3.6 Implications of fast rates of estimating V^* and/or gap between policy classes

We previously discussed the recent results that prove PAC [48] and regret [54] results for model selection in RL given knowledge of V^*. We now show an analogous result for our setting. We use the framework of Algorithm 10 but set the probability of forced exploration to zero, i.e. set $\kappa = \infty$. Then, the test is modified to check the following condition for eliminating model $\widehat{\imath}_t$:

$$\sum_{t' \in \mathcal{T}_t^{\widehat{\imath}_t}} V^* - g_{t'} > \mathcal{W}_{V^*}(|\mathcal{T}_t^{\widehat{\imath}_t}|, \mathcal{R}_{\widehat{\imath}}, d_{\widehat{\imath}_t}, \delta)$$

where

$$\begin{aligned} \mathcal{W}_{V^*}(\Delta, \mathcal{R}, d, \delta) = C_{\mathcal{W}} \cdot \mathcal{R}(d, H, \log(1/\delta)) \cdot \sqrt{\Delta} \\ + C_{\mathcal{W}} \cdot H \sqrt{\Delta \cdot \log(1/\delta)} \end{aligned}$$

for a sufficiently large constant $C_{\mathcal{W}_{V^*}} > 0$. The test effectively measures the regret of $\mathcal{A}_{\widehat{\imath}_t}$ up to noise in g_t and rejects when we are confident that the regret does not match the nominal.

Proposition 3.6.1. *Given side information of the optimal value V^* for MDP $\mathcal{M}$, the above model selection algorithm $\mathcal{A}$ guarantees regret*

$$\mathsf{Reg}_T(\mathcal{A}) = \widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(LT/\delta')) \cdot \sqrt{LT}\right)$$

with probability at least $1 - \delta'$.

Proof. The proof is identical to that of Theorem 3.4.2 except for the handling of the misspecified case. For any model $j < i_*$ for which there is a time when the test succeeds,

$$\begin{aligned}
\sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} &= \sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} (V^* - g_t) + \sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&\le \mathcal{W}_{V^*}(|\mathcal{T}^j_t|, \mathcal{R}_j, d_j, \delta) + \sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&= O\left(\left(\mathcal{R}_{i_*} + H \log^{1/2}(1/\delta)\right) \cdot \sqrt{|\mathcal{T}^j_t|}\right)
\end{aligned}$$

Summing over all $j < i_*$ and using Jensen's inequality again shows that the dominant term remains $O(\mathcal{R}_{i_*}\sqrt{T})$ instead of $O(\mathcal{R}_{i_*}T^{2/3})$. □

This regret optimally matches the regret of the base algorithms in both $\mathcal{R}_{i_*}$ and T, but a dependence on L is still included.

Unfortunately, it is unclear whether such an assumption of knowing V^* is realistic in practice. An immediate alternative solution is to try to estimate V^* without first finding the optimal policy. The original test in Section 3.3 attempts this: the average returns of the algorithms in B_t act as a noisy lower bound of V^*. The test, however, is sensitive to the amount of exploration allocated to the base algorithms, and, since we are comparing to the nominal regret, the flat dependence on $\mathcal{R}$ is unlikely to improve. We hypothesize that better estimates of V^* can significantly improve the model selection guarantee.

In the following subsections, we consider the implications of having access to fast estimators, either of the optimal value $V^* := V^*_{i_*}$ or *gaps* between optimal values of different model orders, i.e. $\Delta_{i,j} := V^*_i - V^*_j$. We employ our instance-dependent analysis to show that improved regret rates can be obtained in both cases when the gap between the value of the optimal policy class and others is relatively large (i.e. constant). These consequences are demonstrated for the special case of linear contextual bandits, where such fast estimators are known to be available [21, 39, 40, 65].

Implications for access to a fast rate of estimating *gaps* in policy class optimal values

We first consider the possibility of fast rates in estimating the *gap* in optimal policy values, i.e. $\Delta_{i,j} := V^*_j - V^*_i$ for all $i < j$. In this section, we show that a modification of our ECE algorithm with a direct estimator of the gap in maximal values would yield improved model selection rates if there is a constant gap between all lower-order

Algorithm 11: Explore-Commit-Eliminate With Fast Gap Estimator And Forced Exploration Routines(ECE-Gap)

Input: $\{\mathcal{A}_i, \widetilde{\mathcal{A}}_i, \mathcal{F}_i, \mathcal{V}_i, d_i\}_{i\in[L]}, T, \delta', \tau_{\min}(\cdot), \kappa$

$\delta \leftarrow \frac{\delta'}{10LT^2\log_2 T}$, $\widehat{\imath}_t \leftarrow 1$, $\mathcal{T}_0^i = \emptyset$ for $i \in [L]$, $B_1 = [2, L]$

$U_t = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{t^\kappa} \\ 1 & \text{w.p. } \frac{1}{t^\kappa} \end{cases}$ for all $t \in [T]$.

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

  if $U_t = 0$ **then**

    Set $j \leftarrow \widehat{\imath}$.

  else

    Sample $J_t \sim \text{Unif}\{B_t\}$

    Set $j \leftarrow J_t$

  $\mathcal{T}_t^j \leftarrow \mathcal{T}_{t-1}^j \cup \{t\}$ and $\mathcal{T}_t^k \leftarrow \mathcal{T}_{t-1}^k$ for all $k \neq j$.

  IF $U_t = 0$: Rollout policy π_t from $\mathcal{A}_j$.

  ELSE: Rollout policy π_t from $\widetilde{\mathcal{A}}_j$.

  Observe $z_t := (s_{t,1}, u_{t,1}, \ldots, u_{t,H}, s_{t,H+1})$ and $g_t := \sum_{h\in[H]} r_{t,h}$

  Update $\mathcal{A}_j$ if $U_t = 0$, else update $\widetilde{\mathcal{A}}_j$ with t, z_t, g_t

  if $t \geq \tau_{\min}(\delta)$ *and there exists* $j \in B_t$ *such that* $\widehat{\Delta}_{\widehat{\imath}_t,j}(\mathcal{T}_t^j) > \mathbb{Z}(|\mathcal{T}_t^j|, \mathcal{V}_j)$ **then**

    $\widehat{\imath}_{t+1} \leftarrow \widehat{\imath}_t + 1$ $B_{t+1} \leftarrow B_t \setminus \{\widehat{\imath}_{t+1}\}$ If $\widehat{\imath}_{t+1} = L$, break and run $\mathcal{A}_L$ to end of time

  else

    $B_{t+1} = B_t$

models and the true model, i.e. $\Delta_{i,i_*} > 0$ for all i. Along with the replaced estimator, the radius of the statistical test is also modified according to the faster estimation error rate in the policy gap. For the special case of linear contextual bandits, these modifications will correspond *exactly* to the ModCB algorithm proposed by [27].

Since our focus is on instance-dependent analysis, we carry over the assumptions from Section 3.5, and further assume model nested-ness in the sense that $V_j^* = V^*$ for $j \geq i_*$. Thus, we get $\Delta_{i_*,i} = 0$ for all $i \geq i_*$, and $\Delta_{i,i_*} > 0$ for all $i < i_*$. To estimate the gap during exploration episodes, rather than running $\mathcal{A}_i$ directly, we allow an exploration algorithm $\widetilde{\mathcal{A}}_i$ to be run. In the case of [27] for contextual bandits, this would be an exploration policy that picks an arm uniformly at random from the set of K arms. Finally, we make the following assumption on the estimation error rate of the gaps.

Assumption 3.6.1. *For any $i < j$, we define $\widehat{\Delta}_{i,j}^{(n)}$ as an estimate of $\Delta_{i,j}$ that is a functional of the (context and reward) feedback obtained after running n exploration episodes for $\widetilde{\mathcal{A}}_j$. Then, we say that our estimate is $\mathcal{V}_j := \mathcal{V}(d_j, H, \log(1/\delta))$-consistent if, for some positive constant $C > 1$, we have*

$$|\widehat{\Delta}_{i,j}^{(n)} - \Delta_{i,j}| \leq \frac{\Delta_{i,j}}{C} + \frac{\mathcal{V}_j}{\sqrt{n}} \text{ for all } n \in [T] \text{ and } i < j \tag{3.3}$$

with probability at least $1 - \delta$. As with the earlier definition[5], $\mathcal{V}_j$ is poly and non-decreasing in d_j, H, $|\mathcal{U}|$, and $\log(LT/\delta)$).

The original estimator used in the ECE algorithm satisfies the above assumption with $\mathcal{V} := \mathcal{R}$. In what follows, we want to exploit situations in which we have available an estimator $\widehat{\Delta}_{i,j}$ with guarantee $\mathcal{V} \ll \mathcal{R}$; in particular, the dependence of the function $\mathcal{V}$ on dimension d could be significantly improved over any regret bound. While constructing such estimators is in general an open problem in RL, we do have one example for the linear contextual bandit problem where this is known to be possible.

Example 3.6.2. *[Linear contextual bandits.] Consider the stochastic d^{th}-order linear contextual bandits model as in [19], parameterized by K context distributions $\{\Sigma_i\}_{i=1}^K$, reward parameter $\theta^* \in \mathbb{R}^d$, and σ-sub-Gaussian noise in the rewards. Further, we carry over the assumptions from [27] of τ-sub-Gaussianity of the contexts and $\lambda_{min}(\overline{\Sigma}) \geq \nu > 0$ where $\overline{\Sigma} := \frac{1}{K}\sum_{i=1}^K \Sigma_i$ is the action-averaged covariance matrix. We assume that τ, ν are universal positive constants. Then, Assumption 3.6.1 holds with the choice of forced exploration $\widetilde{\mathcal{A}}_i$ that chooses arms uniformly at random from the set $[K]$ (regardless of round index t and model index i), with the choices $C = 2$ and $\mathcal{V}_i(d_i, \log(1/\delta))$ scaling as $\widetilde{O}(d_i^{1/4})$ for the estimator based on the square loss gap, used in [27]. Meanwhile, the regret bound for the base algorithms (e.g. instances of Exp4-IX) would give $\mathcal{R}_i$ scaling as $\widetilde{O}(d_i^{1/2})$. Further, note that Algorithm 2 exactly becomes the ModCB algorithm for this case.*

We now described the modified ECE algorithm, ECE-Gap, to work with a plugged-in estimate of $\Delta_{i,j}$ with the above guarantees. Note that the input now has extra "exploration algorithms" $\widetilde{\mathcal{A}}_i$, and what was earlier defined as regret bound leading factors, i.e. $\mathcal{R}_i$, is replaced by $\mathcal{V}_i$, the leading factors in the gap estimation error. Importantly, we are now using the fast estimator $\widehat{\Delta}_{i,j}(t)$ in place of the earlier estimator $\mathcal{G}_t(j, i)/|\mathcal{T}_t^j|$.

[5] Similar to $\mathcal{R}$, the definition of $\mathcal{V}_j$ can be general and include other problem dependent parameters as well.

Moreover, the threshold is now defined as:

$$\mathbb{Z}(n, \mathcal{V}) := \frac{\mathcal{V}}{\sqrt{n}}$$

Note that the threshold *is always applied to the more complex model* $d := d_i$ for $i > j$. The algorithm is stated formally in Algorithm 11. We derive the following instance-dependent result for this algorithm.

Proposition 3.6.3. *For a given $\mathcal{M}$, let Assumption 3.6.1 hold and let $\{\Delta_{i,i_*}\}_{i<i_*}$ be the gaps. Then, with probability at least $1-\delta'$, ECE-Gap in Algorithm 11 satisfies the regret bound*

$$\widetilde{O}\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{i=1}^{i_*-1} \min\{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{2}{1-\kappa}}\Delta_{i,i_*}^{-\frac{1+\kappa}{1-\kappa}}, \Delta_{i,i_*}T\}\right),$$

where regret is measured with respect to the optimal value V^*.

Before proving Proposition 3.6.3, let us consider its implication for the linear contextual bandits setting, ignoring dependence on $K = |\mathcal{U}|$ for now. Here, the modified ECE algorithm will essentially correspond to ModCB.

By choosing $\kappa = 1/3$ and using the gap estimator from [27], we can achieve an instance-dependent result with lower d_{i_*} dependence than that of Theorem 3.5.2 for the same setting of κ under the assumption of constant gaps. Furthermore, in the case the case of variable gaps, this result can immediately imply a minimax guarantee that matches that of [27].

Corollary 3.6.1. *For the linear contextual bandit problem, under the same setting as Corollary 3.6.2, with probability at least $1-\delta'$, Algorithm 11 with $\kappa = 1/3$ and constant gaps satisfies the instance-dependent regret bound*

$$\widetilde{O}\left(LT^{2/3} + \sqrt{d_{i_*}LT} + L^{3/2}d_{i_*}^{3/4}\sum_{i<i_*}\Delta_{i,i_*}^{-2}\right) = \widetilde{O}\left(LT^{2/3} + \sqrt{d_{i_*}LT}\right). \tag{3.4}$$

Furthermore, for variable gaps, let $\mathsf{Reg}_T(\mathcal{A}; \mathcal{M}, \{\Delta_{i,i_*}\}_i)$ denote the regret as a function of the gaps. Since $\min\{L^{3/2}\mathcal{V}_{i_*}^3\Delta_{i_*,i}^{-2}, \Delta_{i_*,i}T\} \leq L^{1/2}\mathcal{V}_{i_*}T^{2/3}$, ECE-Gap also satisfies the minimax regret bound

$$\sup_{\Delta_{i,i_*}>0\;:\;i<i_*} \mathsf{Reg}_T\left(\textit{ECE-Gap}; \mathcal{M}, \{\Delta_{i,i_*}\}_i\right) = \widetilde{O}\left(Ld_{i_*}^{1/4}T^{2/3} + \sqrt{d_{i_*}LT}\right).$$

The equality in (3.4) uses $d_i \ll T$ for all $i \in [L]$ and the constant gap assumption. If we knew *a priori* that the gaps are constant, the instance-dependent bound in (3.4) can be improved by a more aggressive choice of $\kappa = 1/2$, as in Theorem 3.5.2. We can then achieve the desired regret rate of $\widetilde{O}(\sqrt{d_{i_*}T})$ regret *if and only if* the gaps are constant. Again there is only sub-optimal d_{i_*}-dependence on the term independent of T.

Corollary 3.6.2. *For the linear contextual bandit problem under Assumption 3.6.1 with constant gaps $\{\Delta_{j,i_*}\}_{j<i_*}$, let $\mathcal{V}_{i_*} := \widetilde{O}(d_{i_*}^{1/4})$ and $\mathcal{R}_{i_*}^{\Pi_{i_*}} := \widetilde{O}(d_{i_*}^{1/2})$. Then, with probability at least $1-\delta'$, Algorithm 11 with $\kappa = 1/2$ satisfies the regret bound*

$$\widetilde{O}\left(L\sqrt{T} + \sqrt{d_{i_*}LT} + L^2 d_{i_*} \sum_{i<i_*} \Delta_{i,i_*}^{-3}\right) = \widetilde{O}\left(L\sqrt{T} + \sqrt{d_{i_*}LT}\right).$$

In summary, Proposition 3.6.3 not only recovers the minimax rate, but shows an improved instance-dependent guarantee for more favorable cases when the gap between optimal policy values is larger.

Let us now prove the proposition.

Proof. Let $\widehat{\Delta}_{i,j}^t := \widehat{\Delta}_{i,j}^{(|\mathcal{T}_t^i|)}$. First, we show that under the intersection of the event of Equation (3.3) and event E' of Theorem 3.5.2, we will never reach $\widehat{\imath}_t > i_*$. For every $i > i_*$, and all $t \geq 1$, Equation (3.3) gives us

$$\widehat{\Delta}_{i_*,i}^t \leq \frac{\mathcal{V}_i}{\sqrt{|\mathcal{T}_t^i|}}$$

Thus, model order i_* is never rejected under this event, and higher order models have no contribution to the overall regret.

Next, we bound the regret arriving from the misspecified models $i < i_*$. We do this by bounding the number of rounds during which model order $i < i_*$ is used, given by $|\mathcal{T}_T^i|$. From Equation (3.3), we get

$$\begin{aligned}
\Delta_{i,i_*} &\leq \widehat{\Delta}_{i,i_*}^t + \frac{\Delta_{i,i_*}}{C} + \frac{\mathcal{V}_{i_*}}{\sqrt{|\mathcal{T}_t^{i_*}|}} \\
\implies \Delta_{i,i_*} &\leq \frac{C}{C-1}\left(\widehat{\Delta}_{i_*,i}^t + \frac{\mathcal{V}_{i_*}}{\sqrt{|\mathcal{T}_t^{i_*}|}}\right) \\
&\leq \frac{C\mathcal{V}_{i_*}}{(C-1)\sqrt{|\mathcal{T}_t^{i_*}|}}
\end{aligned}$$

where the last inequality follows because the condition in the test has not yet been violated. More-over, since model i_* has not been selected yet, we have $|\mathcal{T}_t^{i_*}| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^i|^{1-\kappa}}{8L}$. This gives us

$$\Delta_{i,i_*} \leq \frac{8(CL)^{1/2}\mathcal{V}_{i_*}}{\sqrt{C-1}|\mathcal{T}_t^i|^{\frac{1-\kappa}{2}}}$$
$$\implies |\mathcal{T}_t^i| = \mathcal{O}\left(\frac{L^{\frac{1}{1-\kappa}}(\mathcal{V}_{i_*})^{\frac{2}{1-\kappa}}}{\Delta_{i,i_*}^{\frac{2}{1-\kappa}}}\right)$$

Thus, the total contribution to the regret from the misspecified model i is given by

$$T^{1-\kappa} + |\mathcal{T}_t^i|\Delta_{i,i_*} + \mathcal{R}_i^{\Pi_i}\sqrt{|\mathcal{T}_t^i|}$$
$$\leq T^{1-\kappa} + |\mathcal{T}_t^i|\Delta_{i,i_*} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{|\mathcal{T}_t^i|}.$$

The first term comes from the forced exploration, and the last term is equivalent to the regret we would pay anyway if we knew $i_* = 2$ beforehand. Focusing on the second term, the contribution to regret is upper bounded by

$$\min\left\{\Delta_{i,i_*}T, \left(\frac{C_{\mathbb{Z}}L^{1/2}\mathcal{V}_{i_*}}{\Delta_{i,i_*}}\right)^{\frac{2}{1-\kappa}} \cdot \Delta_{i,i_*}\right\}$$

□

Implications for a fast rate of estimating V^*

An alternative setting is one where we have access to an estimator of V^* instead of an estimator of the gap. Corollary 1 of [40] shows that an ϵ-close approximation of V^* is possible in $\widetilde{O}\left(\sqrt{d}/\epsilon^2\right)$ interactions in the disjoint linear bandit setting (where there is a different parameter vector for each arm) under Gaussian assumptions. Whether or not such fast estimators exist or are practical for other general settings is still open, but future work on this problem could be applied to the instance dependent results here.

We will retain the same problem assumptions as the previous subsection. We also assume there is $\widehat{V}_i$ for each $i \in [L]$. Each estimator offers a high-probability guarantee on the estimation error as a function of the number of exploration episodes using corresponding exploration algorithms $\{\widetilde{\mathcal{A}}_i\}$.

Assumption 3.6.2. *For all $i \in [L]$, we define the $\widehat{V}_i^{(n)}$ where $n \in [T]$ as the estimator of V_i^* given n exploration rounds with $\widetilde{\mathcal{A}}_i$. We assume with probability at least $1-\delta$, for all $i \geq i_*$, the estimator $\widehat{V}_i^{(n)}$ satisfies*

$$|V^* - \widehat{V}_i^{(n)}| \leq \frac{\mathcal{V}_i}{n^\alpha} + \frac{\mathcal{V}_i'}{n^\beta} \tag{3.5}$$

where $\mathcal{V}_i$ and $\mathcal{V}_i'$ are poly and increasing in d, H, $|\mathcal{U}|$, and $\log(LT/\delta))$ and $\alpha, \beta \in (0,1)$.

Let $\widehat{V}_i^t := \widehat{V}_i^{(|\mathcal{T}_t^i|)}$. The algorithm will be of the same form as Algorithm 11, but instead we leverage the following alternative test:

$$\sum_{t \in \mathcal{T}_t^{\widehat{\imath}_t}} \widehat{V}_j^t - g_{t'} \leq \mathbb{Z}_{\widehat{\imath}}(|\mathcal{T}_t^{\widehat{\imath}_t}|, \mathcal{V}_j, \mathcal{V}_j') \tag{3.6}$$

where

$$\mathbb{Z}_i(t, \mathcal{V}, \mathcal{V}') := C_{\mathbb{Z}} \left(\mathcal{V}_j L^\alpha t^{1-(1-\kappa)\alpha} + \mathcal{V}_j' L^\beta t^{1-(1-\kappa)\beta} + H\sqrt{t \log(1/\delta)} + \mathcal{R}_i^{\Pi_i} \sqrt{t} \right)$$

for a sufficiently large constant $C_{\mathbb{Z}} > 0$. That is, if the above inequality holds, then ECE continues to use $\widehat{\imath}_t$; otherwise, ECE switches to $\widehat{\imath}_t + 1$ for round $t+1$. First, we prove an analogous result to Lemma 3.4.1, showing that the test will not fail under the good event E''. Here, we let $E'' = E' \cap E_4$ where E' is the event from Theorem 3.5.2 and event E_4 is the following.

Event E_4: Let $\{\widehat{V}_i\}$ be the estimators from Assumption 3.6.2. For all $i \geq i_*$ and $n \in [T]$, equation (3.5) is satisfied.

Note that E_4 holds with probability at least $1-\delta$ by assumption. Therefore E'' still holds with probability at least $1 - 10LT^2\delta \log_2(T)$.

Lemma 3.6.4. *Given that event E' holds, then for all $t \geq \tau_{\min}$ and $j \in [i_*+1, L]$, it holds that $\sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_t^j - g_{t'} \leq \mathbb{Z}_{i_*}(|\mathcal{T}_t^{i_*}|, \mathcal{V}_j, \mathcal{V}_j')$*

Proof. Since $j > i_*$, we use the assumption on the estimator $\widehat{V}_j$ to write the difference in terms of regret, estimation error and noise:

$$\begin{aligned}
\sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} &\leq \sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - V^{\pi_{t'}} - \epsilon_{t'} \\
&\leq \frac{\mathcal{V}_j |\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|^\alpha} + \frac{\mathcal{V}_j' |\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|^\beta} + \sum_{t' \in \mathcal{T}_t^{i_*}} V^* - V^{\pi_{t'}} - \epsilon_{t'}
\end{aligned}$$

Then note that $\sum_{t'\in|\mathcal{T}_t^{i_*}|}\epsilon_{t'} \leq H\sqrt{2|\mathcal{T}_t^{i_*}|\log(2/\delta)}$ and $\sum_{t'\in|\mathcal{T}_t^{i_*}|} V^* - V^{\pi_{t'}} \leq \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{|\mathcal{T}_t^{i_*}|}$ under event E'. Furthermore, under E', we have $|\mathcal{T}_t^j| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^{i_*}|^{1-\kappa}}{8L}$, which implies

$$\sum_{t'\in\mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} \leq C_{\mathbb{Z}}\left(\mathcal{V}_j L^{\alpha}|\mathcal{T}_t^{i_*}|^{1-(1-\kappa)\alpha} + \mathcal{V}_j' L^{\beta}|\mathcal{T}_t^{i_*}|^{1-(1-\kappa)\beta} + H\sqrt{|\mathcal{T}_t^{i_*}|\log(2/\delta)} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{|\mathcal{T}_t^{i_*}|}\right)$$

for $C_{\mathbb{Z}}$ large enough. Therefore, it holds that $\sum_{t'\in\mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} \leq \mathbb{Z}_{i_*}(|\mathcal{T}_t^{i_*}|, \mathcal{V}_j, \mathcal{V}_j')$. □

The main proposition states that a better instance-dependent rate is available under less restrictive assumptions on "realizability" by utilizing the test based on the V^* estimators.

Proposition 3.6.5. *For a given $\mathcal{M}$, let Assumption 3.6.2 hold some for α, β and $i \geq i_*$ and let $\kappa \in (0, 1/2]$. Then, with probability at least $1-\delta'$, ECE in Algorithm 10 with the modified test (Equation 3.6) satisfies the regret bound*

$$\widetilde{O}\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}}, \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}'^{\frac{1}{(1-\kappa)\beta}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}}, \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}} + H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\right)$$

Proof. As discussed previously, the sufficient events occur with probability at least $1-\delta'$. Similar to Theorem 3.5.2, we now show that the gaps Δ_{j,i_*} can be bounded by using the estimation error of $\widehat{V}^{i_*}$ and the concentration bounds from E'. Let t be such that $\widehat{\imath}_t = j$ and the test succeeds. Then,

$$\begin{aligned}
\Delta_{j,i_*} &= V^* - V_j^* \\
&\le \widehat{V}_{i_*}^t + \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}'_{i_*}}{|\mathcal{T}_t^{i_*}|^\beta} - \frac{1}{|\mathcal{T}_t^j|}\sum_{t'\in\mathcal{T}_t^j} V^{\pi_{t'}} \\
&\le \widehat{V}_{i_*}^t + \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}'_{i_*}}{|\mathcal{T}_t^{i_*}|^\beta} - \frac{1}{|\mathcal{T}_t^j|}\sum_{t'\in\mathcal{T}_t^j} g_{t'} + \frac{1}{|\mathcal{T}_t^j|}\sum_{t'\in\mathcal{T}_t^j} \epsilon_{t'} \\
&\le C_{\mathbb{Z}}\left(\mathcal{V}_{i_*}L^\alpha|\mathcal{T}_t^j|^{-(1-\kappa)\alpha} + \mathcal{V}'_{i_*}L^\beta|\mathcal{T}_t^j|^{-(1-\kappa)\beta} + H\sqrt{\frac{\log(1/\delta)}{|\mathcal{T}_t^j|}} + \frac{\mathcal{R}_{i_*}^{\Pi_{i_*}}}{\sqrt{|\mathcal{T}_t^j|}}\right) + \\
&\qquad \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}'_{i_*}}{|\mathcal{T}_t^{i_*}|^\beta} + H\sqrt{\frac{\log(1/\delta)}{|\mathcal{T}_t^j|}}
\end{aligned}$$

Again noting that $|\mathcal{T}_t^{i_*}| \ge \frac{t^{1-\kappa}}{8L} \ge \frac{|\mathcal{T}_t^j|^{1-\kappa}}{8L}$, the above can be simplified to

$$\begin{aligned}
\Delta_{j,i_*} &\le C'_{\mathbb{Z}}\cdot\left(2\mathcal{V}_{i_*}L^\alpha|\mathcal{T}_t^j|^{-(1-\kappa)\alpha} + 2\mathcal{V}'_{i_*}L^\beta|\mathcal{T}_t^j|^{-(1-\kappa)\beta} + \frac{2H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}}}{|\mathcal{T}_t^j|^{1/2}}\right) \\
&\le 6C'_{\mathbb{Z}}\cdot\max\left\{\frac{\mathcal{V}_{i_*}L^\alpha}{|\mathcal{T}_t^j|^{(1-\kappa)\alpha}},\ \frac{\mathcal{V}'_{i_*}L^\beta}{|\mathcal{T}_t^j|^{(1-\kappa)\beta}},\ \frac{H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}}}{|\mathcal{T}_t^j|^{1/2}}\right\}
\end{aligned}$$

where $C'_{\mathbb{Z}} = \max\{1, C_{\mathbb{Z}}\}$. Then, we can consider the three potential cases to upper bound $|\mathcal{T}_t^j|$. Depending on the maximal term, one of the three possible cases occurs:

$$|\mathcal{T}_t^j| \le \left(\frac{6C'_{\mathbb{Z}}\mathcal{V}_{i_*}L^\alpha}{\Delta_{j,i_*}}\right)^{\frac{1}{(1-\kappa)\alpha}}, \qquad |\mathcal{T}_t^j| \le \left(\frac{6C'_{\mathbb{Z}}\mathcal{V}'_{i_*}L^\beta}{\Delta_{j,i_*}}\right)^{\frac{1}{(1-\kappa)\beta}},$$
$$|\mathcal{T}_t^j| \le \left(\frac{6C'_{\mathbb{Z}}(H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}})}{\Delta_{j,i_*}}\right)^2$$

The regret during the misspecified phase becomes

$$\begin{aligned}
&\mathsf{Reg}_{\tau_{\min}(\delta):\tau_*} \\
&= O\Bigg(HLT^{1-\kappa} + Hi_* + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \\
&\quad \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}},\ \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}'^{\frac{1}{(1-\kappa)\beta}}_{i_*}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}},\ \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}} + H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\Bigg)
\end{aligned}$$

The total regret is

$$O\left(HL^{\frac{2}{1-\kappa}}\log^{\frac{1}{1-\kappa}}(1/\delta)+HLT^{1-\kappa}+Hi_*\right)+O\Bigg(\mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT}+ \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}},\ \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}'^{\frac{1}{(1-\kappa)\beta}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}},\ \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}}+H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\Bigg)$$

□

Consider again the implications of this bound in the contextual bandit setting. It is possible that to estimate an upper bound of V^* with rate $\widetilde{O}\left(\frac{d_j^{1/4}}{n^{1/2}}+\frac{1}{n^{1/4}}\right)$, where n is the number of samples and $j \geq i_*$ [27, 39]. However, this would only give a one-sided estimation error bound. If a two-sided guarantee of the same form were possible, we would have $\alpha = 1/2$, $\beta = 1/4$, and $\mathcal{V}_{i_*} = \widetilde{O}\left(d^{1/4}\right), \mathcal{V}'_{i_*} = \widetilde{O}\left(1\right)$. We now state the following immediate corollary in this setting with constant gaps under the hypothesis that such an estimator for this problem exists and is given.

Corollary 3.6.3. *For the linear contextual bandit problem under Assumption 3.6.2 with constant gaps $\{\Delta_{j,i_*}\}_{j<i_*}$, let $\alpha = 1/2$, $\beta = 1/4$, $\mathcal{V}_{i_*} = \widetilde{O}(d_{i_*}^{1/4})$ and $\mathcal{V}'_{i_*} = \widetilde{O}(1)$. Let the exploration parameter $\kappa = 1/2$. Then with probability at least $1-\delta'$, ECE in Algorithm 10 with the modified test (Equation 3.6) satisfies the regret bound*

$$\widetilde{O}\left(\sqrt{T}+\sqrt{d_{i_*}T}+\sum_{j<i_*}\max\left\{d_{i_*}\Delta_{j,i_*}^{-3},\ \Delta_{j,i_*}^{-7},\ d_{i_*}\Delta_{j,i_*}^{-1}\right\}\right)=\widetilde{O}\left(\sqrt{T}+\sqrt{d_{i_*}T}+d_{i_*}\right)$$

where $\widetilde{O}$ hides dependence on the number of models L, the number of actions $K = |\mathcal{U}|$, and log factors.

For constant gaps, the scalings in d and T are nearly same for this estimator and the gap estimator of the previous section. The main difference arises in the dependence on the gap, $O(\Delta_{\min}^{-5})$ in this case compared to $O\left(\Delta_{\min}^{-2}\right)$ in the previous case. In this case, it is clearly suboptimal.

3.7 Applications of ECE

Though Theorem 3.4.2 is stated generally for any RL algorithms with nominal anytime regret bounds, we can easily specialize it to several important problem settings without knowing the optimal model class *a priori*. In this section, we expand on the applications of Theorem 3.4.2 to paradigms of function approximation in RL.

Linear MDPs Consider the setting of [38] which we mentioned as an example in Section 3.2. In this setting, we assume access to a set of nested features $\phi_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_i}$ for $i \in [L]$ such that $d_i \leq d_{i+1}$ and the first d_i components of ϕ_{i+1} are the same as ϕ_i. These features generate linear model classes of the form

$$\mathcal{F}_i = \left\{(s,a) \mapsto \langle \phi_i(s,a), \theta \rangle \ : \ \theta \in \mathbb{R}^{d_i}\right\} \tag{3.7}$$

Nested-ness of the features ensures that $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ for all i. In accordance with the setting of [38], we assume that there exists some minimal i_* such that for any $\mathcal{F}_i$ with $i \geq i_*$ there exist $\mu(\cdot)$ and $\omega_{i,h} \in \mathbb{R}^{d_i}$ that predict exactly the transition probabilities P and reward r:

$$\begin{aligned} P(s'|s,u) &= \langle \phi_i(s,u), \mu_i(s') \rangle \\ r_h(s,u) &= \langle \phi_i(s,u), \omega_{i,h} \rangle \end{aligned} \tag{3.8}$$

Here, $\mu_i(\cdot)$ is a d_i-dimensional vector of measures on $\mathcal{S}$. Let $\{\mathcal{A}_i\}$ be instances of LSVI-UCB equipped with the doubling trick and model classes $\{\mathcal{F}_i\}$. We further assume that the features and parameters for each of the models with $i \geq i_*$ satisfies the regularity conditions of Assumption A of [38], i.e. bounded ℓ_2 norms, $r \in [0,1]$.

[38] guarantees that for $i \geq i_*$ and $t \in [T]$ with probability at least $1 - \delta_0$, $\mathsf{Reg}_t(\mathcal{A}_i) = O(\sqrt{d_i^3 H^4 t \cdot \log^2(d_i TH/\delta_0)})$. Adapting this to the framework of ECE, we let $\mathcal{R}_i = O\left(\sqrt{d_i^3 H^4 \cdot \log^2(d_i TH/\delta)}\right)$, which ensures $\mathcal{R}_i \leq \mathcal{R}_{i+1}$. A model selection corollary immediately follows from Theorem 3.4.2.

Corollary 3.7.1. *In the linear MDP setting of (3.8) with LSVI-UCB, ECE guarantees with probability at least* $1 - \delta'$

$$\mathsf{Reg}_T = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^4 \log^2(d_{i_*} LTH/\delta')} \cdot L^{5/6} T^{2/3}\right)$$

[68] consider a similar setting of linear MDPs where the transition dynamics P are linear. We again assume access to nested linear models but of the form

$$\mathcal{F}_i = \left\{(s,u,s') \mapsto \phi_i(s,u)^\top M \psi_i(s') \ : \ M \in \mathbb{R}^{d_i \times d_i'}\right\}$$

where $\{\phi_i\}_{i \in [L]}$ and $\{\psi_i\}_{i \in [L]}$ are nested features of dimension d_i and d_i' respectively. [68] assume that there is some minimal i_* such that for any $i \geq i_*$, there is $M \in \mathbb{R}^{d_i \times d_i'}$ such that

$$P(s'|s,u) = \phi_i(s,u)^\top M \psi_i(s') \tag{3.9}$$

for all $s, s' \in \mathcal{S}$, $u \in \mathcal{U}$. We further adhere to the regularity conditions of Assumption 2 of [68], who guarantee the MatrixRL $\mathcal{A}_i$ with model $\mathcal{F}_i$ has regret $\mathsf{Reg}_t(\mathcal{A}_i) = \widetilde{O}\left(\sqrt{d_i^3 H^5 t} \cdot \log(d_i TH/\delta_0)\right)$ with probability at least $1 - \delta_0$. Letting $\mathcal{R}_i = \widetilde{O}\left(\sqrt{d_i^3 H^5} \cdot \log(d_i TH/\delta)\right)$, we have the following model selection guarantee.

Corollary 3.7.2. *In the linear MDP setting of (3.9) with MatrixRL, ECE guarantees with probability at least* $1 - \delta'$

$$\mathsf{Reg}_T = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^5 \log^2(d_{i_*} LTH/\delta')} \cdot L^{5/6} T^{2/3}\right)$$

The final linear setting we consider is that of low inherent Bellman error studied by [69]. We let $\mathcal{F}_i$ be defined as it is in (3.7) and let $\mathcal{B} = \{\theta \in \mathbb{R}^{d_i} \ : \ \|\theta\| \leq D\}$ for some $D > 0$. Then assume there is a minimal i_* such that for any $i \geq i_*$ and $\theta_{h+1} \in \mathcal{B}$, there is θ_h such that

$$\langle \phi_i(s, u), \theta_h \rangle - \mathbf{B}_h Q_{h+1}(\theta_{h+1})(s, u) = 0$$

for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$, where $Q_h(\theta)$ is the linear action-value function parameterized by θ (with features ϕ_i) and $\mathbf{B}_h$ is the Bellman operator with reward r_h. In other words, this condition asserts that $\mathcal{F}_{i_*}$ has zero inherent Bellman error. Under the same regularity conditions, for $i \geq i_*$, [69] guarantees ELEANOR achieves $\mathsf{Reg}_t(\mathcal{A}_i) = \widetilde{O}\left(d_i \sqrt{H^4 t}\right)$ with probability at least $1 - \delta_0$. Letting $\mathcal{R}_i = \widetilde{O}\left(d_i \sqrt{H^4}\right)$, we have the following model selection guarantee.

Corollary 3.7.3. *In the inherent Bellman error setting with ELEANOR, ECE guarantees with probability at least* $1 - \delta'$

$$\mathsf{Reg}_T = \widetilde{O}\left(d_{i_*} \sqrt{H^4} \cdot L^{5/6} T^{2/3}\right)$$

where $\widetilde{O}$ hides polylog dependencies.

Low Bellman Rank Another class of algorithms using more general function approximation considers the setting of MDPs with low Bellman rank [36]. In this setting, a finite model class $\mathcal{F} : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ realizes $\mathcal{M}$ if there exists $f^* \in \mathcal{F}$ such that $Q_h^*(s, a) = f^*(s, a)$, where Q^* is the optimal action-value function for all $h \in [H]$. For any $f \in \mathcal{F}$, define π_f as the greedy policy with respect to f, and the Bellman error at $h \in [H]$ as

$$\mathcal{E}(f, \pi, h) := \mathbb{E}\left[f(s, \pi_f(s)) - r(s, \pi_f(s)) - f(s', \pi_f(s'))\right],$$

where the expectation is over s from the state distribution of π at h and $s' \sim P(\cdot|s, \pi_f(s))$. In this setting, it is assumed that there is a Bellman rank $M \ll |\mathcal{F}|$ such that for any $f, g \in \mathcal{F}$, we have $\mathcal{E}(f, \pi_g, h) = \langle \nu_h(g), \xi_h(f) \rangle$ for $\nu_h(g), \xi_h(f) \in \mathbb{R}^M$ and $\|\nu\|\|\xi\| \leq \zeta$. We assume access to a set of finite model classes $\{\mathcal{F}_i\}_{i\in[L]}$ such that there is at least one that realizes $\mathcal{M}$, and the complexity of $\mathcal{F}_i$ is a function of its cardinality $|\mathcal{F}_i|$ and induced Bellman rank M_i. We consider instances of the AVE algorithm $\{\mathcal{A}_i\}$ of [22] with the doubling trick, which has nominal regret $\widetilde{O}\left(\sqrt{M_i^2|\mathcal{U}|H^4 t \log^3|\mathcal{F}_i|}\right)$. Choose $\mathcal{R}_{\mathcal{F}_i} = \widetilde{O}\left(\sqrt{M_i^2|\mathcal{U}|H^4\log^3(|\mathcal{F}_i|)}\right)$ and let i_* be the smallest index that realizes $\mathcal{M}$. This yields the following corollary.

Corollary 3.7.4. *In the low Bellman rank setting with AVE, the model selection algorithm guarantees with probability at least* $1 - \delta'$

$$\mathsf{Reg}_T(\mathcal{A}) = \widetilde{O}\left(\sqrt{M_{i_*}^2|\mathcal{U}|H^4\log^3(|\mathcal{F}_{i_*}|)} \cdot L^{5/6}T^{2/3}\right).$$

3.8 Omitted Proofs of Section 3.4

In this section, we collect proofs for Theorem 3.4.2 that were omitted from the previous sections of the chapter.

Proof of Lemma 3.4.3

Here, we restate and prove Lemma 3.4.3.

Lemma 3.4.3. *The event* $E = \bigcap_{i\in\{1,2,3\}} E_i$ *holds with probability at least* $1 - 10LT^2\delta\log_2 T$.

Proof. We will show that each of the three events holds with high probability and the apply the union bound.

Corollary 3.9.1 of Section 3.9 shows event E_1 holds with probability at least $1 - 4LT^2\delta\log_2 T$.

For event E_2, i_* is the index of the algorithm that is $\mathcal{R}_{i_*}$-compatible and anytime. Let $\pi_{(k)}^{i_*}$ denote the policy played by $\mathcal{A}_{i_*}$ at the k^{th} call to i_*. For $K \in [T]$, these properties guarantee its regret bound holds, with probability at least $1-\delta$,

$$\sum_{k\in[K]} V^* - V^{\pi_{(k)}^{i_*}} \leq \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{K}$$

Taking the union bound over all $K \in [T]$ shows that event E_2 holds with probability at least $1 - T\delta$.

As in the previous case, we can view the process $\epsilon^i_{(1)}, \ldots, \epsilon^i_{(T)}$ as the pre-drawn differences between the observed and expected returns for the 1 through (at most) T times of playing model $\mathcal{A}_i$. Applying the Azuma-Hoeffding inequality with $|\epsilon^i_{(k)}| \leq H$ and taking the union bound over all $K \in [T]$,

$$|\sum_{k \in [K]} \epsilon^i_{(k)}| \leq H\sqrt{2K \log(2/\delta)}$$

with probability at least $1 - T\delta$. Taking the union bound over all models, event E_3 occurs with probability at least $1 - LT\delta$.

Taking these events together and $\delta' = 10LT^2\delta \log_2 T$, event E holds with probability at least $1 - \delta'$. □

Full Proof of Theorem 3.4.2

Here, we restate and complete the proof of Theorem 3.4.2.

Theorem 3.4.2. *Let the model exploration parameter $\kappa = 1/3$. Then, the model selection algorithm ECE satisfies the regret bound*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}(d_{i_*}, H, \log(LT/\delta')) \cdot i_*^{1/3} L^{1/2} T^{2/3}\right).$$

with probability at least $1 - \delta'$, where $\widetilde{O}$ hides logs and terms independent of T and $\mathcal{R}$.

Proof. Let $\tau_* := \tau_{i_*}$ denote the time that $\mathcal{A}_{i_*}$ is chosen as the candidate. Recall that $\delta = \frac{\delta'}{10LT^2 \log_2 T}$. The analysis can be divided into three phases when conditioned on the event E.

1. $t < \tau_{\min}(\delta)$: the test to determine switching to i_* is not valid yet.

2. $\tau_{\min}(\delta) < t \leq \tau_*$: the test is eligible but ECE is still switching among incompatible algorithms.

3. $t > \tau_*$: ECE has switched to $\mathcal{A}_{i_*}$.

Note that it is possible that $\tau_* \geq T$. That is, the algorithm only uses incompatible algorithms; however, we will show that this case still guarantees regret that adapts to the optimal algorithm i_*.

Case 1: Invalid Test We require $t \geq \tau_{\min}(\delta)$ in order for the condition in Lemma 3.4.1 to hold under E when $\widehat{\imath}_t = i_*$. Therefore, we can view this period $t < \tau_{\min}(\delta)$ as an unavoidable burn-in period. The regret during this interval can then be upper bounded in the worst case as

$$\mathsf{Reg}_{1:\tau_{\min}(\delta)-1} = \sum_{t=1}^{\tau_{\min}-1} V^* - V^{\pi_t} \leq H\tau_{\min} = O\left(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)\right)$$

Case 2: Misspecified Case In the second phase, the test is valid, but ECE is either utilizing algorithms below i_* or switching among them in the event the test fails. The regret can be decomposed across each set $\mathcal{T}^j_{\tau_*}$ of times playing $\mathcal{A}_j$ up to time τ_*:

$$\begin{aligned}
\mathsf{Reg}_{\tau_{\min}(\delta):\tau_*} &= \sum_{j\in[L]} \sum_{t\in\mathcal{T}^j_{\tau_*}} V^* - V^{\pi_t} \\
&\leq 4H(L-i_*)\tau_*^{1-\kappa} + \sum_{j<i_*} \sum_{t\in\mathcal{T}^j_{\tau_{j+1}}} V^* - V^{\pi_t} \\
&\leq 4H(L-i_*)\tau_*^{1-\kappa} + Hi_* + \sum_{j<i_*} \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t}
\end{aligned}$$

The second line follows from the fact that for $j > i_*$, algorithm j is not selected yet (if ever), so maximal regret is paid for those algorithms during exploration. Event E_1 upper bounds the number of times that can be in $\mathcal{T}^j_{\tau_*}$ at time τ_*, since the regret due to j is only due to exploration. Furthermore, for $j < i_*$, once j is rejected, it is never used for exploration again, so we can replace $\mathcal{T}^j_{\tau_*}$ with $\mathcal{T}^j_{\tau_{j+1}}$ for $j < i_*$. The third line is necessary as no guarantee is given during episodes when a test fails and there can be at most i_* failing tests since the condition in Lemma 3.4.1 is always true under event E.

Then, we focus on bounding the right-hand term. Fix $j < i_*$. Observe that for $t \in \mathsf{T}^j_{\tau_{j+1}-1}$ the tests succeed for all comparisons including with i_*:

$$\mathcal{G}_{\tau_{j+1}-1}(j, i) \leq \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta)$$

for all $i > j$. Therefore, since $i_* > j$, the definition of $\mathcal{G}$ can be used the bound the following:

$$
\begin{aligned}
\sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} &= \sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} (V^* - g_t) + \sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&\leq \frac{|\mathsf{T}^j_{\tau_{j+1}-1}|}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathsf{T}^{i_*}_{\tau_{j+1}-1}} (V^* - g_t) + \mathcal{W}(|\mathsf{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) + \sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&\leq \frac{|\mathsf{T}^j_{\tau_{j+1}-1}|}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathsf{T}^{i_*}_{\tau_{j+1}-1}} (V^* - V^{\pi_t}) + \mathcal{W}(|\mathsf{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) \\
&\quad + \sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} \epsilon_t + \frac{|\mathsf{T}^j_{\tau_{j+1}-1}|}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathsf{T}^{i_*}_{\tau_{j+1}-1}} \epsilon_t
\end{aligned}
$$

Now we can use the fact that E_2 and E_3 hold to bound the regret and estimation errors:

$$
\begin{aligned}
\sum_{t\in\mathsf{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} \leq O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{\frac{|\mathsf{T}^j_{\tau_{j+1}-1}|^2}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|}}\right) + \mathcal{W}(|\mathsf{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) \\
+ O\left(H\sqrt{|\mathsf{T}^j_{\tau_{j+1}-1}| \cdot \log(1/\delta)}\right) + O\left(H\sqrt{\frac{|\mathsf{T}^j_{\tau_{j+1}-1}|^2}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|} \cdot \log(1/\delta)}\right)
\end{aligned}
\tag{3.10}
$$

Using E_1 and the fact that $\tau_{\min}(\delta) \leq \tau_{j+1} - 1 \leq \tau_*$, we have that

$$
|\mathcal{T}^{i_*}_{\tau_{j+1}-1}| \geq \frac{(\tau_{j+1}-1)^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|^{1-\kappa}}{8L}.
$$

Then the terms in (3.10) that contain $|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|$ in the denominator can be upper bounded:

$$
\begin{aligned}
O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{\frac{|\mathsf{T}^j_{\tau_{j+1}-1}|^2}{|\mathsf{T}^{i_*}_{\tau_{j+1}-1}|}}\right) \leq O\left(L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot |\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}}\right) \\
O\left(H\sqrt{\frac{|\mathsf{T}^j_{\tau_{j+1}-1}|^2}{|\mathsf{T}^j_{\tau_{j+1}-1}|} \cdot \log(1/\delta)}\right) \leq O\left(HL^{1/2}|\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)
\end{aligned}
$$

The bound then becomes

$$\sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} \leq O\left(L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot |\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}}\right) + \mathcal{W}(|\mathsf{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) + O\left(H|\mathsf{T}^j_{\tau_{j+1}-1}|^{1/2} \cdot \log^{1/2}(1/\delta)\right) + O\left(HL^{1/2}|\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)$$

Since $\mathcal{R}_j \leq \mathcal{R}_{i_*}$, the regret for j in this case is

$$\sum_{t \in \mathsf{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} \leq O\left(L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot |\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} + HL^{1/2}|\mathsf{T}^j_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)$$

Observe that $\sum_{j<i_*} |\mathcal{T}^j_{\tau_{j+1}-1}| \leq T$ and the right-hand side is a sum of concave functions of each $|\mathcal{T}^j_{\tau_{j+1}-1}|$. Using Jensen's inequality with the uniform distribution over $|\mathcal{T}^j_{\tau_{j+1}-1}|$ for $j < i_*$ and then upper bounding by T yields the bound:

$$\mathsf{Reg}_{\tau_{\min}(\delta):\tau_*} \leq O\left(HLT^{1-\kappa} + Hi_* + \left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) + H\log^{1/2}(1/\delta)\right) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} \cdot T^{\frac{1+\kappa}{2}}\right)$$

Case 3: Selecting $\mathcal{A}_{i_*}$ Starting at $\tau_* + 1$, $\mathcal{A}_{i_*}$ is selected. Note that the condition in Lemma 3.4.1 holds under event E, so ECE will never reject i_*. Then

$$\begin{aligned}
\mathsf{Reg}_{\tau_*+1:T} &\leq \sum_{j \in [i_*+1, L]} H|\mathcal{T}^j_T| + \sum_{t \in \mathcal{T}^{i_*}_T} V^* - V^{\pi_t} \\
&\leq \sum_{j \in [i_*+1, L]} H|\mathcal{T}^j_T| + O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta) \cdot \sqrt{T}\right) \\
&\leq O\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta) \cdot \sqrt{T}\right)
\end{aligned}$$

Adding the terms from these three phases gives the final bound:

$$\mathsf{Reg}_T = O\left(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta) + HLT^{1-\kappa} + Hi_* + \left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) + H\log^{1/2}(1/\delta)\right) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} \cdot T^{\frac{1+\kappa}{2}}\right)$$

Then we choose $\kappa = 1/3$ to recover the statement in the theorem. □

Proof of Theorem 3.5.2

Here, we restate an prove Theorem 3.5.2.

Theorem 3.5.2. *For a given $\mathcal{M}$, let $(\mathcal{A}_i, \mathcal{F}_i)$ be $\mathcal{R}_i^{\Pi_i}$-compatible with respect to Π_i for all $i \in [L]$. Then, with probability at least $1-\delta'$, ECE with $\kappa = 1/3$ satisfies the regret bound with respect to policy class Π_{i_*}:*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^{3/2}(\mathcal{R}_{i_*}^{\Pi_{i_*}})^3 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-2}\right)$$

If $\kappa = 1/2$, then it satisfies

$$\widetilde{O}\left(HL\sqrt{T} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^2(\mathcal{R}_{i_*}^{\Pi_{i_*}})^4 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-3}\right)$$

Proof. First we will show that the sufficient events to prove this result occur with high probability. While the other events remain the same. we must modify event E_2 from Lemma 3.4.3 slightly because we are interested in the case when all algorithms are compatible with respect to their own policy classes. Let E_2' denote the following event: for all $t \in [T]$ and $i \in [L]$,

$$\textstyle\sum_{t' \in \mathcal{T}_t^i} V_i^* - V^{\pi_{t'}} \leq \mathcal{R}_i^{\Pi_i}(d_i, H, \log(T/\delta))\sqrt{|\mathcal{T}_t^i|}$$

As in Lemma 3.4.3, this almost follows from Definition 3.5.1; however, we also union bound over all algorithms. Thus E_2' occurs with probability at least $1 - LT\delta$. Let $E_1' = E_1$ and $E_3' = E_3$. Then $E' = \bigcap_{i \in 1,2,3} E_i'$ occurs with probability at least $1 - 10LT^2\delta \log_2 T$, as before.

Recall that $i_* = \min B_*$ where B_* is the set of indices that achieve maximal value, $\operatorname{argmax}_i V_i^*$. For shorthand, we will let $\mathcal{R}_j := \mathcal{R}_j^{\Pi_j}(d_j, H, \log(T/\delta))$. We now verify that the statistical test will not fail once ECE reaches some $i_* \in B_*$. This is nearly identical to Lemma 3.4.1, but we must verify it with respect to values that are not the optimal value.

Lemma 3.8.1. *Let $(\mathcal{A}_i, \mathcal{F}_i)$ be an $\mathcal{R}_i^{\Pi_i}$-compatible algorithm with respect to Π_i for all $i \in [L]$ and let $i_* = \min B_*$. Given that event E' holds and $t \geq \tau_{\min}(\delta)$, then, for all $j \in [i_* + 1, L]$, it holds that $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$.*

Proof. From the definition of $\mathcal{G}$,

$$\begin{aligned}\mathcal{G}_t(i_*, j) &= \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} g_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} g_{t'} = \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} (V^{\pi_{t'}} + \epsilon_{t'}) - \sum_{t' \in \mathcal{T}_t^{i_*}} (V^{\pi_{t'}} + \epsilon_{t'}) \\ &\leq \sum_{t' \in \mathcal{T}_t^{i_*}} \left(V_{i_*}^* - V^{\pi_{t'}}\right) + \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} \epsilon_{t'}\end{aligned}$$

where the last step uses the fact that $V^*_{i_*} = \max_i V^*_i$. Since $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}^{\Pi_{i_*}}_{i_*}$ compatible, the remainder of the proof is identical to that of Lemma 3.4.1 by applying the conditions in E'. □

As before, in the full proof we handle three cases: (1) before the test is valid, (2) while $i < i_*$ is chosen, (3) after i_* is chosen. In the first case, we again pay the burn-in period regret of $\mathsf{Reg}_{1:\tau_{\min}(\delta)-1} = O(H\tau_{\min}(\delta))$. In the third, we showed that the test will never fail once $\widehat{\imath}_t = i_*$. Therefore, $\mathsf{Reg}_{\tau_*:T} = O\left(HLT^{1-\kappa} + \mathcal{R}_{i_*} \cdot \sqrt{T}\right)$.

To bound the regret during the misspecified phase, we construct an upper bound on the number of times $\mathcal{A}_j$ can be played for $j < i_*$. Let t be a time such that $\widehat{\imath}_t = j < i_*$ and the test succeeds. First, we bound the size of the gaps.

Note that by definition $V^*_j \geq \frac{1}{|\mathcal{T}^j_t|} \sum_{t' \in \mathcal{T}^j_t} V^{\pi_{t'}}$ and event E' ensures that $V^*_{i_*} \leq \frac{\mathcal{R}_{i_*}}{|\mathcal{T}^{i_*}_t|^{1/2}} + \frac{1}{|\mathcal{T}^{i_*}_t|} \sum_{t' \in \mathcal{T}^{i_*}_t} V^{\pi_{t'}}$. Then,

$$\begin{aligned}
\Delta_{j,i_*} &= V^*_{i_*} - V^*_j \\
&\leq \frac{1}{|\mathcal{T}^{i_*}_t|} \sum_{t' \in \mathcal{T}^{i_*}_t} V^{\pi_{t'}} + \frac{\mathcal{R}_{i_*}}{|\mathcal{T}^{i_*}_t|^{1/2}} - \frac{1}{|\mathcal{T}^j_t|} \sum_{t' \in \mathcal{T}^j_t} V^{\pi_{t'}} \\
&= \frac{\mathcal{R}_{i_*}}{|\mathcal{T}^{i_*}_t|^{1/2}} + \frac{1}{|\mathcal{T}^{i_*}_t|} \sum_{t' \in \mathcal{T}^{i_*}_t} (g_{t'} - \epsilon_{t'}) - \frac{1}{|\mathcal{T}^j_t|} \sum_{t' \in \mathcal{T}^j_t} (g_{t'} - \epsilon_{t'}) \\
&\leq \frac{\mathcal{W}(|\mathcal{T}^j_t|, \mathcal{R}_j, d_j, \delta)}{|\mathcal{T}^j_t|} + \frac{\mathcal{R}_{i_*}}{|\mathcal{T}^{i_*}_t|^{1/2}} - \frac{1}{|\mathcal{T}^{i_*}_t|} \sum_{t' \in \mathcal{T}^{i_*}_t} \epsilon_{t'} + \frac{1}{|\mathcal{T}^j_t|} \sum_{t' \in \mathcal{T}^j_t} \epsilon_{t'}
\end{aligned}$$

And therefore,

$$\begin{aligned}
\Delta_{j,i_*} \leq C_{\mathcal{W}} \cdot &\left(\frac{\mathcal{R}_j}{|\mathcal{T}^j_t|^{1/2}} + H\sqrt{\frac{16L\log(2/\delta)}{|\mathcal{T}^j_t|^{1-\kappa}}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}^j_t|}} \right) \\
&+ \frac{\mathcal{R}_{i_*}}{|\mathcal{T}^{i_*}_t|^{1/2}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}^{i_*}_t|}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}^j_t|}}
\end{aligned}$$

where we have applied the definition of $\mathcal{W}$ and event E_3 to bound the noise of the returns. Let $C'_{\mathcal{W}} = \max\{1, C_{\mathcal{W}}\}$. Since i_* has not been selected yet $|\mathcal{T}^{i_*}_t| \geq \frac{t^{1-\kappa}}{8L} \geq$

$\frac{|\mathcal{T}_t^j|^{1-\kappa}}{8L}$. Then, since $\mathcal{R}_j \le \mathcal{R}_{i_*}$,

$$\Delta_{j,i_*} \le C'_{\mathcal{W}} \cdot \left(\frac{2\sqrt{8L}\mathcal{R}_{i_*}}{|\mathcal{T}_t^j|^{\frac{1-\kappa}{2}}} + H \frac{2\sqrt{16L\log(2/\delta)}}{|\mathcal{T}_t^j|^{\frac{1-\kappa}{2}}} \right)$$

Rearranging gives

$$|\mathcal{T}_t^j| = O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H \log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{2}{1-\kappa}}} \right)$$

Now this bound can be used to bounding the regret with dependence on the gap. The regret during this phase is again

$$\begin{aligned} \mathsf{Reg}_{\tau_{\min}(\delta):\tau_*} &\le H(L-i_*)\tau_*^{1-\kappa} + \sum_{j<i_*} \sum_{t \in \mathcal{T}^j_{\tau_{j+1}}} V^*_{i_*} - V^{\pi_t} \\ &\le H(L-i_*)\tau_*^{1-\kappa} + Hi_* + \sum_{j<i_*} \sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} V^*_{i_*} - V^{\pi_t} \end{aligned}$$

As in the proof of Theorem 3.4.2, we focus on bounding the right-hand term. For a fixed $j < i_*$, at time $\tau_{j+1}-1$ we have that the test succeeds so $\mathcal{G}_{\tau_{j+1}-1}(j, i_*) \le \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$. Then, applying the bound on the number of times j can be played,

$$\begin{aligned} \sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} V^*_{i_*} - V^{\pi_t} &\le \Delta_{j,i_*} |\mathcal{T}^j_{\tau_{j+1}-1}| + \mathcal{R}_j \cdot \sqrt{|\mathcal{T}^j_{\tau_{j+1}-1}|} \\ &\le O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H \log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} + \right. \\ &\qquad \left. \frac{\mathcal{R}_{i_*} L^{\frac{1}{2(1-\kappa)}} \left(\mathcal{R}_{i_*} + H \log^{1/2}(1/\delta) \right)^{\frac{1}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1}{1-\kappa}}} \right) \\ &= O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H \log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} \right) \end{aligned}$$

Therefore, the regret in this phase can be upper bounded by

$$\mathsf{Reg}_{\tau_{\min}(\delta):\tau_*} \le O\left(H(L-i_*)T^{1-\kappa} + Hi_* + L^{\frac{1}{1-\kappa}}\left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta)\right)^{\frac{2}{1-\kappa}}\sum_{j<i_*}\frac{1}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}}\right)$$

Combining these three phases, the total regret is

$$O\Big(HL^{\frac{2}{1-\kappa}}\log^{\frac{1}{1-\kappa}}(1/\delta) + HLT^{1-\kappa} + Hi_* +$$
$$L^{\frac{1}{1-\kappa}}\left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta)\right)^{\frac{2}{1-\kappa}}\sum_{j<i_*}\frac{1}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} + \mathcal{R}_{i_*}\sqrt{T}\Big)$$

Choosing either $\kappa = 1/3$ or $\kappa = 1/2$ gives us the statements of Theorem 3.5.2. This completes the proof. □

3.9 Anciliary Technical Results

In this section, we use a Freedman inequality to lower and upper bound with high probability the number of times a particular algorithm is played both during exploration and while it is chosen by the meta-algorithm (Lemma 3.4.3). First, we state a variant of the Freedman inequality from [11].

Lemma 3.9.1 (Lemma 2, [11])**.** *Suppose $X_1, \cdots, X_T$ is a martingale difference sequence with $|X_s| \le b$. We define*

$$\mathrm{Var}_s X_s = \mathbf{Var}(X_s | X_1, \cdots, X_{s-1})$$

Further, let $V_T = \sum_{s=1}^{T}\mathrm{Var}_s X_s$ be the sum of conditional variances of $X_s's$, and $\sigma_T = \sqrt{V_T}$. Then we have, for any choice of $\delta < 1/e$ and $T \ge 4$:

$$\mathbb{P}\left(\sum_{s=1}^{T} X_s > 2\max(2\sigma_T, b\sqrt{\ln(1/\delta)})\sqrt{\ln(1/\delta)}\right) \le \log_2(T)\delta \qquad (3.11)$$

Recall that B_s denotes the indices of algorithms that have not been selected by time s. Note that $|B_s| \le L$. For all $i \in [L]$ and $t \in [T]$, define the event

$$\mathcal{E}_{i,t} := \begin{cases} ||\mathcal{T}_t^i| - \sum_{s\in[t]}\frac{1}{|B_s|s^\kappa}| \le 4\sqrt{\sum_{s\in[t]}\frac{1}{s^\kappa}\log(1/\delta)} & \tau_i \ge t \\ ||\mathcal{T}_t^i| - \sum_{s\in[\tau_i]}\frac{1}{|B_s|s^\kappa} - \sum_{s\in[\tau_i+1,t]}\left(1-\frac{1}{s^\kappa}\right)| \le 4\sqrt{\sum_{s\in[t]}\frac{1}{s^\kappa}\log(1/\delta)} & \tau_i < t \end{cases}$$

Lemma 3.9.2. *The event* $\mathcal{E} = \cap_{i\in[L],t\in[T]}\mathcal{E}_{i,t}$ *holds with probability at least* $1 - 4LT^2\delta\log_2 T$

Proof. Define

$$S_i(t,t') = \sum_{s\in[t']} Y_{s,i} + \sum_{s\in[t'+1,t]} \overline{Y}_{s,i}$$

where $Y_{s,i} \sim \mathrm{Ber}\left(\frac{1}{s^\kappa|B_s|}\right)$ and $\overline{Y}_{s,i} \sim \mathrm{Ber}\left(1-\frac{1}{s^\kappa}\right)$. Then define

$$Z_i(t,t') := \sum_{s\in[t]} \mathbf{1}_{s\le t'} \cdot \left(Y_{s,i} - \frac{1}{|B_s|s^\kappa}\right) + \mathbf{1}_{s>t'}\left(\overline{Y}_{s,i} - \left(1-\frac{1}{s^\kappa}\right)\right)$$
$$V_i(t,t') := \sum_{s\in[t]} \mathbf{Var}_s\left(\mathbf{1}_{t\le t'} \cdot \left(Y_{s,i} - \frac{1}{|B_s|s^\kappa}\right) + \mathbf{1}_{t>t'} \cdot \left(\overline{Y}_{s,i} - \left(1-\frac{1}{s^\kappa}\right)\right)\right)$$

where $\mathbf{Var}_s$ denotes the conditional variance up to time s. By definition, $\{Z_i(t,t')\}_{t\ge1}$ is a martingale sequence and $V_i(t,t') \le \sum_{s\in[t]} \frac{1}{s^\kappa}$. By the Freedman inequality from Lemma 3.9.1,

$$\Pr\left(|Z_i(t,t')| \ge 4\sqrt{\sum_{s\in[t]}\frac{1}{s^\kappa}\cdot\log(1/\delta)} + 4\log(1/\delta)\right) \le 2\delta\log_2 T$$

Let this event be denoted by $\overline{\mathcal{E}}_i(t,t')$ for each $i\in[L]$ and $t,t'\in[T]$. Then, by the union bound, the event $\bigcup_{i,t,t'}\overline{\mathcal{E}}_i(t,t')$ holds with probability at most $4LT^2\delta\log_2 T$. Therefore, $\bigcap_{t,t'\ge1}\mathcal{E}_i(t,t')$ holds with probability at least $1-4LT^2\delta\log_2 T$, and this event implies for all $i\in[L]$ and $t\in[T]$, if $t>\tau_i$, then

$$||\mathcal{T}_t^i| - \sum_{s\in[\tau_i]}\frac{1}{|B_s|s^\kappa} - \sum_{s\in[\tau_i+1,t]}\left(1-\frac{1}{s^\kappa}\right)| \le 4\sqrt{\sum_{s\in[t]}\frac{1}{s^\kappa}\log(1/\delta)} + 4\log(1/\delta)$$

and if $\tau \le \tau_i$, then

$$||\mathcal{T}_t^i| - \sum_{s\in[t]}\frac{1}{|B_s|s^\kappa}| \le 4\sqrt{\sum_{s\in[t]}\frac{1}{s^\kappa}\log(1/\delta)} + 4\log(1/\delta)$$

□

Corollary 3.9.1. *With probability at least* $1-4LT^2\delta\log_2 T$, *for all* $i\in[L]$ *and* $t\in[T]$ *such that* $t\ge\tau_{\min}(\delta)$, *the following is true:*

1. *If* $t \le \tau_i$*, then* $\frac{t^{1-\kappa}}{8L} \le |\mathcal{T}_t^i| \le 4t^{1-\kappa}$.
2. *If* $t > \tau_i$*, then* $|\mathcal{T}_t^i| \le t - \tau_i + 4t^{1-\kappa}$.

Proof. Note that when $t \le \tau_i$, it is also the case that $|B_s| \ge 1$ for all $s \le t$. We condition on the event $\mathcal{E}$ from above, which occurs with probability at least $1 - 4LT^2\delta \log_2 T$. Given this event, it follows that if $t \le \tau_i$, then

$$\begin{aligned}
|\mathcal{T}_t^i| &\ge \sum_{s\in[t]} \frac{1}{s^\kappa |B_s|} - 4\sqrt{\sum_{s\in[t]} \frac{1}{s^\kappa} \log(1/\delta)} - 4\log(1/\delta) \\
&\ge \frac{1}{2L} \sum_{s\in[t]} \frac{1}{s^\kappa} - 32L\log(1/\delta) \\
&\ge \frac{1}{2L}\left(t^{1-\kappa} - 2\right) - 32L\log(1/\delta) \\
&\ge \frac{t^{1-\kappa}}{4L} - 32L\log(1/\delta) \\
&\ge \frac{t^{1-\kappa}}{8L}
\end{aligned}$$

The second inequality uses the AM-GM inequality and that $|B_s| \le L$, which implies

$$\sqrt{\sum_{s\in[t]} \frac{1}{Ls^\kappa} \cdot 16L\log(1/\delta)} \le \frac{1}{2L} \sum_{s\in[t]} \frac{1}{s^\kappa} + 8L\log(1/\delta)$$

The third applies the integral approximation of the sum. The last two follow from the condition that $t \ge \tau_{\min}(\delta) = C_{\min} \cdot L^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)$ for a large enough constant $C_{\min} > 0$. The other side follows similarly with

$$|\mathcal{T}_t^i| \le 3t^{1-\kappa} + 32\log(1/\delta) \le 4t^{1-\kappa}$$

when $t \ge (32\log(1/\delta))^{\frac{1}{1-\kappa}}$. Similarly, for $t > \tau_i$, event $\mathcal{E}$ guarantees

$$\begin{aligned}
|\mathcal{T}_t^i| &\le \sum_{s\in[\tau_i]} \frac{1}{s^\kappa |B_s|} + \sum_{s\in[\tau_i+1,t]} \left(1 - \frac{1}{s^\kappa}\right) + 4\sqrt{\sum_{s\in[t]} \frac{1}{s^\kappa} \log(1/\delta)} + 4\log(1/\delta) \\
&\le t - \tau_i + 32\log(1/\delta) + \frac{3}{2} \sum_{s\in[\tau_i]} \frac{1}{s^\kappa} \\
&\le t - \tau_i + 32\log(1/\delta) + 3t^{1-\kappa} \\
&\le t - \tau_i + 4t^{1-\kappa}
\end{aligned}$$

when $t \ge \tau_{\min}(\delta)$. □

Chapter 4

Simple Regret Balancing

4.1 Introduction

We study the problem of choosing among a set of learning algorithms in sequential decision-making problems with partial feedback. Learning algorithms are designed to perform well when certain favorable conditions are satisfied. However, the learning agent might not know in advance which algorithm is more appropriate for the current problem that the agent is facing.

As an example, consider the application of stochastic bandit algorithms in personalization problems, where in each round a user visits the website and the learning algorithm should present the item that is most likely to receive a click or be purchased. When contextual information (such as location, browser type, etc) is available, we might decide to learn a click model given the user context. If the context is not predictive of the user behavior, using a simpler non-contextual bandit algorithm might lead to a better performance. As another example, consider the problem of tuning the exploration rate of bandit algorithms. Typically, the exploration rate in an ϵ-greedy algorithm has the form of c/t, where t is time and the optimal value of constant c depends on unknown quantities related to reward vector. The decision rule of the UCB algorithm also involves an exploration bonus [9]. Choosing values smaller than the theoretically suggested value can lead to better performance in practice if the theoretical value is too conservative. However, if the exploration bonus is too small, the regret can be linear. It is desirable to have a model selection strategy that finds a near-optimal parameter value in an online fashion.

A model selection strategy can also be useful in finding effective reinforcement learning methods. There has been a great number of reinforcement learning algorithms proposed and studied in the literature [62, 63]. In some specialized domains, we might have a reasonable idea of the type of solution that can perform well. In general, however, designing a reinforcement learning solution can be a daunting task as the solution often involves many components. In fact, in some problems it is not even clear if we should use a reinforcement learning solution or a simpler contextual bandit solution. For example, bandit algorithms are used in many personalization and recommendation problems, although the decisions of the learning system can potentially change the future traffic and inherently we face a Markov decision process. In such problems, the available data might not be enough to solve the problem using an RL algorithm and a simpler bandit solution might be preferable. The complexity of the RL problem is often not known in advance and we would like to adapt to the complexity of the problem in an online fashion.

While model selection is a well-studied topic in supervised learning, results in the bandit and RL setting are scarce. [49] propose a method for the model selection problem based on EXP4 with additional uniform exploration. [5] obtain improved results by an online mirror descent method with a carefully selected mirror map. The algorithm is called CORRAL, and under a stability condition, it is shown to enjoy strong regret guarantees. Many bandit algorithms that are designed for stochastic environments (such as UCB, Thompson sampling, etc) do not satisfy the stability condition and thus cannot be directly used as base algorithms for CORRAL. Although it might be possible to make these algorithm stable by proper modifications, the process can be tedious. To overcome this issue, [54] propose a generic smoothing procedure that transforms nearly any stochastic algorithm into one that is stable. Results of [5] and [54] require the knowledge of the optimal base regret. [27] study bandit model selection among linear bandit algorithms when the dimensionality of the underlying linear reward model, and thus the optimal base regret, is not known. A related problem is studied by [18].

In this chapter, we propose a model selection method for bandit and RL problems in stochastic environments. We call our method "regret balancing" because it maintains regret estimates of base algorithms and tries to keep the *empirical regret* of all algorithms roughly the same. All algorithms maintain an empirical estimator of their regret computed as the difference of an optimistic estimator of the optimal policy's reward and the algorithm's collected reward. The method achieves regret balancing by playing the base algorithm with the smallest empirical regret. An algorithm can have small empirical regret for two reasons: either it chooses good actions, or it has not been played enough. By playing the algorithm with the smallest empirical regret, the model selection procedure finds an effective trade-off between exploration and exploitation.

The proposed approach has several notable properties. First, no stability condition is needed and any base algorithm without any modifications can be used. Note that when applied to stochastic bandit algorithms, [5] and [54] modify the base algorithms to ensure certain stability conditions. Second, our approach is intuitive and almost as simple as a UCB rule. By contrast, many existing model selection approaches have a complicated form. Finally, the approach can be readily applied to reinforcement learning problems.

The proposed approach, similar to a number of existing solutions, requires the knowledge of the regret of the optimal base algorithm. We show that, in general, any model selection strategy that achieves a near-optimal regret requires either the optimal base regret or direct sampling from the arms. We show that by adding a forced exploration scheme, and hence direct access to the arms, the regret balancing strategy can achieve near-optimal regret in a class of problems without the knowledge of the optimal base regret. Further, we show a class of problems where any near-optimal model selection procedure is indeed implementing a regret balancing method, possibly implicitly.

As we will show, the regret of our model selection strategy is $\Omega(T)$, where T is time horizon. This regret is minimax optimal, given the existing lower bound for the model selection problem that scales as $\Omega(\sqrt{T})$ [54]; Even if it is known that a base algorithm has logarithmic regret, the fast logarithmic regret cannot be preserved in general.

We show a number of applications of the proposed approach for model selection. We show how a near-optimal regret can be achieved in the class of ϵ-greedy algorithms without any prior knowledge of the reward function. We also show how the proposed approach can be used for representation learning in bandit problems. Further, we show a model selection strategy to choose among reinforcement learning algorithms. As a consequence for reinforcement learning, if a set of feature maps are given and the value functions are known to be linear in a feature map belonging to this set, we can use the regret balancing strategy to achieve a regret that is near-optimal up to a constant factor. Finally, the proposed regret balancing strategy can also be used as a bandit algorithm. We show how the approach is implemented as an algorithm for linear stochastic bandits.

4.2 Problem Statement

For an integer A, we use $[A]$ to denote the set $\{1, 2, \ldots, A\}$. A contextual bandit problem is a sequential game between a learner and an environment. We consider a set of learners $[M]$. The game is specified by a context space S, an action set $[K]$ of size K, a reward function $r : S \times [K] \to [0, 1]$, and a time horizon T. In round $t \in [T]$, the learner $i \in [M]$ observes the context $s_t \in S$ and chooses an action $a_t \in [K]$ from the action set. Then the learner observes a reward $r_t = r(s_t, a_t) + \eta_t$, where for a positive constant σ, η_t is a σ-sub-Gaussian random variable, meaning that for any $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \eta_t}] \le e^{\lambda^2 \sigma^2/2}$. In the special case of linear contextual bandits [41], we are given a feature map $\phi : S \times [K] \to \mathbb{R}^d$ such that $r(s, a) = \phi(s, a)^\top \theta_*$ for an unknown vector $\theta_* \in \mathbb{R}^d$. Let $\mu_{*,t} = \mathbb{E}(\max_a r(s_t, a))$ be the expected reward of the optimal action at time t, where expectation is taken with respect to the randomization in s_t and η_t. The goal is to have small regret, defined as $C_{i,T} = \sum_{t=1}^{T}(\mu_{*,t} - r_t)$. If $\{s_t\}_{t=1}^{T}$ is an IID sequence, then $\mu_{*,t}$ is the same constant for all rounds and we use μ_* to denote this value. The game is challenging as the reward function is not known in advance. If S contains only one element, then the problem reduces to the multi-armed bandit problem. If an action influences the distribution of the next context, then the problem is a Markov decision process (MDP) and it is more suitable to define regret with respect to the *policy* that has the highest total (or stationary) reward (See Section 4.4 for more details).

A bandit model selection problem is specified by a class of bandit problems and a set of bandit algorithms. Let M be the number of bandit algorithms (called base algorithms in what follows). As defined above, $C_{i,T}$ is the regret of the ith base in the underlying bandit problem if the base algorithm is executed alone. In a bandit model selection problem, the decision making is a two step process. In round t, the learner chooses base i_t from the set of M bandit algorithms, the base observes the context s_t and selects an action a_t from the set of K actions, and the reward r_t of the action is revealed to the learner. Then the internal state of the base i_t is updated using reward r_t. The regret of the overall model selection strategy is defined with respect to $\mu_{*,t}$:

$$\text{Regret}_T = \sum_{t=1}^{T}(\mu_{*,t} - r_t) \ .$$

Let i_* be the optimal base with the smallest regret if it is played in all rounds, $i_* = \arg\min_i C_{i,T}$. We would like to ensure that $\text{Regret}_T = O(C_{i_*,T})$. A reinforcement learning model selection problem is defined similarly (See Section 4.4 for more details).

4.3 Regret Balancing

At a high level, the main idea is to estimate the empirical regret of the base algorithms during the rounds that the algorithms are played, and ensure that all base algorithms suffer roughly the same empirical regret. This simple idea ensures a good trade-off between exploration and exploitation: if a base algorithm is played only for a small number of rounds, or if it plays good actions, then its empirical regret will be small and will be chosen by the model selection procedure.

Bandit Model Selection

In this section, we present the regret balancing model selection method. Consider a bandit model selection problem in a stochastic environment. Let $N_{i,t}$ be the number of rounds that base i is played up to but not including round t, and let $R_{i,t}$ be the total reward of this base during these $N_{i,t}$ rounds. With an abuse of notation we also use $N_{i,t}$ to denote the set of rounds that base i is selected. Let $S_{i,t}$ be all data in the rounds that base i is played, $S_{i,t} = \{(s_t, a_t, r_t) : t \in N_{i,t}\}$. Let $\mathbb{H}$ be the space of all such histories for all i and t. We use $R_{*,t}$, $N_{*,t}$, and $S_{*,t}$ to denote the quantities related to the optimal base, which was defined earlier in the problem definition. Regret of base i during the $N_{i,t}$ rounds is $G_{i,t} = \sum_{\tau \in N_{i,t}} \mu_{*,\tau} - R_{i,t}$. We assume that a high probability (possibly data-dependent) upper bound on the regret of the optimal base algorithm is known: a function $U : \mathbb{R} \times \mathbb{H} \to \mathbb{R}$ is given so that for any $\delta \in (0,1)$, with probability at least $1-\delta$, $G_{i_*,t} \le U(\delta, S_{*,t})$ for any t.[1] For example, for the UCB algorithm we have $U(\delta, S_{*,t}) = \widetilde{O}(\sqrt{Kt\log(1/\delta)})$,[2] and for the OFUL algorithm we have $U(\delta, S_{*,t}) = \widetilde{O}(\log(\det(V_t)/\delta)\sqrt{t})$, where V_t is an empirical covariance matrix [1]. Given that $G_{i_*,t}$ is defined with respect to the realized rewards $R_{i_*,t}$, the regret upper bound U should be at least of order $\Omega(\sqrt{t})$.

Next, we describe the model selection strategy. In round t, let j_t be the *optimistic* base and b_t be the optimistic value,

$$j_t = \arg\max_{i\in[M]} \frac{R_{i,t}}{N_{i,t}} + \frac{U(\delta, S_{i,t})}{N_{i,t}}\,, \qquad b_t = \frac{R_{j_t,t}}{N_{j_t,t}} + \frac{U(\delta, S_{j_t,t})}{N_{j_t,t}}\,. \tag{4.1}$$

Variable b_t estimates the value of the best action. Define the empirical regret of base i by

$$\widehat{G}_{i,t} = N_{i,t} b_t - R_{i,t}\,.$$

Recall the true regret defined by $G_{i,t} = \sum_{\tau \in N_{i,t}} \mu_{*,\tau} - R_{i,t}$. Notice that we have $N_{j_t,t} b_t - R_{j_t,t} = U(\delta, S_{j_t,t})$, i.e. b_t is chosen so that the empirical regret of the optimistic base scales as the target regret of the optimal base. Throughout the game, we play bases to ensure that the empirical regrets of all bases are roughly the same. To be more precise, in time t, we choose the base with the smallest empirical regret:

$$i_t = \arg\min_{i\in[M]} \widehat{G}_{i,t}\,. \tag{4.2}$$

This choice will most likely increase the empirical regret of base i_t. Next theorem shows the model selection guarantee of the regret balancing strategy.

[1] We can use different probabilistic guarantees here, and any form used here will also appear in Theorem 4.3.1.

[2] We use $\widetilde{O}$ notation to hide polylogarithmic terms.

Theorem 4.3.1. *If $\mu_{*,t} = \mu_*$ for a constant μ_* regardless of time t, and if with probability at least $1-\delta$, $G_{i_*,t} \le U(\delta, S_{i_*,t})$ for any t, then $Regret_T \le M \max_i U(\delta, S_{i,T})$ with probability at least $1-\delta$.*

Proof. First, we show that b_t is an optimistic estimate of the average optimal reward. By (4.1) and the regret guarantee of the optimal base,

$$b_t = \frac{R_{j_t,t}}{N_{j_t,t}} + \frac{U(\delta, S_{j_t,t})}{N_{j_t,t}} \ge \frac{R_{*,t}}{N_{*,t}} + \frac{U(\delta, S_{*,t})}{N_{*,t}} \ge \frac{\sum_{\tau \in N_{*,t}} \mu_{*,\tau}}{N_{*,t}} = \mu_* \ . \tag{4.3}$$

Let i_t be the base chosen at time t and j_t be the optimistic base. The cumulative regret of base i_t at time t can be bounded as

$$\begin{aligned} G_{i_t,t} &= N_{i_t,t}\mu_* - R_{i_t,t} & \\ &\le N_{i_t,t} b_t - R_{i_t,t} & \text{By (4.3)} \\ &\le N_{j_t,t} b_t - R_{j_t,t} & \text{By definition of } i_t \\ &= U(\delta, S_{j_t,t}) \ . & \text{By definition of } j_t \text{ and } b_t \end{aligned} \tag{4.4}$$

Let T_i be the last time step that base i is played. Given that the instantaneous regret is upper bounded by 1, by (4.4) the regret can be bounded as

$$\sum_{i=1}^{M} G_{i,T} = \sum_{i=1}^{M} G_{i,T_i} \le \sum_{i=1}^{M} U(\delta, S_{j_{T_i},T_i}) \le M \max_i U(\delta, S_{i,T}) \ .$$

□

The condition that $\mu_{*,t} = \mu_*$ for a constant μ_* regardless of time t is needed to ensure that $b_t \ge \sum_{\tau \in N_{i,t}} \mu_{*,\tau} / N_{i,t}$ for any base i. The condition holds in the following model selection problems: choosing a feature mapping in a stochastic bandit problem, and choosing the optimal exploration rate among a number of ϵ-greedy algorithms. The condition is also satisfied for choosing between multi-armed bandits and stochastic linear contextual bandits, where $\mu_{*,t} = \mathbb{E}\left[\max_{i \in [K]} \phi(s_t, i)^\top \theta_*\right]$ is a time-independent constant value for IID context s_t.

As we mentioned earlier, the regret upper bound U should be of order $\Omega(\sqrt{T})$. Thus, our approach can achieve the regret of the optimal base as long as the optimal regret is at least $\Omega(\sqrt{T})$. This observation is consistent with the lower bound argument of [54] who show that, in general, $O(\sqrt{T})$ is the best rate that can be achieved by any model selection strategy. Unfortunately, this lower bound implies that in a model selection setting, we can no longer hope to achieve the logarithmic regret bounds that can be usually obtained in stochastic bandit problems. Notice that such logarithmic bounds are shown for the *pseudo-regret* and not for the regret as defined above. The pseudo-regret is the difference of the expected rewards of the optimal arm and the arm played, and is not directly observed by the learner, and it can be estimated only up to an error of order $\Omega(\sqrt{T})$.

In fact we can show that a simple modification to the base selection strategy of Theorem 4.3.1 yields a model selection guarantee with a sublinear dependence in the number of models M.

Let $\rho \in (0, 1]$ and let i_t the base's index defined in Equation 4.2. Define j_t as in Equation 4.1. Instead of always playing i_t as above, we analyze a strategy that plays algorithm i_t if $\widehat{G}_{i_t,t} \leq \rho \widehat{G}_{j_t,t}$, and j_t if the opposite is true. The next theorem shows the model selection guarantee of this strategy.

Theorem 4.3.2. *If $\mu_{*,t} = \mu_*$ for a constant μ_* regardless of time t, and if with probability at least $1 - \delta$, $G_{i_*,t} \leq U(\delta, S_{i_*,t})$ for any t, then if $\rho = \frac{1}{\sqrt{M}}$ and $U(\delta, S_{i,t}) = c\sqrt{|S_{i,t}| \log(|S_{i,t}|/\delta)}$ for all $i \in [M]$, the refined regret balancing model selection strategy described above satisfies $\text{Regret}_T \leq 2c\sqrt{MT \log(T/\delta)}$ with probability at least $1 - \delta$.*

Proof. Let l_t be the base chosen at time t. Observe that whenever $l_t = i_t$ the following inequalities hold:

$$G_{l_t,t} = G_{i_t,t} \leq \widehat{G}_{i_t,t} \leq \rho \widehat{G}_{j_t,t} = \rho U(\delta, S_{j_t,t}) \leq \rho U(\delta, t)$$

Similarly, when $l_t = j_t$ the following inequalities hold:

$$G_{l_t,t} = G_{j_t,t} \leq \widehat{G}_{j_t,t} = U(\delta, S_{j_t,t})$$

For any $u \in [M]$ let T_u be the last time step that base u has been played before and including time T. Let $\mathcal{M}_i$ be the set of arms such that $u \in \mathcal{M}_i$ satisfy $l_{T_u} = i_{T_u}$. Similarly let $\mathcal{M}_j$ be the set of arms such that $u \in \mathcal{M}_j$ satisfy $l_{T_u} = j_{T_u}$. Notice that $\mathcal{M}_i \cup \mathcal{M}_j = [M]$. The sum of regrets of $u \in \mathcal{M}_i$:

$$\sum_{u \in \mathcal{M}_i} G_{u,T} \leq |\mathcal{M}_i| \rho U(\delta, T) \leq M \rho U(\delta, T).$$

The sum of regrets of $u \in \mathcal{M}_j$:

$$\sum_{u \in \mathcal{M}_j} G_{u,T} \leq \sum_{u \in \mathcal{M}_j} U(\delta, S_{j_{T_u},T})$$

Notice that $\sum_{u \in \mathcal{M}_j} S_{j_{T_u},T} \leq T$. Therefore if $U(\delta, t) = c\sqrt{t \log(t/\delta)}$, we get:

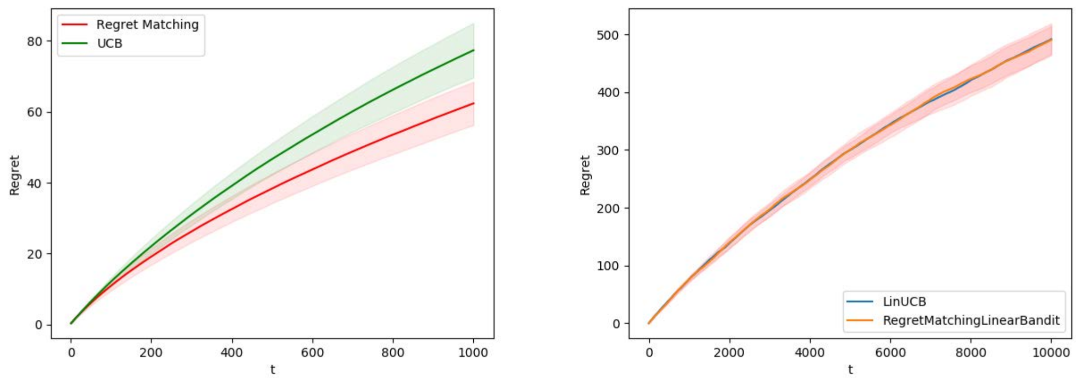


Figure 4.1: Regret Balancing vs UCB and OFUL. Mean and standard deviation of 2000 and 20 runs.

$$\sum_{u \in \mathcal{M}_j} U(\delta, S_{j_{T_u},T}) \le c\sqrt{MT\log(T/\delta)}.$$

Setting $\rho = \frac{1}{\sqrt{M}}$ the result follows.

□

4.4 Applications of Regret Balancing

In this Section, we show some applications of the regret balancing strategy.

Regret Balancing for Bandits

The regret balancing strategy can be used as a bandit algorithm. To use as a multi-armed bandit algorithm, we treat each arm as a base algorithm and we choose $U(\delta, t) = \sqrt{(t/2)\log(1/\delta)}$ as the regret of the optimal arm a_*. To see this, notice that by the sub-Gaussianity of the noise, with probability at least $1-\delta$, $G_{a_*,t} = \sum_{\tau \in N_{a_*,t}} \mu_* - R_{a_*,t} = \sum_{\tau \in N_{a_*,t}} (\mu_* - \mu_* + \eta_t) = \sqrt{(t/2)\log(1/\delta)}$. In Figure 4.1-Left, we compare regret balancing with the UCB algorithm [9] on a 4-armed Bernoulli bandit with means $\{0.1, 0.2, 0.3, 0.4\}$. In regret balancing, we treat each arm as a base algorithm and so we use $U(t) = \sqrt{(t/2)\log(1/\delta)}$ with $\delta = 0.1$ as the target regret.

Next, we show the implementation of the strategy as an algorithm for the linear stochastic bandits. Consider the following problem. In round t, the learner chooses action x_t from a (possibly time varying) decision space that is a subset of the unit sphere $D_t \subset \mathbb{S}^d$ and observes a reward $y_t = x_t^\top \theta_* + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is an unknown parameter vector and η_t is a σ-sub-Gaussian noise term.[3] Let $x_{t,*}$ be the optimal action at time t defined as $x_{t,*} = \arg\max_{x \in D_t} x^\top \theta_*$. The objective is to have small regret defined as $\text{Regret}_T = \sum_{t=1}^T (x_{t,*}^\top \theta_* - x_t^\top \theta_*)$.

We state some notation before defining the bandit method. For a regularization parameter $\lambda > 0$, let $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$ be the empirical covariance matrix, and let $\|z\|_V = \sqrt{z^\top V z}$ be the weighted ℓ^2-norm of vector z. Let $\widehat{\theta}_t = V_t^{-1} \sum_{k=1}^{t-1} x_k y_k$ be the regularized least-squares estimate. Let $\beta_t(\delta) = O(\sqrt{d \log(t)})$ be as defined in Section 4.7. Let $y_t = \arg\max_{x \in D_t} x^\top \widehat{\theta}_t + \beta_t(\delta) \|x\|_{V_t^{-1}}$ be the "optimistic" choice in round t. A UCB approach would take action y_t next. Regret balancing, however, uses the optimistic choice to estimate the empirical regrets of different choices. Let $b_t = y_t^\top \widehat{\theta}_t + \beta_t(\delta) \|y_t\|_{V_t^{-1}}$, which will be shown to be an upper bound on the value of the best action. In time t, we choose the action with the smallest *empirical regret*,

$$x_t = \arg\min_{x \in D_t} \widehat{G}_{x,t} \,, \qquad \widehat{G}_{x,t} = \frac{b_t - x^\top \widehat{\theta}_t}{\|x\|^2_{V_t^{-1}}} \,.$$

Intuitively, $b_t - x^\top \widehat{\theta}_t$ is an estimate of the instantaneous regret of action x and $1/\|x\|^2_{V_t^{-1}}$ is roughly the number of times that x is played.[4] Next theorem bounds the regret of the regret balancing strategy.

Theorem 4.4.1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, $\text{Regret}_T = \widetilde{O}(d^{3/2}\sqrt{T})$. Here $\widetilde{O}$ hides polylogarithmic terms in T, d, λ, and $1/\delta$.*

Proof. By Theorem 4.7.1, for all t, $x_{t,*}^\top \widehat{\theta}_t + \beta_t(\delta) \|x_{t,*}\|_{V_t^{-1}} \ge x_{t,*}^\top \theta_*$ with probability at least $1-\delta$. In what follows, we condition on the high probability event that these inequalities hold.

First, we show that b_t is an optimistic estimate of $x_{t,*}^\top \theta_*$. By definition of y_t,

$$b_t = y_t^\top \widehat{\theta}_t + \beta_t(\delta) \|y_t\|_{V_t^{-1}} \ge x_{t,*}^\top \widehat{\theta}_t + \beta_t(\delta) \|x_{t,*}\|_{V_t^{-1}} \ge x_{t,*}^\top \theta_* \,. \tag{4.5}$$

[3] This formulation includes the special case of linear contextual bandits with $D_t = \{\phi(s_t, a) : a \in [K]\}$.

[4] In multi-armed bandits, where actions are fixed axis aligned unit vectors, $1/\|x\|^2_{V_t^{-1}}$ counts the number of times an action is played.

We upper bound the instantaneous regret,

$$\begin{aligned}
r_t &= x_{t,*}^\top \theta_* - x_t^\top \theta_* \\
&\le b_t - x_t^\top \theta_* && \text{By (4.5)} \\
&\le b_t - x_t^\top \widehat{\theta}_t + \beta_t(\delta) \|x_t\|_{V_t^{-1}} && \text{By (4.10)} \\
&\le \beta_t(\delta)\|x_t\|_{V_t^{-1}} + \|x_t\|_{V_t^{-1}}^2 \left(\frac{b_t - y_t^\top \widehat{\theta}_t}{\|y_t\|_{V_t^{-1}}^2} \right) && \text{By definition of } x_t \\
&= \beta_t(\delta)\|x_t\|_{V_t^{-1}} + \|x_t\|_{V_t^{-1}}^2 \cdot \frac{\beta_t(\delta)}{\|y_t\|_{V_t^{-1}}} && \text{By (4.5)}\ .
\end{aligned}$$

Using the fact that $\lambda_{\max}(V_t) \le \text{trace}(V_t) = \lambda d + \sum_{k=1}^{t-1} \|x_t\|^2 \le \lambda d + t$, and hence $\lambda_{\min}(V_t^{-1}) = \frac{1}{\lambda_{\max}(V_t)} \ge 1/(\lambda d + t)$, we get that $\|y\|_{V_t^{-1}}^2 \ge 1/(\lambda d + t)$ for any $y \in D_t$. Thus,

$$r_t \le \beta_t(\delta)\|x_t\|_{V_t^{-1}} + \beta_t(\delta)\|x_t\|_{V_t^{-1}}^2 \sqrt{\lambda d + t}\ .$$

Thus, by Cauchy–Schwarz inequality and Lemma 4.7.2,

$$\begin{aligned}
\text{Regret}_T &= \sum_{t=1}^{T} \left(\beta_t(\delta)\|x_t\|_{V_t^{-1}} + \beta_t(\delta)\|x_t\|_{V_t^{-1}}^2 \sqrt{\lambda d + t} \right) \\
&\le \beta_T(\delta) \left(\sqrt{T \sum_{t=1}^{T} \|x_t\|_{V_t^{-1}}^2} + 2d \log(1 + T/(\lambda d)) \sqrt{\lambda d + T} \right) \\
&\le \beta_T(\delta) \left(\sqrt{2dT \log(1 + T/(\lambda d))} + 2d \log(1 + T/(\lambda d)) \sqrt{\lambda d + T} \right)\ .
\end{aligned}$$

□

The regret bound in the theorem is slightly worse than the minimax optimal rate of $\widetilde{O}(d\sqrt{T})$, however and as we show next, regret balancing strategy can be a competitive linear bandit algorithm in practice. In Figure 4.1-Right, we compare regret balancing, as described above, with the OFUL algorithm [1] on a contextual linear bandit problem with two arms: for $i \in \{1, 2\}$, let $\theta_i \in \mathbb{R}^3$ drawn uniformly at random from $[0, 1]^3$ at the beginning of the experiment. In round t, the reward of arm $i \in \{1, 2\}$ is $\theta_i^\top s_t + \xi$ where $\xi \sim N(0, 1)$ and context $s_t \in \mathbb{R}^3$ is drawn uniformly at random from $[0, 1]^3$ with $s_t[0] = 1$.

Optimizing the Exploration Rate

Next, we consider the performance of regret balancing as a bandit model selection strategy. First, consider optimizing the exploration rate in an ϵ-greedy algorithm. The ϵ-greedy is a simple and popular bandit method. In round t, the algorithm plays an action chosen uniformly at random with a small probability ϵ_t, and plays the empirically best, or greedy, choice otherwise. For a well-chosen ϵ_t, this simple strategy can be very competitive. The optimal value of ϵ_t however depends on the unknown reward function: It is known that the optimal value of ϵ_t is $\min\{1, \frac{5K}{\Delta^2 t}\}$ where Δ is the smallest gap between the optimal reward and the sub-optimal rewards [41]. By this choice of exploration rate, the regret scales as $\widetilde{O}(\sqrt{T})$ for $K = 2$ and $\widetilde{O}(T^{2/3})$ for $K > 2$.

We apply the regret balancing strategy to find a near-optimal exploration rate. The result directly follows from Theorem 4.3.1. A similar result, but for a different algorithm, is shown by [54].

Corollary 4.4.1. *Let T be the time horizon. Let $B = \{1, 2, \ldots, \lfloor \log(T) \rfloor\}$. For $i \in B$, let B_i be the ϵ-greedy algorithm with exploration rate $\epsilon_t = 2^i/t$ in round t. By the choice of $U(t) = t^{1/2}$ for $K = 2$ (or $U(t) = t^{2/3}$ for $K > 2$), the regret balancing model selection with the set of base algorithms B achieves $\widetilde{O}(\sqrt{T})$ regret for $K = 2$ (or $\widetilde{O}(T^{2/3})$ for $K > 2$).*

Next, we evaluate the performance of regret balancing in finding a near optimal exploration rate. Consider a bandit problem with two Bernoulli arms with means $\{0.5, 0.45\}$. Consider 18 ϵ-greedy base algorithms with exploration $\epsilon_t = c/t$, where values of c are on a geometric grid in $[1, 2T]$. Apply regret balancing with the target regret bound $U(t) = \sqrt{t}$, and the set of ϵ-greedy base algorithms. The experiment is repeated 20 times. Figure 4.2-Left shows the performance of regret balancing strategy.

Representation Learning

The sublinear regret bounds of linear bandit algorithms are valid as long as the reward function is truly a linear function of the input feature representation. Assume it is known that the reward function is linear in one of the M feature maps $\{\phi_i : D_t \to \mathbb{R}^d : i \in [M]\}$, but the identity of the true feature map is unknown. By applying

[5] The shaded areas around UCB and CORRAL are the std. The shaded areas around the ϵ-greedy bases are 0.1 of std . For small ϵ, ϵ-greedy has a very high variance because it either commits to the optimal arm or the sub-optimal arm at the beginning, so plotting the whole std would make the plot unreadable.

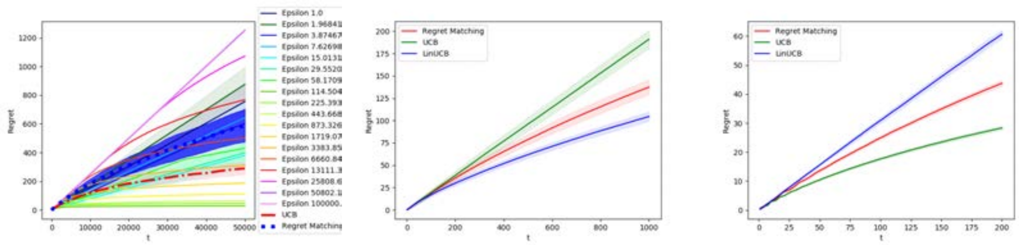


Figure 4.2: Left: Optimising the exploration rate with regret balancing (Mean and standard deviation of 20 runs[5]), Middle and Right: Regret balancing to choose between UCB and LinUCB (Mean and standard deviation of 500 and 200 runs).

Theorem 4.3.1 to M OFUL algorithms, each using one of the feature maps, we obtain a regret that scales as $\widetilde{O}(Md\sqrt{T})$.

As an application, we consider the problem of choosing between UCB and OFUL. Contexts are drawn from the standard normal distribution, but the first element in the context vector is always 1. The noise is $\xi \sim N(0, \sigma^2 = 0.1)$. First, consider a problem with $K = 2$ arms, each having a reward vector in $\mathbb{R}^{10}$ drawn uniformly at random from $[0, 1/3]^{10}$ at the beginning. We use regret balancing with target function $U(t) = \sqrt{2t}$ to perform model selection between UCB and OFUL. Results are shown in Figure 4.2-Middle. In this experiment, OFUL performs better than UCB, and performance of regret balancing is in between. Next we consider a problem with $K = 5$ arms. Mean reward of arm $i \in [K]$, denoted by μ_i, is generated uniformly at random from $[0, 1]$ at the beginning. In each round, we observe a context $s_t \in \mathbb{R}^{10}$, but the expected reward of arm i in each round is μ_i. We use target regret $U(t) = \sqrt{5t}$. Figure 4.2-Right shows that in this setting UCB performs better than OFUL, and performance of regret balancing is again in between.

Choosing Among Reinforcement Learning Algorithms

We consider the model selection problem in finite-horizon reinforcement learning problems. The ideas can be easily extended to average-reward setting as well, but we choose a finite-horizon setting to simplify the presentation.

A finite-horizon reinforcement learning problem is specified by a horizon H, a state space S that is partitioned into H disjoint sets, an action space A, a transition dynamics P that maps a state-action pair to a distribution over the states in the next stage, and a reward function r that assigns a scalar value to each state-action pair. The objective is to find a policy π, that is a mapping from states to distributions on actions, that maximizes the total reward.

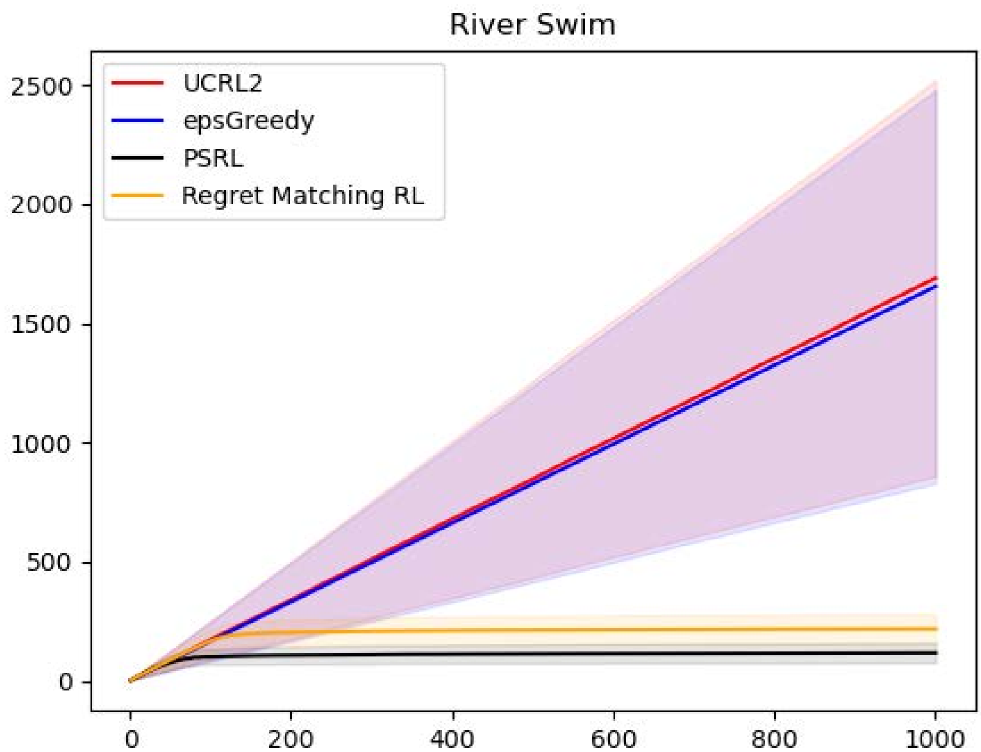


Figure 4.3: Regret balancing for model selection among ϵ-greedy, UCRL, and PSRL. Mean and 0.2 of standard deviation of 10 runs.

The model selection problem is defined next. In episode t, the learner chooses base i_t from a set of M RL algorithms, the base is executed for H rounds, and the rewards of the actions are revealed to the learner. Let $V_{*,t}$ be the total reward of the optimal policy in the underlying reinforcement learning problem. Quantities $N_{i,t}$, $R_{i,t}$, $S_{i,t}$, i_*, U, etc are defined similar to the bandit case. For example, $N_{i,t}$ is the number of episodes that base i is played up to episode t. The regret balancing strategy is defined next. In episode t, let $j_t = \arg\max_{i\in[M]} \frac{R_{i,t}}{N_{i,t}} + \frac{U(\delta,S_{i,t})}{N_{i,t}}$ be the optimistic base. Let b_t such that $N_{j_t,t}b_t - R_{j_t,t} = U(\delta, S_{j_t,t})$. Define the empirical regret of base i by $\widehat{G}_{i,t} = N_{i,t}b_t - R_{i,t}$. In episode t, we choose the base with the smallest empirical regret: $i_t = \arg\min_i \widehat{G}_{i,t}$. The next theorem shows the model selection guarantee for the regret balancing strategy. The analysis is almost identical to the analysis of the bandit model selection in the previous section.

Theorem 4.4.2. *If $V_{*,t} = V_*$ for a constant V_* regardless of round t, and if for any $\delta \in (0,1)$ with probability at least $1-\delta$, $G_{i_*,t} \le U(\delta, S_{i_*,t})$ for any t, then $Regret_T \le M \max_i U(\delta, S_{i,T})$ with probability at least $1-\delta$.*

In Figure 4.3, we perform model selection with base algorithms UCRL2 [34], a Q-learning method with ϵ-greedy exploration and $\epsilon = 0.1$, and PSRL [52] in the River Swim domain [61]. Regret balancing adapts to the best performing strategy (PSRL in this case).

As another application, consider the problem of choosing state representation in reinforcement learning. Many existing theoretical results hold under the assumption that a correct state representation (or feature map) is given. As examples, [3] show sublinear regret bounds under the assumption that the value function of any policy is linear in a given feature vector, while [38] show sublinear regret bounds for linear MDPs, i.e. when the transition dynamics and the reward function are known to be linear in a given feature vector. Given M candidate feature maps, one of which is fully aligned with the true dynamics of the MDP, we can apply the regret balancing strategy and by Theorem 4.4.2, the performance will be optimal up to a factor of M.

Corollary 4.4.2. *Let $\mathcal{M} = (S, A, H, P, r)$ be a linear MDP parametrized by an unknown feature map $\{\Phi^* : S \times A \to \mathbb{R}^d\}$. Let $F = \{\Phi_i(s,a)\}_{i=1}^M$ be a family of feature maps with $\Phi_i(s,a) \in \mathbb{R}^d$ and satisfying $\Phi^* \in F$. For regret balancing with target $U(t) = d^{3/2}H^{3/2}T^{1/2}$ and with a class of LSVI-UCB base algorithms [38], each instantiated with a feature map in F, the regret is bounded as $Regret_T \le \widetilde{\mathcal{O}}\left(M\sqrt{d^3H^3T}\right)$.*

[46, 47, 50] study a closely related but different problem where M state representation functions are given and with at least one such function, the resulting state evolution is Markovian.

4.5 Lower Bound

In this section we show that for any model selection algorithm there are problem instances where the algorithm must do regret balancing. For simplicity we restrict ourselves to the case $M = 2$, and to a simple class of problem instances, although it is possible to extend the argument to richer families and beyond two base algorithms.

Let $\mathcal{M}$ be a model selection algorithm with expected regret $\mathcal{R}(t)$ up to time t. We say an algorithm "model selects" w.r.t. a class of algorithms $\mathcal{B}$ if for any two base algorithms $A, B \in \mathcal{B}$ with expected regret $\mathcal{R}_A$ and $\mathcal{R}_B$, there exists $T_0 > 0$ such that for all $T \geq T_0$, $\mathcal{R}(T) \leq \mathcal{O}(\min(\mathcal{R}_A(T), \mathcal{R}_B(T)))$. We say that algorithm $\mathcal{M}$ is regret balancing for base algorithms (A, B) if for all $\delta \in (0, 1)$ there exists $T(\delta)$ such that for all $T \geq T(\delta)$, with probability at least $1 - \delta$,

$$\log\left(\max\left(\frac{\widetilde{\mathcal{R}}_A(T)}{\widetilde{\mathcal{R}}_B(T)}, \frac{\widetilde{\mathcal{R}}_B(T)}{\widetilde{\mathcal{R}}_A(T)}\right)\right) \leq o(\log(T)), \tag{4.6}$$

where $\widetilde{\mathcal{R}}_A(T)$ and $\widetilde{\mathcal{R}}_B(T)$ are the empirical regrets of algorithms A and B, respectively. The main result of this section is to show there exist problem and algorithm classes such that any model selection strategy must be regret balancing.

Theorem 4.5.1. *There exists two algorithm classes $\mathcal{B}_1, \mathcal{B}_2$ with $\mathcal{B}_1 \subseteq \mathcal{B}_2$ such that any model selection strategy $\mathcal{M}$ for class $\mathcal{B}_2$ must satisfy the condition in (4.6) for all $\mathcal{A}, B \in \mathcal{B}_1$ whose regrets are distinct.*

Proof. Let $\mathcal{B}_1, \mathcal{B}_2$ be two classes of algorithms defined as follows: if $B \in \mathcal{B}_1$ then there exists a value b such that B has a deterministic instantaneous regret of b during all time steps. If $B \in \mathcal{B}_2$, then there is a time index t_0 and two values b_1 and b_2 such that B has a deterministic instantaneous regret of b_1 for all $t \leq t_0$ and a deterministic instantaneous regret of b_2 for all $t > t_0$.

Let $A \in \mathcal{B}_1$ be an algorithm that for all time-steps $t \in [T]$ plays a policy achieving (deterministically) an instantaneous regret of $\frac{1}{T^{1-a}}$ for some $a \in [0, 1]$. Similarly let $B \in \mathcal{B}_1$ be an algorithm that for all time-steps $t \in [T]$ plays a policy with a deterministic instantaneous regret of $\frac{1}{T^{1-b}}$ for some $b \in [0, 1]$.

We proceed by contradiction. If $\mathcal{M}$ is not regret balancing for (A, B), then, there exists an $\epsilon > 0$ such that with probability at least ϵ,

$$\max\left(\frac{\widetilde{R}_A(T)}{\widetilde{R}_B(T)}, \frac{\widetilde{R}_B(T)}{\widetilde{R}_A(T)}\right) \geq CT^c \tag{4.7}$$

for some positive constants $C, c > 0$, and for infinitely many $T > T(\epsilon)$. Wlog the condition in (4.7) implies that for infinitely many $T \geq T(\epsilon)$ with probability at least $\epsilon/2$,

$$\widetilde{R}_A(T) \geq C\widetilde{R}_B(T)T^c\ . \tag{4.8}$$

Call this event $\mathcal{E}_T$ for any such T. Let T_A and T_B be the random number of times that algorithm A (respectively algorithm B) was called by $\mathcal{M}$. In this case, by (4.8), with probability at least $\frac{\epsilon}{2}$,

$$CT_b \frac{1}{T^{1-b}} T^c \leq T_a \frac{1}{T^{1-a}}\ . \tag{4.9}$$

This, in turn, implies $T_a \geq CT^c T_b T^{b-a}$. Additionally, since $T_a \leq T$, with probability at least $\epsilon/2$ we have $T_b \leq \frac{1}{C} T^{1+a-b-c}$. We now proceed to show a lower bound for the regret in each of two cases, $a > b$ and $b > a$.

Case $a > b$: Let $\mathcal{E}_T = \mathcal{E}_T^1 \cup \mathcal{E}_T^1$ where $\mathcal{E}_T^1 = \{T_a \geq \frac{T}{2}\} \cap \mathcal{E}_T$ and $\mathcal{E}_T^1 = \{T_a < \frac{T}{2}\} \cap \mathcal{E}_T$. Notice that $\max(\mathbb{P}(\mathcal{E}_T^1), \mathbb{P}(\mathcal{E}_T^2)) \geq \frac{\epsilon}{4}$. In $\mathcal{E}_T^1$ we have $\widetilde{R}_a(T) \geq \frac{T^a}{2}$. In $\mathcal{E}_T^2$ we have $T_b \geq \frac{T}{2}$ which in turn by (4.9) implies that $T_a \geq CT^{1+c+b-a}$, and therefore in $\mathcal{E}_T^2$ it holds that $\widetilde{R}_a(T) \geq C\frac{T^{c+b}}{2}$. Since $\mathcal{R}(T) = \mathbb{E}[\widetilde{R}_a(T) + \widetilde{R}_b(T)]$, we conclude that $\mathcal{R}(T) \geq \frac{\epsilon}{4} \min\left(C\frac{T^{c+b}}{2}, \frac{T^a}{2}\right)$.

Case $b > a$: Assume $\mathcal{M}$ has model selection guarantees (in expectation) w.r.t. algorithm A. Therefore $\mathcal{R}(T) \leq C'' T^a$. As a consequence of (4.9), with probability at least $\frac{\epsilon}{2}$, it holds that $T_b \leq \frac{1}{C} T^{1+a-b-c} = o(T)$.

This analysis shows that in case $\mathcal{M}$ does not satisfy regret balancing, then it must be the case that:

1. If $a > b$, then $\mathcal{M}$ must incur an expected regret of at least $\frac{\epsilon}{2} \min\left(C\frac{T^{c+b}}{2}, \frac{T^a}{2}\right)$ for some $C > 0$, and thus precluding any model selection guarantees for $\mathcal{M}$.
2. If $b > a$, then with probability at least $\frac{\epsilon}{2}$ it follows that $T_b \leq \frac{1}{C} T^{1+a-b-c}$ for some constants C, c. Furthermore, if $\mathcal{M}$ is assumed to satisfy model selection guarantees, it must be the case that for T large enough, with probability at least $\frac{\epsilon}{2}$, $T_a \geq T/2$. We focus on this case to find a contradiction.

Two alternative worlds: Having analyzed what happens if a model selection strategy does not do regret balancing with algorithms A and B, we proceed to show our lower bound. Let (A, B) two base algorithms defined as above and let (A', B') be two base algorithms defined as:

1. A' acts exactly as A does.
2. B' acts as B does only up to time $t' = \min(\frac{1}{C} T^{1+a-b-c} + 1, T)$, and afterwards it pulls the optimal arm deterministically.

Let $T_{a'}$ and $T_{b'}$ be the random number of times A' and B' are played by $\mathcal{M}$. Suppose the model selection strategy $\mathcal{M}$ is presented with (A'', B'') sampled uniformly at random between (A, B) and (A', B'). Let $b > a$. Note that environment (A', B') is indistinguishable from environment (A, B) in the probability at least $\epsilon/2$ event that $T_b < t'$. This implies that in environment (A', B'), and with probability at least $\frac{\epsilon}{2}$,

$T_{b'} < t' = o(T)$. In this same event and for T large enough since $T_{a'} + T_{b'} = T$ it must be the case that $T_{a'} \geq T/2$ and $T_{b'} \leq \frac{1}{C} T^{1+a-b-c}$, and therefore,

$$\mathbb{E}_{(A'',B'')}[\mathcal{R}(T)|(A'',B'') = (A',B')] \geq \frac{\epsilon}{8} T^a \, .$$

Since for T large enough the optimal regret for (A', B') is $\frac{1}{C} T^{1+a-b-c} \cdot \frac{1}{T^{1-b}} = \frac{1}{C} T^{a-c}$, and for T large enough, $\frac{1}{C} T^{a-c} = o(T^a)$, we conclude that $\mathcal{M}$ couldn't have possibly satisfied the model selection condition. □

4.6 Regret balancing without knowledge of the optimal regret

In this section, we show that by adding forced exploration, and hence direct access to the arms, the regret balancing strategy can achieve near-optimal regret in a class of problems without the knowledge of the optimal base regret.

Let M denote the number of base algorithms, k^* denote the best base, and $M_{k,t}$ denote the number of times base k is selected up to time t. We use K, i^*, and $N_{i,t}$ to denote the number of arms, the best arm, and the number of times arm i is selected. Let μ_i and $\widehat{\mu}_{i,t}$ denote the true mean and the empirical mean of arm i at time t. Let $R_{k,t}$ and $G_{k,t}$ be the reward and regret of base k up to time t. The model selection algorithm is as follows: in the first phase, pull each arm $\left(\frac{TM}{K}\right)^{2/3}$ times. In the second phase, for each time step $t \in \{(TM)^{2/3} K^{1/3} + 1, \ldots, T\}$, play base $k_t = \text{argmin}_k \, \widehat{G}_{k,t}$, where $\widehat{G}_{k,t} = M_{k,t} b_t - R_{k,t}$, $b_t = \widehat{\mu}_{j_t,t} + \frac{1}{\sqrt{N_{j_t,t}}}$, and $j_t = \text{argmax}_i \, \widehat{\mu}_{i,t} + \frac{1}{\sqrt{N_{i,t}}}$. The regret from the initial exploration phase is $(T\mathcal{M})^{2/3} K^{1/3}$. Next we analyze the regret from the second phase. The cumulative regret of base k_t at time t when it is selected can be bounded as,

$$\begin{aligned}
G_{k_t,t} = M_{k_t,t} \cdot \mu_{i^*} - R_{k_t,t} \quad & \overset{(i)}{\leq} M_{k_t,t} \cdot \left(\widehat{\mu}_{i^*,t} + \frac{1}{\sqrt{N_{i^*,t}}} \right) - R_{k_t,t} \\
\overset{(ii)}{\leq} M_{k_t,t} b_t - R_{k_t,t} \quad & \overset{(iii)}{\leq} M_{k^*,t} b_t - R_{k^*,t} \\
\overset{(iv)}{=} M_{k^*,t} \cdot \left(\widehat{\mu}_{j_t,t} + \frac{1}{\sqrt{N_{j_t,t}}} \right) - R_{k^*,t} \quad & \overset{(v)}{\leq} M_{k^*,t} \cdot \left(\mu_{j_t} + \frac{2}{\sqrt{N_{j_t,t}}} \right) - R_{k^*,t} \\
\overset{(vi)}{\leq} M_{k^*,t} \cdot \left(\mu_{i^*} + \frac{2}{\sqrt{N_{j_t,t}}} \right) - R_{k^*,t} \quad & \leq M_{k^*,t} \cdot \mu_{i^*} - R_{k^*,t} + \frac{M_{k^*,t}}{\sqrt{N_{j_t,t}}} \\
= G_{k^*,t} + \frac{M_{k^*,t}}{\sqrt{N_{j_t,t}}} \quad & \overset{(vii)}{\leq} G_{k^*,t} + T^{2/3} K^{1/3} M^{-1/3} \qquad ,
\end{aligned}$$

where inequality (i) holds because $\mu_{i^*} \leq \widehat{\mu}_{i^*,t} + \frac{1}{\sqrt{N_{i^*,t}}}$ w.h.p by Hoeffding, inequality (ii) is a result of the definition of b_t, (iii) follows by definition of k_t, (iv) follows again by definition of b_t, (v) holds because $\widehat{\mu}_{j_t,t} - \frac{1}{\sqrt{N_{j_t,t}}} \leq \mu_{j_t}$ w.h.p by Hoeffding, (vi) is satisfied because $\mu_{j_t} \leq \mu_{i^*}$, and the last inequality $(viii)$ holds because $M_{k^*,t} \leq t \leq T$ and $N_{j_t,t} \geq \left(\frac{TM}{K}\right)^{2/3}$ given the initial exploration phase. Let T_k be the last time step that base k is picked. The regret from the second phase can be bounded as,

$$
\begin{aligned}
\sum_k G_{k,T} &= \sum_k G_{k,T_k} \\
&\leq \sum_k (G_{k^*,T_k} + T^{2/3} K^{1/3} M^{-1/3}) \\
&\leq \sum_k (G_{k^*,T} + T^{2/3} K^{1/3} M^{-1/3}) \\
&= M(G_{k^*,T} + T^{2/3} K^{1/3} M^{-1/3}) \ .
\end{aligned}
$$

The total regret is at most

$$
\text{Regret}_T \leq 2T^{2/3} M^{2/3} K^{1/3} + M G_{k^*,T} \ .
$$

If $G_{k^*,T} = \Omega(T^{2/3} K^{1/3} M^{-1/3})$ then the total regret is the same as the regret of the best base.

4.7 Ancillary Technical Results

We state a result on the error of the least-squares estimate.

Theorem 4.7.1 (Theorem 2 of [1]). *Assume $\|\theta_*\| \leq S$. Let*

$$
\beta_t(\delta) = R\sqrt{\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \ .
$$

For any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$ and any $x \in \mathbb{R}^d$,

$$
|x^\top(\widehat{\theta}_t - \theta_*)| \leq \beta_t(\delta)\|x\|_{V_t^{-1}} \ . \tag{4.10}
$$

Lemma 4.7.2 (Lemma 11 of [1]). *Let $\{X_t\}_{t=1}^\infty$ be a sequence in $\mathbb{R}^d$ and define $V_t = \lambda I + \sum_{k=1}^t X_k X_k^\top$ for a regularizer $\lambda \geq 1$. If $\|X_t\| \leq 1$ for all t, then*

$$
\sum_{k=1}^t \|X_k\|^2_{V_{k-1}^{-1}} \leq 2\log\frac{\det(V_t)}{\det(\lambda I)} \leq 2d\log(1 + t/(\lambda d)) \ .
$$

Chapter 5

Regret Bound Balancing and Elimination

5.1 Introduction

Multi-armed bandits are a general framework of sequential decision making that has in the last two decades received a lot of attention. The main aspect of this framework is a sequence of T *rounds* of interaction between a learning agent and an unknown environment. During each round, the learner picks an action from a set of available actions on that round, and the environment consequently generates a feedback (e.g., in the form of a *reward* value) associated with the chosen action. Given a class of benchmark policies, the goal of the learning agent is to accumulate during the course of the T rounds a total reward which is not much smaller than that of the best policy in hindsight within the benchmark class. Multi-armed bandits have found applications in a wide variety of domains, like clinical trials (e.g., [66]), online advertising (e.g., [57]), recommendation systems (e.g., [44]), and beyond.

Since many bandit methods are often deployed at scale in industrial applications, the complexity and diversity of the involved learning solutions typically require being able to select among several alternatives, like selecting the best within a pool of algorithms, or even alternative configurations of the same algorithm (as in, e.g., hyper-parameter optimization). Hence, the problem of *model selection* in bandit algorithms has become chiefly important in order to simplify the development of data processing pipelines at scale while simultaneously achieving improved statistical performance.

In this chapter, we study the problem of online model selection among a set of alternative learning algorithms, these algorithms being themselves bandit algorithms.

Each such algorithm is designed to work well only when favorable conditions are satisfied. Yet, the algorithm designer may not know in advance which one of them is more appropriate for the problem at hand.

As a simple example, many known multi-armed bandit algorithms, such as UCB (e.g., [41, Ch. 7]), rely on a confidence interval width as prescribed by a theoretical recipe. However, it has been observed multiple times in practice that setting this width smaller than theoretically suggested can lead to substantial performance improvements. On the other hand, picking too small a width can lead to a dramatic degradation in performance that may translate into a linear regret. It is therefore desirable to design theoretically sound model selection procedures that can help us find an optimal parameter setting in an online fashion.

Another simple example comes from trying to distinguish between a contextual and a non-contextual environment. In e-commerce problems, even if contextual information is available about users and the transaction at hand, it may prove more beneficial to use a simple UCB style algorithm that ignores the context or that only uses part of the context information. A model selection strategy that selects when or to what extent making use of contextual information can lead to better performance for contextual bandit algorithms.

Related Work and our Contribution

In this chapter we aim to develop a general purpose model selection master algorithm (that is, aggregation approach) that can be combined with multiple base bandit algorithms, and is able to obtain regret guarantees competitive with respect to the best base algorithm.

The problem of online model selection for bandit algorithms has received a lot of recent attention, as witnessed by a flurry of recent works (e.g., [4, 5, 7, 12, 18, 26, 27, 28, 43, 54]).

These previous works on model selection can be broadly split into two approaches: (i) Approaches that make use of an adversarial master algorithm, and (ii) approaches that rely on a statistical test which is able to detect when a base algorithm is misspecified. Our approach, called *Regret Balancing and Elimination*, falls squarely in the second camp.

Within the first category are the so-called *corraling* algorithms. These yield statistical guarantees of the form $\mathcal{O}(d_\star^\alpha T^\beta)$ for some $\alpha \geq 1, \beta < 1$, where $d_\star$ depends generally on the complexity of the best model class or algorithm and other problem parameters. The original Corraling Algorithm of [5] relies on an adversarial master algorithm based on mirror descent that can be combined with many base algorithms (both stochastic and adversarial), provided these base algorithms satisfy a stability guarantee. In this case, the base algorithms are fed with an importance-weighted estimator of the reward, hence they have to be robust to potentially wide fluctuations in the reward scaling, due to the evolving nature of the master algorithm's distribution over base algorithms. Unfortunately, in order to show that a base algorithm can be combined with the corralling master to satisfy a valid model selection regret guarantee, it is necessary to verify that the above-mentioned stability condition holds, something that has to be done on a case-by-case basis. The model selection guarantee is of the form $\mathcal{O}\left(\sqrt{MT} + MR_{i_\star}(T)\right)$, where M is the number of base algorithms and $R_{i_\star}(T)$ is the regret guarantee of any of the base algorithms. Yet, this is achieved only if the master's learning rate is set as a function of $R_{i_\star}(T)$, a quantity which is typically unknown.

Some of the shortcomings of the original Corralling Algorithm have been addressed by the more recent work of [54] (see Chapter 2). The authors propose a generic model selection procedure to combine stochastic bandit algorithms with an adversarial master. As opposed to the corralling algorithm of [5], the *Stochastic CORRAL* method in [54] allows the use of any stochastic bandit algorithm in stochastic contextual environments (the contexts are i.i.d.), provided it satisfies a high probability regret guarantee, thus relaxing the stability condition in [5]. [54] obtain the following model selection guarantees: When the base algorithms have a regret bound of the form $\{d_i T^\alpha\}_{i=1}^M$, Stochastic CORRAL achieves a regret guarantee of $\widetilde{\mathcal{O}}(\sqrt{MT}+M^\alpha T^{1-\alpha}+M^{1-\alpha}T^\alpha d_{i_\star}^{1/\alpha})$ when using a Corralling Algorithm as master, and a rate of $\widetilde{\mathcal{O}}(\sqrt{MT} + M^{\frac{1-\alpha}{2-\alpha}} T^{\frac{1}{2-\alpha}} d_{i_\star})$ under a forced exploration EXP3 (e.g., [41, Ch. 11]) master. Despite these advances, it remains unclear how to avoid the $\sqrt{MT}$ cost of a corralling approach. Our approach recovers and improves on the guarantees obtained by [5] and [54] in two ways. First, we propose a general purpose stochastic master algorithm that can be used in combination with any set of stochastic bandit algorithms. As opposed to the adversarial master algorithms of [5] and [54], ours is much more interpretable and transparent. Second, due to the stochastic nature of our master algorithm, we are able to prove *gap-dependent* bounds, thereby departing from the inherent $\sqrt{T}$ limit of adversarial master approaches. Furthermore, the memory requirements of [54] are very onerous, since their algorithm requires to store all the policies played by the base algorithms. Our algorithm's memory requirements are minimal in comparison.

There exist other related approaches in the literature that make use of an adversarial corralling master algorithm as a means of performing model selection. [7] propose an approach based on a Tsallis-INF adversarial master, which is able to recover gap-dependent regret guarantees for stochastic bandit problems. Nevertheless, their approach suffers from the drawback that whenever the rates of the input base algorithms are of the form $\{d_i T^\alpha\}_{i=1}^M$, where $d_1 \leq \cdots \leq d_M$, they obtain a regret guarantee for their master algorithm of the form $d_M T^\alpha$, a quantity that could be substantially worse than the regret achievable by the optimal base algorithm $d_{i_\star} T^\alpha$, since $d_{i_\star}$ might be much smaller than d_M. In contrast, our approach achieves a rate of $d_{i_\star}^2 T^\alpha$. Other related approaches that make use of a Tsallis-INF adversarial master have also been proposed, e.g., [26] achieve optimal rates for selecting the the misspecification level in the setting of contextual linear bandits. In the setting of stochastic linear bandits with adversarial contexts, our approach can be seen to achieve the same model selection rates as [26] for the problem of selecting the best level of misspecification.

As for the approaches that rely on a statistical test to perform model selection, minimax-optimal guarantees have been shown under strong eigenvalue assumptions on the context distribution by leveraging the special structure of the stochastic linear contextual bandit setting [18, 27]. These algorithms work by maintaining a set of active base learners, and playing a low complexity algorithm/model within the set. If enough information is gathered to conclude that a higher complexity model better describes the observed data, they eliminate the low complexity model from the active set, and proceed to play a more complex one. Unlike those papers, we are able to get results for the nested linear class problem (initially studied by [27]), but without resorting to eigenvalue assumptions on the context distribution, and without relying on the finiteness of the action space.

In the more general task of selecting among different stochastic bandit algorithms operating in a stochastic environment (with i.i.d. contexts), the recent work [4] has taken some steps towards proposing a stochastic master algorithm that can combine multiple stochastic base bandit algorithms, and obtain regret guarantees of the same nature or better than Stochastic CORRAL. [4] introduce an intriguing new technique for model selection referred to as Regret Balancing. At a high level, the main idea is to estimate the empirical regret of the base algorithms during the rounds that the algorithms are played, and ensure that all base algorithms suffer roughly the same empirical regret. As opposed to [18, 27] the Regret Balancing approach of [4] does not eliminate any base algorithm.[1] Unfortunately, in order for this approach to work, the exact scaling of the target optimal regret guarantee is required, which is again typically unknown. Our approach to model selection expands on the fundamental insights of regret balancing but, in contrast to [4], we are able to obtain results when model selecting among multiple base algorithms with different regret guarantees.

In [43] the authors propose ECE (Explore Commit Exploit), a model selection algorithm on stochastic contextual bandit algorithms. ECE can be thought of as

[1]Technically speaking the methods in [18, 27] do not eliminate base algorithms, but reject a statistical hypothesis on the base algorithms' model complexity.

an epsilon-greedy approach to the problem of model selection. Correspondingly, the regret guarantees of ECE have a dependence on T of the order of $T^{2/3}$, in contrast to our typical $T^{1/2}$ dependence. A regret of the form $T^{2/3}$ is the same as the one achievable by a forced exploration EXP3 master in [54]. [43] also present gap-dependent guarantees under the same assumptions as in [7] (see also [12]): each algorithm satisfies a valid regret guarantee w.r.t. its own policy class. Our work does not rely on this restrictive assumption, in that we only require the optimal algorithm to be well behaved and satisfy its theoretical regret guarantee. This is because we admit the presence of regret-misspecified base algorithms in the pool, and compete against the best among the well-specified ones. When the rates of the base algorithms are of the form $\{d_i T^\alpha\}_{i=1}^M$ and in the regime where T is much larger than d_i, our approach strictly dominates ECE's rates. Other works provide model selection results for specific bandit models, most notably, [28] consider the problem of selecting over nested feature structures and an unknown parameter norm in the case of contextual linear bandits over a sphere. Our results recover model selection rates for these problems without requiring restrictive assumptions on the nature of the contexts.

Content of the Chapter

Building on Chapter 4, we study a general regret bound balancing and elimination algorithm (Section 5.3) for selection among a pool of base bandit algorithms, each coming with a *presumed* regret bound that may or may not hold. The master algorithm does not know a priori the identity of the base algorithms whose regret bounds hold. Under these general assumptions, we show that this master algorithm enjoys general regret guarantee (Section 5.4) that can be specialized to either the gap-independent or the gap-dependent case. Then, we specialize to relevant application examples with nested model classes (Section 5.5) that consider linear contextual bandits or linear Markov decision processes as base learners). We also consider therein the unknown misspecification case, as well as the practically relevant problem of optimally tuning linear contextual bandit algorithms like OFUL. Finally, we specifically focus on the nested linear contextual bandit setting, and extend our balancing and elimination technique to the case where the context information is generated adversarially (Section 5.6). Despite we do not show this explicitly, similar extensions can be exhibited for Linear Markov Decision Processes with Nested Model Classes, Linear Bandits and MDPs with Unknown Approximation Error and tuning the confidence parameter in OFUL.

In the next section, we introduce our basic setup and notation for stochastic contexts. For the adversarial context case, further elements of the setup will be given in Section 5.6. Most of our proofs are provided Sections 5.7 and 5.8.

5.2 Problem Statement

We consider contextual sequential decision making problems described by a context space $\mathcal{X}$, an action space $\mathcal{A}$, and a policy space $\Pi = \{\pi : \mathcal{X} \to \mathcal{A}\}$. At each round t, a context $x_t \in \mathcal{X}$ is drawn[2] i.i.d. from some distribution, the learner observes this context, picks a policy $\pi_t \in \Pi$, thereby playing action $a_t = \pi_t(x_t) \in \mathcal{A}$, and receives an associated *reward* $r_t \in [0, 1]$ drawn from some fixed distribution $\mathcal{D}_{a_t,x_t}$ that may depend on the current action and context.

Base learners. Our learning policy in fact relies on base learner which are in turn learning algorithms operating in the same problem $\langle \mathcal{X}, \mathcal{A}, \Pi \rangle$. Specifically, there are M base learners which we index by $i \in [M] = \{1, \ldots, M\}$. In each round t, we select one of the base learners to play, and receive the reward associated with the action played by the policy deployed by that base learner in that round. Let us denote by $T_i(t) \subseteq \mathbb{N}$ the set of rounds in which learner i was selected up to time $t \in \mathbb{N}$. Then the pseudo-regret Reg_i our algorithm incurs over rounds $k \in T_i(t)$ due to the selection of base learner i is

$$\mathsf{Reg}_i(t) = \sum_{k \in T_i(t)} \left(\max_{\pi' \in \Pi} \mathbb{E}\left[r_k | \pi'(x_k), x_k\right] - \mathbb{E}\left[r_k | \pi_k(x_k), x_k\right] \right) , \tag{5.1}$$

and the total pseudo-regret Reg of our algorithm is then $\mathsf{Reg}(t) = \sum_{i=1}^{M} \mathsf{Reg}_i(t)$.

Candidate regret bounds. Each base learner i comes with a *candidate* regret (upper) bound $R_i \colon \mathbb{N} \to \mathcal{R}_+$, which is a function of the number of rounds this base learner has been played. This bound is typically known a-priori to us, and can also be random as long as the current value of the bound is observable, that is, we assume $R_i(n_i(t))$ is observable for all $i \in [M]$ and $t \in \mathbb{N}$, being $n_i(t) = |T_i(t)|$ the number of rounds learner i was played after t total rounds. Without loss of generality, we shall assume each candidate regret bound is non-decreasing, and increases by at most 1 from one play to the next, i.e.,

$$0 \le R_i(n) - R_i(n-1) \le 1 , \tag{5.2}$$

for all number of rounds $n \in \mathbb{N}$ and base learner $i \in [M]$, with $R_i(0) = 0$.

[2] This assumption will actually be relaxed in Section 5.6.

Well- and misspecified learners. We call learner i *well-specified* if $\mathsf{Reg}_i(t) \leq R_i(n_i(t))$ for all $t \in [T]$, with high probability over the involved random variables (see later sections for more details and examples), and otherwise *misspecified* (or *bad*). A well-specified base learning i is then one for which the candidate regret bound $R_i(\cdot)$ is a reliable upper bound on the actual regret of that learner.

For a given set of base learners and corresponding regret upper bounds, we denote the set bad learners by $\mathcal{B} \subseteq [M]$, and the set of well-specified ones by

$$\mathcal{W} = \{i \in [M] \colon \forall t \in [T] \ \mathsf{Reg}_i(t) \leq R_i(n_i(t))\} = [M] \setminus \mathcal{B} \ .$$

Notice that sets $\mathcal{W}$ and $\mathcal{B}$ are random sets. As a matter of fact, these sets do also depend on the time horizon T, but we leave this implicit in our notation. We assume in our regret-analysis that there is always a well-specified learner, that is $\mathcal{W} \neq \varnothing$. We will show that in the applications we consider, this happens with high probability. The index $i^\star \in \mathcal{W}$ (or just $\star$ in superscripts) will be used for any well-specified learner.

Consistent with the previous notation, we denote the total reward accumulated by base learner i after a total of t rounds as

$$U_i(t) = \sum_{k \in T_i(t)} r_k \ ,$$

and the total sum of rewards as $U(t) = \sum_{i \in [M]} \sum_{k \in T_i(t)} r_k$. The expected reward of the optimal policy at the context x_t at round t will be denoted by

$$\mu_t^\star = \max_{\pi' \in \Pi} \mathbb{E}\left[r | \pi'(x_t), x_t\right]$$

and, when contexts are stochastic, the expectation of $\mu_t^\star$ over contexts simply as $\mu^\star = \mathbb{E}_x\left[\mu_t^\star\right]$ which is a fixed quantity and independent of the round t.

Problem statement. Our goal is to perform model selection in this setting: We devise sequential decision making algorithms that have access to base learners as subroutines and are guaranteed to have regret that is comparable to the smallest regret bound among all well-specified base learners despite not knowing a-priori which base learners that are.

5.3 Regret Bound Balancing and Elimination

Our main algorithm follows the basic principle of regret bound *balancing*. The algorithm chooses the base learner in each round so as to make all presumed regret bounds evaluated at the number of rounds that the respective base learner was played to be roughly equal. To see why this achieves good total regret, assume for now all base learners are well-specified, so that they all satisfy their presumed regret bounds. Then, because the regret accrued by each base learner is bounded by its presumed regret bound, and these regret bounds are approximately equal, the total regret our algorithm incurs is at most M times worse than had we only played the algorithm with the best presumed regret bound:

$$\mathsf{Reg}(T) = \sum_{i=1}^{M} \mathsf{Reg}_i(T) \le \sum_{i=1}^{M} R_i(n_i(T)) \approx M \min_{i\in[M]} R_i(n_i(T)) \le M \min_{i\in[M]} R_i(T) \ .$$

Yet, the above only works if all base learners are well specified, which may not be the case. Besides, if we know all such learners are well specified, we could simply single out at the beginning of the game the learner whose regret bound is lowest at time T, and select that learner from beginning to end. Our task becomes more interesting in the presence of learners that may violate their presumed regret bound, when we do not know the identity of such learners. In this case, a reasonable goal for our policy would be to compete in the regret sense against the best well-specified base learner.

In order to handle this more involved situation, we pair the above regret bound balancing principle with a misspecification test to identify and eliminate misspecified base learners. This test compares the time-average rewards $U_i(t)/n_i(t)$ and $U_j(t)/n_j(t)$ achieved by two base learners i and j, and relies on the following concentration argument. While $U_i(t)$ is random and observable, the optimal average reward $\mu^\star$ is deterministic and unknown. We consider the event where, for each base learner i and each round t, the difference between $U_i(t)/n_i(t)$ and μ^* is close to the corresponding regret:

$$\mathcal{G} = \left\{ \forall i \in [M],\ \forall t \in \mathbb{N} \colon |n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t)| \le c\sqrt{n_i(t)\, \ln \frac{M \ln n_i(t)}{\delta}} \right\} .$$

We show in Lemma 5.7.1 in the appendix that for an appropriate absolute constant c, this event has probability $1-\delta$. This holds because, for each fixed t, $U_i(t)$ concentrates around $\sum_{k\in T_i(t)} \mathbb{E}\left[r_k|\pi_k(x_k), x_k\right]$, while $\sum_{k\in T_i(t)} \max_{\pi'\in\Pi} \mathbb{E}\left[r_k|\pi'(x_k), x_k\right]$ concentrates around $n_i(t)\,\mu^\star$, since contexts x_k are generated in an i.i.d. fashion. Now, since the pseudo-regret Reg_i cannot be negative by definition, the conditions defining $\mathcal{G}$ yield a lower-bound on $\mu^\star$ based on the rewards of each learner i :

$$\mu^\star \ge \frac{U_i(t)}{n_i(t)} - c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} \ . \tag{5.3}$$

When the provided regret bound $\mathsf{Reg}_i(t) \leq R_i(n_i(t))$ for learner i holds (that is, when i is well specified), then $\mathcal{G}$ also yields an upper-bound for $\mu^\star$:

$$\mu^\star \leq \frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \ . \tag{5.4}$$

Thus, if at any round t the upper bound for $\mu^\star$ from learner i contradicts the lower-bound from any other learner j,

$$\frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} < \frac{U_j(t)}{n_j(t)} - c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}},$$

then we conclude that the upper bound on μ^* provided by learner i is false, thereby showing that i is misspecified, and can safely be eliminated. Conversely, this also shows that no well-specified learner $i \in \mathcal{W}$ can be eliminated. Combining the elimination criterion with regret bound balancing yields our main algorithm, whose pseudocode is presented as Algorithm 12. The algorithm is an action elimination scheme that maintains over time a set $\mathcal{I}_t$ of active learners/actions at time t, and undergoes an elimination procedure as described above. The way base learner i_t is selected at each round guarantees the regret bound equalization we alluded to at the beginning of this section.

5.4 Regret Analysis

We first derive a general upper-bound on the regret of Algorithm 12 that depends on the ratios $\frac{n_i(t_i)}{n_\star(t_i)}$ of how often a learner i has been played compared to the best base learner. We will later bound this quantity for specific forms of candidate regret bounds R_i and provide simpler and more interpretable regret bounds.

Theorem 5.4.1. *With probability at least $1-\delta$, the total regret of Algorithm 12 is bounded for all rounds T as follows:*

$$\begin{aligned}\mathsf{Reg}(T) \leq \sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M \\ + 2c\sum_{i\in\mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \ ,\end{aligned} \tag{5.5}$$

where t_i is the last round where learner i passed the elimination test, $\star \in \mathcal{W}$ is any well-specified learner, and c is a universal positive constant.

Algorithm 12: Regret Bound Balancing and Elimination Algorithm (RBBE)

$\mathcal{I}_1 \leftarrow [M]$; `// set of active learners`

$U_i(0) = n_i(0) = 0$ for all $i \in [M]$

for *round* $t = 1, 2, \dots, T$ **do**

  Pick the base learner as $i_t \in \operatorname{argmin}_{i \in \mathcal{I}_t} R_i(n_i(t-1))$

  Play learner i_t and receive reward r_t

  Update base learner i with r_t

  Update $n_i(\cdot)$ and $U_i(\cdot)$:

  - $U_{i_t}(t) \leftarrow U_{i_t}(t-1) + r_t$

  - $n_{i_t}(t) \leftarrow n_{i_t}(t-1) + 1$

  $\mathcal{I}_{t+1} \leftarrow \mathcal{I}_t$

  foreach *active base learner* $i \in \mathcal{I}_t$ **do**

    Test for misspecification by checking

    $\frac{U_i(t)}{n_i(t)} + \frac{R_i(n_i(t))}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} < \max_{j \in \mathcal{I}_t} \frac{U_j(t)}{n_j(t)} - c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}}$

    if *above condition is triggered* **then**

      $\mathcal{I}_{t+1} \leftarrow \mathcal{I}_{t+1} \setminus \{i\}$

In order to prove this statement, we first show that Algorithm 12 indeed keeps all candidate regret bounds approximately equal (Lemma 5.4.2) and that the regret of any learner that has not been eliminated can be upper-bounded in terms of $R_\star(\cdot)$, the smallest regret upper bound among the well-specified learners (Lemma 5.4.3).

Lemma 5.4.2 (Regret Bound Balancing)**.** *In Algorithm 12, the regret bounds of all active learners are balanced at all times, i.e.,*

$$R_i(n_i(t)) \leq R_j(n_j(t)) + 1$$

for all $i, j \in \mathcal{I}_t$ *and* $t \in \mathbb{N} \cup \{0\}$.

Proof. At $t = 0$, the regret bound for all learners is 0 and the statement holds. For the sake of contradiction, assume now the claim is violated for the first time in round t, i.e., there is a $i, j \in \mathcal{I}_t$ such that $R_i(n_i(t)) > R_j(n_j(t)) + 1$. Then $i, j \in \mathcal{I}_{t-1}$ and i must have been played in round t. Further, by assumption on the candidate regret bounds

$$R_i(n_i(t-1)) \geq R_i(n_i(t)) - 1 > R_j(n_j(t)) = R_j(n_j(t-1)) \ ,$$

where the strict inequality follows from the violated claim and the equality holds because j was not played at time t. The resulting inequality $R_i(n_i(t-1)) > R_j(n_j(t-1))$ contradicts the claim that i was played at round t. □

Lemma 5.4.3. *In Algorithm 12 For any active learner $i \in \mathcal{I}_{t+1}$ and well-specified learner $\star \in \mathcal{W}$, the regret of i is bounded in event $\mathcal{G}$ as*

$$\mathsf{Reg}_i(t) \leq 1 + \left(\frac{n_i(t)}{n_\star(t)} + 1\right) R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}\ , \quad (5.6)$$

where c is a universal constant.

Proof. If $i \in \mathcal{I}_{t+1}$ remains active, then it must have passed the misspecification test in round t and satisfy, for all[3] $\star \in \mathcal{W}$,

$$\frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M\ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \geq \frac{U_\star(t)}{n_\star(t)} - c\sqrt{\frac{\ln(M\ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Subtracting $\mu^\star$ from both sides and rearranging terms gives

$$\mu^\star - \frac{U_i(t)}{n_i(t)} - c\sqrt{\frac{\ln(M\ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \leq \mu^\star - \frac{U_\star(t)}{n_\star(t)} + c\sqrt{\frac{\ln(M\ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Applying the definition of $\mathcal{G}$, we obtain an inequality in terms of pseudo-regrets:

$$\frac{\mathsf{Reg}_i(t)}{n_i(t)} - 2c\sqrt{\frac{\ln(M\ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \leq \frac{\mathsf{Reg}_\star(t)}{n_\star(t)} + 2c\sqrt{\frac{\ln(M\ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Multiplying both terms by $n_i(t)$ and rearranging terms gives

$$\mathsf{Reg}_i(t) \leq 2c\sqrt{\ln\frac{M\ln n_i(t)}{\delta}n_i(t)} + R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)}\mathsf{Reg}_\star(t) + 2c\sqrt{\ln\frac{M\ln n_\star(t)}{\delta}}\sqrt{\frac{n_i(t)}{n_\star(t)}}\ .$$

We now upper-bound the RHS by (i) replacing $\ln n_\star(t) \leq \ln t$ in the log-terms, (ii) using the fact that $\star \in \mathcal{W}$ is well-specified to replace the pseudo-regret $\mathsf{Reg}_\star(\cdot)$ by $R_\star(\cdot)$, and (iii) use the balancing condition from Lemma 5.4.2 to replace $R_i(n_i(t))$ by $R_\star(n_\star(t)) + 1$. This yields

$$\mathsf{Reg}_i(t) \leq 1 + \left(1 + \frac{n_i(t)}{n_\star(t)}\right) R_\star(n_\star(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\ ,$$

which is the claimed bound. □

[3]Recall that, under $\mathcal{G}$, any $\star \in \mathcal{W}$ will remain active.

We are now ready to prove Theorem 5.4.1.

Proof of Theorem 5.4.1. Let t_i be the last round where learner i passed the elimination test. Then the total regret can be bounded as

$$\mathsf{Reg}(T) = \sum_{i=1}^{M} \mathsf{Reg}_i(T) \leq \sum_{i \in \mathcal{W}} R_i(n_i(T)) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(t_i).$$

Applying Lemma 5.4.3 on $\mathsf{Reg}_i(t_i)$ for all $i \in \mathcal{B}$ and the balancing condition from Lemma 5.4.2 on the regret-bound for $i \in \mathcal{W}$ gives the desired bound. Finally, Lemma 5.7.1 in Section 5.7 shows that event $\mathcal{G}$ has probability at least $1 - \delta$. □

The general regret bound contained in Theorem 5.4.1 will be instantiated to more concrete cases for certain classes of candidate regret bounds. This will lead us to explicitly control the ratios $n_i(t_i)/n_\star(t_i)$. We do so in turn in the subsequent discussion.

Gap-Independent Regret Bounds

The regret guarantees in this section hold whenever there is a well-specified learner. These guarantees are independent of how much misspecified learners violate their presumed regret bounds ("gap" of the learner). In the next section, we will show that tighter guarantees can be achieved in cases where the gap is large, that is, when misspecified learners exceed their presumed bounds by a significant amount.

The first class of candidate regret bounds we consider is T^β with $\beta \in (0, 1]$. More concretely, each learner comes with a candidate regret bound of the form

$$R_i(n) = d_i C n^\beta \wedge n \ , \tag{5.7}$$

where $d_i \geq 1$ is some parameter and $C \geq 1$ is some term that does not depend on n or i. Note that the minimum with n is without loss of generality as any learner satisfies the regret bound n by our assumption on rewards being in $[0, 1]$. Consistent with our assumptions from Section 5.2, this minimum ensures that the regret bound can increase by at most 1 in each round. For candidate regret bounds of this form, we can show the following regret bound:

Theorem 5.4.4. *If Algorithm 12 is used with candidate regret bounds in Equation (5.7), then its total regret is bounded with probability at least $1 - \delta$ for all T as*

$$\mathsf{Reg}(T) \leq \left(M + 2B^{1-\beta} d_\star^{\frac{1}{\beta}-1} \right) d_\star C T^\beta + 5 d_\star^{\frac{1}{2\beta}} c \sqrt{BT \ln \frac{M \ln T}{\delta}} + 2M,$$

Presumed Bounds R_i	Regret Guarantee of Algorithm 12	Proof
$d_i C n^{1/3}$	$(M + B^{2/3} d_\star^2) d_\star C T^{1/3} + d_\star^{3/2}\sqrt{BT}$	Theorem 5.4.4
$d_i C n^{2/3}$	$(M + B^{1/3}\sqrt{d_\star}) d_\star C T^{2/3} + d_\star^{3/4}\sqrt{BT}$	Theorem 5.4.4
$d_i C \sqrt{n}$	$(M + \sqrt{B} d_\star) d_\star C \sqrt{T}$	Theorem 5.4.4
$d_i C \sqrt{n \ln \frac{n}{\delta}}$	$(M + \sqrt{B} d_\star) d_\star C \sqrt{T \ln \frac{T}{\delta}}$	Theorem 5.7.6
$\epsilon_i n + C\sqrt{n}$	$M(\epsilon_* T + C\sqrt{T}) + MC^2$	Theorem 5.4.5

Table 5.1: Summary of our gap-independent regret guarantees In all bounds but the one in the 4th line, log factors are omitted for readability. In green is the regret guarantee of the best well-specified learner.

where $\star \in \mathcal{W}$ is any well-specified learner and $B = |\mathcal{B}|$ is the number of misspecified learners.

The first three entries in Table 5.1 summarize this result in the relevant cases where $\beta = \frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$. When $\beta \geq 1/2$, our regret bound can recover the best T^β rate. In particular, the bound of Theorem 5.4.4 recovers the regret bound guarantee of the best well-specified learner up to a multiplicative factor of the form $M + B^{1-\beta} d_\star^{\frac{1}{\beta}-1}$. On the other hand, when $\beta < 1/2$ our bound scales sub-optimally as $\sqrt{T}$. This is not surprising since the lower bound by [54] indicates a $\Omega(\sqrt{T})$ barrier for model-selection based on observed rewards without additional assumptions.

We further show in the appendix that this result can be generalized to the case where the candidate regret bounds scale with additional logarithmic factors in the number of rounds, e.g. $\sqrt{n \ln n}$ as opposed to just $\sqrt{n}$ – see Theorem 5.7.6 in Section 5.7.

We defer the full proof of Theorem 5.4.4 to Section 5.7, but provide a brief sketch of the main argument for the special case of $\beta = \frac{1}{2}$. The general case follows analogously.

Proof sketch of Theorem 5.4.4. The first term of the general regret bound from Theorem 5.4.1 can be written as $\sum_{i=1}^{M} R_\star(n_\star(t_i)) \leq M R_\star(T) \leq M C d_\star \sqrt{T}$, the first inequality using the monotonicity of the candidate regret bound. This yields the first term in Theorem 5.4.4. The second term in Theorem 5.4.1 can be controlled as follows:

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) \overset{(i)}{\leq} \sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} C d_\star \sqrt{n_\star(t_i)} = C d_\star \sum_{i\in\mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i)}$$
$$\overset{(ii)}{\leq} C d_\star \sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{\sum_{i\in\mathcal{B}} n_i(t_i)} \overset{(iii)}{\leq} C d_\star \sqrt{2Bd_\star^2} \sqrt{t_i} \leq C d_\star^2 \sqrt{2BT} \ ,$$

where step (i) applies the definition of the candidate regret bound, step (ii) uses Cauchy-Schwarz inequality and step (iii) follows from the fact that the total number of plays at round t_i is t_i and a bound on the sum of play ratios $\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \leq 2Bd_\star^2$, which we will show below. This yields the second term in the desired regret bound. The remaining terms can handled in a similar manner.

To derive the bound on the play ratios, consider first the case where $n_i(t_i)$ is so large that $R_i(n_i(t_i)) < n_i(t_i)$. Then, by the balancing condition from Lemma 5.4.2,

$$d_i C \sqrt{n_i(t_i)} = R_i(n_i(t_i)) \leq R_\star(n_\star(t_i)) + 1 \overset{(iv)}{\leq} 2R_\star(n_\star(t_i)) \leq 2d_\star C \sqrt{n_\star(t_i)},$$

where (iv) holds because no learner can be eliminated before each learner has been played at least once and thus $R_\star(n_\star(t_i)) \geq 1$. Rearranging this inequality yields $n_i(t_i)/n_\star(t_i) \leq 2d_\star^2/d_i^2$. Analogously, we can show that if $n_i(t_i)$ satisfies $n_i(t_i) = R_i(n_i(t_i))$, then $n_i(t_i)/n_\star(t_i) \leq 2$. This follows from $n_i(t_i) = R_i(n_i(t_i)) \leq 2R_\star(n_\star(t_i)) \leq 2n_\star(t_i)$. Thus, the sum of play ratios is bounded as

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \leq \sum_{i\in\mathcal{B}} 2 \vee 2\frac{d_\star^2}{d_i^2} \leq 2Bd_\star^2 \ .$$

□

Linear regret base learners. When we instantiate Theorem 5.4.4 to the case where candidate regret bounds are linear in n ($\beta = 1$), then the total regret of Algorithm 12 is of order $\widetilde{O}(MCd_\star T)$, which is only a factor M worse than the regret bound for the best well-specified learner. The follow result shows that this is still the case when candidate regret bounds come with an additional $\sqrt{n}$ term common to all learners under the additional assumption that no misspecified algorithm has a larger candidate regret bound than the best well-specified learner. This will be useful when Algorithm 12 is used with contextual bandits or linear MDP algorithms with misspecified function classes (see Section 5.5).

Theorem 5.4.5. *Let the candidate regret bounds for all M base learners be of the form*

$$R_i(n) = C_1\sqrt{n} + \epsilon_i C_2 n \wedge n, \tag{5.8}$$

where $\epsilon_i \in (0, 1]$ and $C_1, C_2 > 1$ are quantities that do not depend on ϵ_i or n. Then, probability at least $1 - \delta$, the total regret of Algorithm 12 is bounded for all rounds T as

$$\mathsf{Reg}(T) = O\left(MC_1\sqrt{T}\sqrt{\ln\frac{M\ln T}{\delta}} + MC_2\epsilon_* T\sqrt{\ln\frac{M\ln T}{\delta}} + BC_1^2\right) ,$$

where $* \in \mathcal{W}$ is any well-specified base learner such that $\epsilon_i \leq \epsilon_*$ for all misspecified learners $i \in \mathcal{B}$.

Proof. This statement is proven analogously to the generic bound in Theorem 5.4.4, but it makes heavy use of a case-by-case analysis of the different regimes of candidate regret bounds provided in Lemma 5.7.9 in Section 5.7. □

Gap-Dependent Regret Bounds

The regret guarantees in the previous section only depend on which learners are well- or misspecified and their presumed regret bounds. In particular, a misspecified learner may violate their presumed regret bound at any time by any amount. However, in many relevant practical cases, a base learner is either well-specified or violates their presumed regret bound by a significant amount. For example in contextual bandits where each base learner has access to a restricted policy class, a learner achieves good $\sqrt{T}$ regret when the optimal policy is contained in its policy class, but has otherwise to suffer linear regret. We now provide tighter guarantees for Algorithm 12 in such cases. Specifically, we assume that if a learner j is misspecified, its regret is lower-bounded by

$$\mathsf{Reg}_j(t) \geq \Delta_j n_j(t)^\alpha$$

for all t, where $\Delta_j > 0$ and α is strictly larger than both $\frac{1}{2}$ and the presumed regret rate β in Eq. (5.7). Since the regret of j grows significantly faster than its presumed regret bound and the regret of the best well-specified learner (that is, Reg_j has a large *gap*), we can show that the elimination test in Algorithm 12 is triggered after playing learner i for a certain number of times. This allows us to prove the following gap-dependent regret-guarantee:

Theorem 5.4.6. *Assume Algorithm 12 is used with candidate regret bounds in Equation (5.7) and that the pseudo-regret of all misspcified learners $j \in \mathcal{B}$ is bounded for all t from below as $\mathsf{Reg}_j(t) \geq \Delta_j n_j(t)^\alpha$, for some constants $\Delta_j > 0$ and $\alpha > \frac{1}{2} \vee \beta$. If $0 < \beta < \frac{1}{2}$ then total regret is bounded with probability at least $1 - \delta$ for all T as*

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i \in \mathcal{B}} C\left((2d_\star)^{\frac{1}{\beta}+\frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}}\right)\left[\frac{20C}{\Delta_i}\ln\frac{M\ln T}{\delta}\right]^{\frac{1}{2\alpha-1}}\right),$$

where $\star \in \mathcal{W}$ is any well-specified learner. If instead $\beta \geq \frac{1}{2}$, then the total regret is bounded with probability at least $1 - \delta$ for all T as

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i \in \mathcal{B}} C\sqrt{\ln\frac{M\ln T}{\delta}}\left(d_\star^{\frac{1}{\beta}+\frac{1}{\alpha-\beta}} + d_\star d_i^{\frac{\beta}{\alpha-\beta}}\right)\left[\frac{20C}{\Delta_i}\right]^{\frac{\beta}{\alpha-\beta}}\right).$$

Although the argument of eventually eliminating base learners with a large gap is similar to a gap-dependent analysis is multi-armed bandits, it is important to note that the notion of gap here is a property of the base learner and not (necessarily) of the action space at hand.

Table 5.2 contains a summary of the guarantees in Theorem 5.4.6 for the special case where $\alpha = 1$ and $\beta = \frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$. Comparing Theorem 5.4.6 to Theorem 5.4.4 (or Table 5.2 to Table 5.1), we see that the multiplicative factor in front of the best well-specified regret bound is only M, as compared to the presence of extra $d_\star$ factors without a gap-assumption. Further, while the additive term in Table 5.2 may have a dependency on a potentially large d_i, this term only scales with T as $\ln\ln T$, and is thus virtually constant. Importantly, this yields the optimal scaling in T even when $\beta < \frac{1}{2}$ (see the first line of Table 5.2) so that the additional $\sqrt{T}$-term occurring in Table 5.1 can be avoided. This result is in contrast with existing approaches such as [54], where the $\sqrt{T}$ dependence cannot be avoided.

5.5 Applications of RBBE

Let's start by reviewing the OFUL Algorithm.

Brief Review of Contextual Linear Bandits and the OFUL Algorithm

One important application of the methods we presented in Section 5.3 and Section 5.4 is the setting of contextual linear bandits, which we now briefly review. To keep

Bounds R_i	Gap-Dependent Regret Guarantee of Algorithm 12	Theorem
$d_i C n^{1/3}$	$M d_\star C T^{1/3} + \sum_{i\in\mathcal{B}} \frac{C^2(d_\star^6 + d_\star d_i)}{\Delta_i} \ln \frac{M \ln T}{\delta}$	5.4.6
$d_i C n^{2/3}$	$M d_\star C T^{2/3} + \sum_{i\in\mathcal{B}} \frac{C^3(d_\star^{4.5} + d_\star d_i^2)}{\Delta_i^2} \sqrt{\ln \frac{\ln T}{\delta}}$	5.4.6
$d_i C \sqrt{n}$	$M d_\star C \sqrt{T} + \sum_{i\in\mathcal{B}} \frac{C^2(d_\star^4 + d_\star d_i)}{\Delta_i} \sqrt{\ln \frac{\ln T}{\delta}}$	5.4.6
$d_i C \sqrt{n \ln \frac{n}{\delta}}$	$M d_\star C \sqrt{T \ln \frac{T}{\delta}} + \sum_{i\in\mathcal{B}} \frac{C^2(d_\star^4 + d_\star d_i)}{\Delta_i} \ln^{3/4} \frac{MT}{\delta} \ln^{3/2} \frac{\ln T}{\delta}$	5.7.8

Table 5.2: Summary of our gap-dependent regret bounds when each misspecified learner has linear pseudo-regret ($\alpha = 1$). Some constant factors are omitted for readability. In green is the regret guarantee of the best well-specified learner.

consistency with previous sections, we shall assume here that contexts are drawn i.i.d. from some distribution over context space $\mathcal{X}$. Yet, the algorithmic solutions we present (specifically, the OFUL algorithm) actually work unchanged even in the more general fixed design or adaptive design scenarios. This will be useful in Section 5.6, when dealing with the *adversarial* contextual bandit setting.

In the contextual bandit setting, context x_t determines the set of actions $\mathcal{A}_t \subseteq \mathcal{A}$ that can be played at time t. When the bandit setting is *linear* the policies we consider are of the form $\pi_\theta(x_t) = \arg\max_{a\in\mathcal{A}_t} \langle a_t, \theta\rangle$, for some $\theta \in \mathbb{R}^d$, and the class of policies Π can then be thought of as a class of d-dimensional vectors $\Pi \subseteq \mathbb{R}^d$. Moreover, rewards are generated according to a noisy linear function, that is, $r_t = \langle a_t, \theta_*\rangle + \xi_t$, where $\theta_* \in \Pi$ is unknown, and ξ_t is a conditionally zero mean $\sigma-$subgaussian random variable. We denote the time-t optimal action as $a_t^\star = \text{argmax}_{a\in\mathcal{A}_t}\langle a, \theta_\star\rangle$. The learner's objective is to control its pseudo-regret:

$$\mathsf{Reg}(T) = \sum_{t=1}^{T} \langle a_t^\star, \theta_\star\rangle - \langle a_t, \theta_\star\rangle \ .$$

OFUL Algorithm. We now recall the relevant components of the OFUL algorithm [1] shown in Algorithm 13. Instances of this algorithm will play the role of base learners in subsequent sections. The OFUL algorithm proceeds by computing a regularized least-squares (RLS) estimator $\hat{\theta}_t$ of the true parameter $\theta_\star$ using the data

Algorithm 13: OFUL [1]

Input: regularization parameter $\lambda > 0$, confidence scaling $\beta_1, \beta_2, \ldots$
for *round* $t = 1, 2, \ldots$ **do**
  Update regularized least-squares estimator $\hat{\theta}_t$ and covariance matrix Σ_t
  Receive context x_t/action space $\mathcal{A}_t$
  Play optimistic action:

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle = \operatorname*{argmax}_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle + \beta_t \|a\|_{\Sigma_t^{-1}}$$

  Receive reward $r_t = \langle a_t, \theta_\star \rangle + \xi_t$.

collected so far:

$$\hat{\theta}_t := \Sigma_t^{-1} \left(\sum_{\ell=1}^{t-1} a_\ell r_\ell \right) \quad \text{where} \quad \Sigma_t = \lambda \mathbb{I} + \sum_{\ell=1}^{t-1} a_\ell a_\ell^\top \ . \tag{5.9}$$

Here, Σ_t is the regularized covariance matrix of the played actions up to the beginning of round t with regularization parameter λ, and $\mathbb{I}$ denotes the $d \times d$ identity matrix. Using $\hat{\theta}_t$ and Σ_t, OFUL proceeds by computing a confidence ellipsoid

$$\mathcal{C}_t := \{\theta \ : \ \|\theta - \hat{\theta}_t\|_{\Sigma_t} \le \beta_t\} \tag{5.10}$$

that should contain the optimal parameter $\theta_\star$. We will discuss a choice of the (possibly data-dependent) scaling factor $\beta_t \in \mathcal{R}_+$ below that ensures that this happens in all rounds with high probability. Algorithm 13 now plays any action that achieves highest expected return in what we refer to as the optimistic model

$$\widetilde{\theta}_t = \operatorname*{argmax}_{\theta \in \mathcal{C}_t} \max_{a \in \mathcal{A}_t} \langle a, \theta \rangle \ . \tag{5.11}$$

This choice of action is equivalent to picking $a_t \in \operatorname{argmax}_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle + \beta_t \|a\|_{\Sigma_t^{-1}}$.

We define the event that the above-mentioned ellipsoidal confidence set $\mathcal{C}_t$ contains θ^* at all times $t \in \mathbb{N}$ as

$$\mathcal{E} = \{\theta_* \in \mathcal{C}_t, \quad \forall t \in \mathbb{N}\} \ . \tag{5.12}$$

In this event $\mathcal{E}$, the optimistic model $\widetilde{\theta}_t$ indeed gives rise to an optimistic estimate of the expected reward in each round

$$\langle a_t, \widetilde{\theta}_t \rangle \ge \max_{a \in \mathcal{A}_t} \langle a, \theta_\star \rangle = \langle a_t^\star, \theta_\star \rangle \ . \tag{5.13}$$

[1] show that the following choice for β_t is sufficient to make $\mathcal{E}$ happen with high probability:

Lemma 5.5.1 (Theorem 1 in [1]). *For any $\delta \in (0,1)$, let the confidence scaling be*

$$\beta_t := \sqrt{2\sigma^2 \ln\left(\frac{\det(\Sigma_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S \leq \sqrt{\sigma^2 d \ln\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S \tag{5.14}$$

where S is a known bound on the parameter norm $\max_{\theta\in\Pi}\|\theta\|_2$ and L is a known bound on the action norm in all rounds, i.e., $\max_{a\in\mathcal{A}_t}\|a\|_2 \leq L$ for all t. Then $\theta_\star$ is contained in the confidence ellipsoid with high probability, i.e., $\mathbb{P}(\mathcal{E}) \geq 1-\delta$.

In event $\mathcal{E}$, one can show that the regret of Algorithm 13 is bounded for all $t \in [T]$ as

$$\mathsf{Reg}(t) \leq 2\beta_{\max}\sqrt{dt\left(1+\frac{L^2}{\lambda}\right)\ln\frac{d\lambda+tL}{d\lambda}}\ ,$$

where $\beta_{\max} = \max_{k\in[t]}\beta_k$. We reproduce a slightly more general version of the standard proof for this regret bound in Lemma 5.9.1 in the appendix. The right side of the above inequality will play the role of our presumed regret bound $R(n_i(t))$ when OFUL is used as a base learner.

In the rest of this section, we present a number of applications of our balancing and elimination machinery to the case where the base learners are instances of the OFUL algorithm.

Linear Bandits with Nested Model Classes

We can apply our regret bound balancing algorithm to linear bandits where the true dimensionality $d_\star$ of the model $\theta_\star$ is unknown a-priori. In this standard scenario, considered by many recent papers in the model selection literature for bandit algorithms [e.g. 27, 54], the learner chooses among actions $\mathcal{A}_t \subseteq \mathcal{R}^{d_{\max}}$ of dimension $d_{\max}$ but only the first $d_\star$ dimensions are relevant (that is, $(\theta_\star)_i = 0$ for $i > d_\star$).

One can learn in this setting as follows: We use $\log_2 d_{\max}$ instances of OFUL as base learners[4]. Each instance i first truncates the actions to dimension $d_i = 2^i$ and then only computes the least-squares estimate and confidence ellipsoid in $\mathcal{R}^{d_i}$. Based

[4]We here assume that $d_\star$ and $d_{\max}$ are powers of 2 for convenience but our results also hold generally up to a constant factor of 2.

on the OFUL regret guarantees in the previous section, we use $R_i(n) = d_i C\sqrt{n} \wedge n$ as putative regret bounds, with constant C set to

$$C = 2\left(\sigma + \sqrt{\lambda}S\right)\sqrt{\left(1 + \frac{L^2}{\lambda}\right)\ln\left(\frac{1 + TL^2/\lambda}{\delta}\right)\ln\frac{\lambda + TL}{\lambda}} .$$

For convenience, we here assume the time horizon T is known and $\ln T$ terms can therefore be absorbed into the constant C common to all base learners, but any-time versions are also possible by setting $n = T$ above at which the regret bound scales as $\sqrt{n}\ln n$ (see Theorem 5.7.6 in appendix). By the regret guarantee of OFUL discussed in the previous section, with probability $1 - M\delta$, any base learner i such that $d_i < d_\star$ will be misspecified, while all remaining i are well specified.

More specifically, we have $M = O(\ln d_{\max})$-many base learners, out of which $B = O(\ln d_\star)$ are misspecified. Then a direct application of Theorem 5.4.4 with $\beta = 1/2$ gives

$$\mathsf{Reg}(T) = O\left(\left(\ln d_{\max} + d_\star\sqrt{\ln d_\star}\right) d_\star C\sqrt{T}\right) \approx O\left(\left(\ln d_{\max} + d_\star\sqrt{\ln d_\star}\right) d_\star\sqrt{T}\ln T\right),$$

where the second expression only retains dependencies on T, $d_\star$ and $d_{\max}$.

If further all misspecified learners suffer linear regret $\mathsf{Reg}_i(t) \geq \Delta n_i(t)$ for some $\Delta > 0$ (e.g. since they cannot represent the observed rewards, they may converge to playing a strictly suboptimal action for most contexts), then applying Theorem 5.4.6 yields

$$\begin{aligned}\mathsf{Reg}(T) &= O\left(\ln(d_{\max})d_\star C\sqrt{T} + \ln(d_\star)\frac{C^2 d_\star^4}{\Delta}\sqrt{\ln\frac{\ln T}{\delta}}\right)\\ &\approx O\left(\ln(d_{\max})d_\star\sqrt{T}\ln T + \frac{d_\star^4\ln d_\star}{\Delta}(\ln T)^2\sqrt{\ln\ln T}\right) ,\end{aligned}$$

where the second expression again only shows dependencies on T, $d_\star$, $d_{\max}$ and Δ. Notice that, as T grows large, the main term of the above bound becomes $d_\star\sqrt{T}$, up to log factors. This is precisely the bound we would achieve had we known in advance dimension $d_\star$, and just played the associated base OFUL from beginning to end.

Remark 5.5.2. *A standard goal in model selection is to obtain sub-linear regret bounds even in the case where the model complexity of the target class is allowed to grow sub-linearly with T – see, e.g., the discussion in [27]. In our case, this would be obtained by regret bounds of the form $d_\star^\alpha T^{1-\alpha}$, for some $\alpha \in (0,1)$, for example a bound of the form $\sqrt{d_\star T}$. It is worth observing that in the setting considered in this chapter this is an impossible goal to achieve since, unlike [27], we are dealing with* infinite *action spaces, and the best one can hope for in this case is indeed $d_\star\sqrt{T}$ (see Section 2 in [56]).*

Linear Markov Decision Processes with Nested Model Classes

We can instantiate the regret bound in Theorem 5.4.4 ($\beta = 1/2$) to the episodic linear MDP setting of [38], again with nested feature classes of doubling dimension. Here, each round t of Algorithm 12 corresponds to one episode of H time steps in the MDP, and contexts x_t are the initial state of the episode in the MDP. [38] prove that their LSVI-UCB algorithm achieves regret $O(H^2\sqrt{d^3K}\ln(dK/\delta))$ after K episodes when used with a realizable function class of dimension d. We deploy $M = O(\ln d_{\max})$ instances of LSVI-UCB as base learners with presumed regret bounds

$$R_i(n) = Hn \wedge H^2\sqrt{d_i^3 n}\ln(d_{\max}T/\delta).$$

Since the total reward per episode (= round) is in $[0, H]$ instead of $[0, 1]$ in this setting, we scale the regret bound as well as the constant c in Algorithm 12 by H. By Theorem 5.4.4 the total regret of Algorithm 12 after T episodes is bounded as

$$\mathsf{Reg}(T) = O\left(\left(\sqrt{d_\star^3 \ln d_\star} + \ln d_{\max}\right) H^2\sqrt{d_\star^3 T}\ln(d_{\max}T/\delta)\right)$$

with probability $1 - M\delta$. Similar to the contextual bandit setting above, we can achieve a tighter bound if all misspecified learners suffer linear regret $\mathsf{Reg}_i(t) \geq \Delta n_i(t)$ for some $\Delta > 0$. Then applying Theorem 5.4.6 yields

$$\mathsf{Reg}(T) = O\left(H^2\sqrt{d_\star^3 T}\ln(d_{\max})\ln(d_{\max}T/\delta) + \frac{H^4 d_\star^6}{\Delta}\ln(d_{\max}T/\delta)^2\sqrt{\ln\frac{\ln T}{\delta}}\right)$$

which, up to log factors and lower order terms, again coincides with the regret bound of the best base learner in hindsight.

Linear Bandits and MDPs with Unknown Approximation Error

[69] presents an algorithm for learning a good policy in episodic MDPs where the value functions are all close to a linear feature space of dimension d. Their algorithm admits a high-probability regret bound of order[5] $\widetilde{O}(Hd\sqrt{T} + H\sqrt{d}\epsilon T)$ for all T when a bound ϵ on the inherent Bellman error is known a-priori. For details of the setting and the exact definition of inherent Bellman error see [69]. Unfortunately, in most

[5]The $\widetilde{O}$ notation is similar to the O-notation but hides poly-logarithmic dependencies.

practical applications, one does not know ϵ ahead of time and picking a conservative value (large ϵ) makes the algorithm over-explore and suffer large regret.

We can address this limitation by applying Algorithm 12 with several instances of their algorithm as base-learners, each associated with a certain value of the inherent Bellman error $\epsilon_i = \frac{2^{1-i}}{\sqrt{d}}$ and the putative regret bound $R_i(n) = (CHd\sqrt{n} + CH\sqrt{d}\epsilon_i n) \wedge Hn$ for an appropriate value C that depends at most logarithmically on d, T or H. It is sufficient to use $M = \lceil 1 + \frac{1}{2}\log_2(T/d^2)\rceil$ base learners since the putative regret bound of learner 1 (with $\epsilon_1 = 1/\sqrt{d}$ and $R_1(n) \geq Hn$) always holds, while the putative regret bound of learner M is at most $R_M(T) \leq 2CHd\sqrt{T}$, which is a constant factor worse than the regret when $\epsilon = 0$.

By Theorem 5.4.5, the total regret of Algorithm 12 with these base learners is

$$\begin{aligned}\mathsf{Reg}(T) &= O\left(MCH(d\sqrt{T} + \sqrt{d}\epsilon_\star T)\sqrt{\ln\frac{M\ln T}{\delta}} + BC^2H^2d^2\right)\\ &= \widetilde{O}\left(Hd\sqrt{T} + H\sqrt{d}\epsilon_\star T + H^2d^2\right)\end{aligned}$$

with probability $1 - M\delta$. Hence, up to at most logarithmic factors and a lower-order additive term, our model-selection framework can recover the best regret bound without requiring knowing the inherent Bellman error ahead of time. Notice also that the special case $H = 1$ recovers the standard linear bandit setting and the algorithm by [69] reduces to OFUL with a confidence ellipsoid that accounts for ϵ_i. In this bandit case $\epsilon_\star$ is the absolute approximation error of expected rewards.

Recently, [26] have shown that an adaptation to unknown approximation errors $\epsilon_\star$ is possible in contextual bandits, but their model-selection approach requires base learners that work with importance weights, and whose importance-weighted regret admits a favorable dependency on ϵ_i. Here we have shown that a similar result (up to logarithmic factors) can be achieved with standard optimistic base learners such as OFUL. Our result also matches the regret-guarantee by [54] but does not require their smoothing procedure for base-learners. Importantly, our result proves that an adaptation to unknown approximation errors $\epsilon_\star$ is also possible without any modification to base learners in the MDP setting where base-learners that achieve the importance-weighted regret guarantee required by [26] are (still) unavailable. Note also that our framework is not specific to instances of the algorithm by [69] as base learners. Our model selection algorithm can, for example, also be used with approximate versions of LSVI-UCB by [38] and achieve similar regret guarantees in their setting and for their notion of approximation error.

Confidence parameter tuning in OFUL

A standard problem that arises in the practical deployment of contextual bandit algorithms like OFUL is that they are extremely sensitive to the tuning of their upper-confidence parameter ruling the actual trade-off between exploration and exploitation. The choice of confidence parameter from Lemma 5.5.1 ensures high-probability regret guarantee but is often too conservative. This can for example be the case when the actual noise variance is smaller than the assumed σ^2 variance. While there are concentration results (empirical Bernstein bounds) that can adapt to such fortunate low-variance noise for scalar parameters (e.g., in unstructured multi-armed bandits), such adaptive bounds are still unavailable for least-squares estimators. Empirically, choosing smaller values for $\beta_1, \dots, \beta_T$ can often achieve significantly better performance but comes at the cost of losing any theoretical performance guarantee. Our model-selection framework can be used to tune the confidence parameter online and simultaneously achieve a regret guarantee.

We will now look at ways to compete against the instance of the OFUL algorithm which is equipped with the optimal scaling of its upper-confidence value, in the sense of the following definition:

Definition 5.5.3. *Denote by $\bar{\beta}_t$ the confidence-parameter choice from Lemma 5.5.1 and let $\kappa \in \mathcal{R}_+$ be a scaling factor. Further, let $\hat{\theta}_S(\kappa)$ and $\Sigma_S(\kappa)$ be the iterates of least squares estimator and covariance matrix obtained by running* OFUL *with scaled confidence parameters $(\kappa\bar{\beta}_t)_{t\in\mathbb{N}}$ on a subset of rounds $S \subseteq [T]$. Then, for a given range $[\kappa_{\min}, 1]$, the optimal confidence parameter scaling for* OFUL *is defined as*

$$\kappa_\star = \min_{\kappa\in[\kappa_{\min},1]} \max_{S\subseteq[T]} \frac{\|\hat{\theta}_S(\kappa) - \theta_\star\|_{\Sigma_S(\kappa)^{-1}}}{\bar{\beta}_{|S|}} .$$

In words, the optimal $\kappa_\star$ is the smallest scaling factor of confidence parameters that ensures that no matter to what subset of rounds we would apply OFUL to, the optimal parameter $\theta_\star$ is always contained in the confidence ellipsoid. Observe that $\kappa_\star$ is a random quantity, i.e., $\kappa_\star$ is the best scaling factor for the given realizations in hindsight. Lemma 5.5.1 ensures that $\mathbb{P}(\kappa_\star \leq 1) \geq 1 - \delta$ and empirical observations suggest that $\kappa_\star$ is much smaller in many events and bandit instances.

Now, Lemma 5.9.1 in Section 5.9 ensures that OFUL with confidence parameters $\kappa\bar{\beta}_t$ admits a regret bound of the form[6] $\mathsf{Reg}(n) \lesssim \kappa d\sqrt{n}\ln(n) \wedge n$ if $\kappa \geq \kappa_\star$. Since $\kappa_\star$ is unknown, we run Algorithm 12 with M instances of OFUL as base learners,

[6] For simplicity of presentation, we set here $\lambda = 1$ and disregarded the dependence on other parameters like L, S, and σ.

each with a scaling factor $\kappa_i = 2^{1-i}$, $i = 1, \ldots, M$, and putative regret bound $R_i(n) \approx \kappa_i d \ln(T)\sqrt{n} \wedge n$. Note that it is sufficient to use $M = 1 + \log_2 \frac{1}{\kappa_{\min}}$.

Then, by Theorem 5.4.4 (with $\beta = 1/2$ therein), the regret of Algorithm 12 is bounded with probability at least $1 - \delta$ as

$$\begin{aligned}\mathsf{Reg}(T) &\lesssim \left(M + \sqrt{B}\frac{\kappa_i}{\kappa_{\min}}\right) R_\star(T) \\ &= O\left(\left(\frac{\kappa_\star}{\kappa_{\min}}\sqrt{\ln\frac{\kappa_\star}{\kappa_{\min}}} + \ln\frac{1}{\kappa_{\min}}\right)\kappa_\star d \ln(T)\sqrt{T}\right).\end{aligned}$$

Note that this is a random and problem-dependent bound because so is $\kappa_\star$. In cases where $\kappa_\star \lesssim \sqrt{\frac{\kappa_{\min}}{\ln(1/\kappa_{\min})}}$, this bound strictly improves on the standard OFUL bound relying on confidence scaling $\kappa = 1$, which is often way too conservative in practice.

5.6 Extension to Adversarial Contexts

In this section, we show that the regret balancing and elimination principle can also be used for model selection when the contexts x_t are generated in an adversarial manner. This requires slightly stronger assumptions on the base learners, which hold in many settings when we select between a hierarchy of optimistic learners such as OFUL or LSVI-UCB. For the sake of concreteness, we present our extension of the regret balancing and elimination algorithm to adversarial contexts for the setting, but our technique for adversarial contexts can be easily adapted to all other bandit applications discussed in Section 5.5 and likely to episodic MDP settings with adversarial start states as well.

Let us briefly recall the setting of Linear Bandits with Nested Model Classes. We consider the problem of linear bandits and are given M instances of OFUL as base learners. Each instance i considers only on the first $d_i = 2^i$ dimensions of the actions, with $d_1 < d_2 < \cdots < d_M$. Since the entries of the true parameter $\theta_\star$ are 0 for all dimensions above $d_{i_\star}$, where $i_\star \in [M]$ is an unknown index, all learners $i_\star, i_\star + 1, \ldots M$ are well-specified with high probability. We focus our analysis on the event $\mathcal{E}$ where this is the case. Unlike in the preceding discussion on Linear Bandits with Nested Model Classes where contexts are assumed to be drawn i.i.d., we here consider the setting where contexts x_t (corresponding to the action set $\mathcal{A}_t$ at round t) are generated adversarially. Since each base learner operates only in a lower-dimensional subspace, we allow the bounds on the action norm L_i, the bound on the parameter norm S_i and the range of expected return $R_i^{\max}$ to vary per base learner i (potentially depend on the number of dimension d_i) but for the sake of simplicity, we assume that all learners use regularization parameter $\lambda = 1$.

Algorithm 14: EpochBalancing

```
Input: set of learners I
for round t = 1, 2, ... do
    Receive context x_t
    foreach learner i ∈ I do
        Ask learner i for a lower bound B_{t,i} on the value of its proposed action
    Sample i_t ~ p ∝ 1/z_i for i ∈ I                    (see Equation (5.15))
    Play learner i_t and receive reward r_t
    Update base learner i_t with r_t
    Test for misspecification by checking
        sum_{i∈I} [U_i(t) + R_i(n_i(t))] + c sqrt(t ln(ln(t)/δ)) < max_{i∈I} sum_{k=1}^{t} B_{k,i}
    if above condition is triggered then
        Return ;        // At least one learner must be misspecified
```

$$\sum_{i\in\mathcal{I}}[U_i(t)+R_i(n_i(t))]+c\sqrt{t\ln\frac{\ln(t)}{\delta}} < \max_{i\in\mathcal{I}}\sum_{k=1}^{t}B_{k,i}$$

Algorithm 12, which assumes stochastic contexts, compares upper- and lower confidence bounds on the optimal return value $\mu^\star$ obtained from learners that were executed on two disjoint subsets of rounds to determine misspecification. This strategy does not work with adversarial contexts since the optimal policy that an algorithm could have achieved depends on the contexts in the rounds that it was played. One algorithm may only have seen "bad" contexts with low $\mu_t^\star$, while another may only encountered favorable contexts with high $\mu_t^\star$. A direct comparison is therefore meaningless.

To be able to handle adversarial contexts and address this challenge, we modify our regret balancing and elimination algorithm in two ways: (1) we randomize the learner choice for regret balancing and (2) we change the misspecfication test to compare upper and lower confidence bounds on the optimal policy value of *all* rounds played to far. The resulting algorithm is presented in Algorithm 15 which operates in epochs where the subroutine in Algorithm 14 is executed. We start by discussing the regret balancing subroutine in the next section before presenting the main algorithm and its regret guarantee afterwards.

The Epoch Balancing Subroutine

This subroutine in Algorithm 14 takes in input a set of active base learners $\mathcal{I} = \{s, s+1, \dots, M\}$ and ensures by randomized regret bound balancing that its total regret is controlled for all rounds until it terminates.

In addition to the putative bound R_i on its regret, Algorithm 14 requires that each learner i can also provide a lower-confidence bound on $\mathbb{E}[r_t|a_{t,i}, x_t]$, the expected reward of the action it would play in the current context x_t. Since each base learner i is an instance of OFUL, we can choose these bounds at round t as

$$R_i(n_i(t)) = 2 \sum_{k \in T_i(t)} \left(\beta_{k,i} \|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge R_i^{\max}\right) \qquad \text{and}$$
$$B_{t,i} = \left(\langle \widehat{\theta}_{t,i}, a_{t,i}\rangle - \beta_{t,i}\|a_{t,i}\|_{\Sigma_{t,i}^{-1}}\right) \vee -R_i^{\max}$$

where $R_i^{\max} \in [1, L_i S_i]$ is the range of expected returns[7] and $L_i \geq \max_t \|a_{t,i}\|$ and $S_i \geq \|\theta^\star\|$ are the norm bounds used by the OFUL base learners. Further, $\widehat{\theta}_{t,i}$, $\Sigma_{t,i}$ and $\beta_{t,i}$ are the parameter estimate (Eq. 5.9), the covariance matrix (Eq. 5.9) and the ellipsoid radius (Eq. 5.11) of base learner i at time t, respectively. In similar spirit,

$$a_{t,i} \in \underset{a \in \mathcal{A}_t}{\operatorname{argmax}} \langle \widehat{\theta}_{t,i}, a\rangle + \beta_{t,i}\|a_{t,i}\|_{\Sigma_{t,i}^{-1}}$$

denotes the action that base learner i would take at time t. Note that we mean here the truncated actions and covariance matrix in $\mathcal{R}^{d_i}$ and $\mathcal{R}^{d_i \times d_i}$.

At each round t, Algorithm 14 first requests these bounds from each base learner to be later used in the misspecification test. The algorithm then selects one of the base learners in $\mathcal{I}$ by sampling an index $i_t \sim \text{Categorical}(p)$ from a categorical distribution with probabilities

$$p_i = \frac{1/z_i}{\sum_{j \in \mathcal{I}} 1/z_j}, \qquad \text{where } z_i = (d_i^2 + d_i S_i^2)\left(R_i^{\max} \wedge L_i^2\right) \qquad \text{for } i \in \mathcal{I}\,. \tag{5.15}$$

Since the regret of OFUL scales roughly at a rate of $\sqrt{z_i T}$, this learner selection rule approximately equalizes the regret of all learners in expectation. The algorithm proceeds by playing the action proposed by i_t, gathering the associated reward r_t, and updating i_t's internal state.[8] Finally, Algorithm 14 performs a misspecification test and terminates if this test triggers. We refer to the execution of Algorithm 14 as an epoch.

Unlike the misspecification test in Algorithm 12 which considers the hypothesis that a *specific* learner i is well specified, the misspecification test in Algorithm 14

[7] We specifically assume that $\mathbb{E}[r_t|a_t, x_t] \in [-R_\star^{\max}, +R_\star^{\max}]$ where $\star$ is the smallest base learner whose model class contains the optimal parameter $\theta_\star$.

[8] We may also pass on the observation all base learners when base learners can accept *off-policy* samples (which do not necessarily come from the proposed action), as is the case for OFUL.

tests the hypothesis that *all* active learners are well-specified. If all OFUL learners $i \in \mathcal{I}$ are well-specified, in the sense that their ellipsoid confidence sets contain $\theta_\star$ for all rounds t so far, then each $B_{t,i}$ is also a lower-bound on the optimal value in round t, since

$$B_{t,i} \le \mathbb{E}\left[r_t | a_{t,i}, x_t\right] \le \max_{a \in \mathcal{A}_t} \mathbb{E}\left[r_t | a, x_t\right] = \mu_t^\star \ .$$

Hence, the right-hand side of the misspecification test in Algorithm 14 is a lower-bound on the optimal value of all rounds to far, that is, it satisfies $\max_{j\in\mathcal{I}} \sum_{k=1}^t B_{k,j} \le \sum_{k=1}^t \mu_t^\star$. Similarly, when all learners are well-specified and satisfy their putative regret bounds, then the left-hand side of the misspecification test is an upper-bound on $\sum_{k=1}^t \mu_k^\star$. We can see this as follows. First, by basic concentration arguments, the realized rewards cannot be much smaller than their conditional expectations with high probability, that is, $\sum_{i\in\mathcal{I}} U_i(t) \ge \sum_{k=1}^t \mathbb{E}\left[r_t|a_t, x_t\right] - c\sqrt{t \ln \frac{\ln(t)}{\delta}}$. This implies that

$$\begin{aligned}
&\sum_{i\in\mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \\
&\ge \sum_{k=1}^t \mathbb{E}\left[r_t|a_t, x_t\right] + \sum_{i\in\mathcal{I}} R_i(n_i(t)) = \sum_{i\in\mathcal{I}} \left[\sum_{k\in T_i(t)} \mathbb{E}\left[r_t|a_t, x_t\right] + R_i(n_i(t)) \right] \\
&\ge \sum_{i\in\mathcal{I}} \left[\sum_{k\in T_i(t)} \mathbb{E}\left[r_t|a_t, x_t\right] + \mathsf{Reg}_i(t) \right] = \sum_{i\in\mathcal{I}} \sum_{k\in T_i(t)} \mu_k^\star = \sum_{k=1}^k \mu_k^\star,
\end{aligned}$$

where the last inequality holds because $R_i(n_i(t)) \ge \mathsf{Reg}_i(t)$ when i is well-specified. Thus, if all learners are well-specified, the misspecification test cannot trigger (with high probability). The following theorem formalizes this argument:

Theorem 5.6.1. *With probability at least $1-\delta$, Algorithm 14 does not terminate if all base learners are well-specified and their elliptical confidence sets contain $\theta^\star$ at all times.*

Therefore, if the test does trigger, at least one learner in $\mathcal{I}$ has to be misspecified, that is, either their putative regret bound R_i or a lower bound $B_{k,i}$ does not hold. However, until the test triggers, the condition in the test is sufficient to control the regret as the following theorem formalizes.

In this result, we assume that the base learner regret bounds z_i (see Eq. (5.15)) are sufficiently apart, i.e., $2z_i \le z_{i+1}$ holds for all $i \in \mathcal{I} \setminus \{M\}$. Note that this assumption can always be ensured by first filtering the base learners. This filtering can increase the regret by at most a factor of 2.

Theorem 5.6.2. *Assume that Algorithm 14 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$ (see Eq. (5.15)). All base learners use expected reward range $R_i^{\max} = 1$ and $\lambda = 1$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right)$$

Here, we highlighted the regret bound of the single best well-specified learner $\star$ in green. We here assumed that the range of expected rewards is known and 1. If this is not the case and we have to rely on the expected reward range induced by the vector norms L_i and S_i, then we have an additional lower-order term:

Theorem 5.6.3. *Assume that Algorithm 14 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i and $R_i^{\max} = L_i S_i$ with $2z_i \leq z_{i+1}$ (see Eq. (5.15)). Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{t} + \sum_{i \in \mathcal{I}} L_i S_i\right) .$$

The proofs of Theorem 5.6.3 and Theorem 5.6.2 are similar to the proof of Theorem 5.4.1 but requires a randomized version of the standard elliptical potential lemma that we prove in Lemma 5.9.4.

Main Algorithm

We now show how to obtain a robust model selection algorithm for adversarial contexts with the help of the Epoch Balancing subroutine from the previous section. Since Theorem 5.6.2 guarantees that the regret of Epoch Balancing is controlled in each epoch, all that is left it to ensure that the number of epochs is small. When Algorithm 14 terminates, we know that one of the base learners must have been misspecified but we do not know which one. We here use the hierarchy of base

Algorithm 15: Regret Bound Balancing and Elimination with Adversarial Contexts

```
for s = 1, ..., M do
    EpochBalancing ({s, s+1..., M}) in Algorithm 14
```

learners: It is safe to remove the learner $i_{\min} = \min_{i \in \mathcal{I}} d_i$ with the smallest dimension as its model class is a subset of the model classes of other base learners. Thus, if there is a model class that fails to contain $\theta^\star$, this must also be the case for $i_{\min}$. Therefore, our main algorithm shown in Algorithm 15 calls Epoch Balancing (Algorithm 14) repeatedly and removes the smallest index from the active learner set each time.

Note that once $d_i \geq d_\star$ for all $i \in \mathcal{I} = \{s, s+1, \dots, M\}$, Epoch balancing will not terminate with high probability because all remaining learners are well-specified and their bounds hold (see Theorem 5.6.1). Therefore, there can only be $i_\star \leq M$ epochs where $d_{i_\star} = d_\star$ and the total regret $\mathsf{Reg}(T)$ of Algorithm 15 is just the sum of the regret in each epoch up to the total number of T rounds. We denote by $t^{(s)}(T)$ the total number of rounds in the first s epochs after a total of T rounds. Note that $t^{(s)}(T)$ are stopping times. The regret in the s-th epoch is referred to as $\mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T))$ where $t^{(s)}(T) - t^{(s-1)}(T)$ is the number of rounds in episode s. Therefore, we can write the total regret as

$$\mathsf{Reg}(T) = \sum_{s=1}^{M} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \ . \tag{5.16}$$

The regret incurred within each epoch can be bound using Theorem 5.6.2, which yields the main result of this section:

Theorem 5.6.4 (Model Selection for Adversarial Contexts in Stochastic Linear Bandits)**.** *Assume that Algorithm 15 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$ (see Eq. (5.15)). All base learners use regularizer $\lambda = 1$. With probability at least $1 - 3(M+1)\delta$ the total regret of Algorithm 15 is bounded for all rounds $T \in \mathbb{N}$ as*

$$\mathsf{Reg}(T) = \widetilde{O}\left(\left(\sqrt{B}d_\star + \sqrt{Bd_\star}S_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{T}\right) \ ,$$

if base learners use a common expected reward range $R_i^{\max} = 1$. Here, B are the number of base learners that use a misspecified model that cannot represent $\theta_\star$, If base learners use instead $R_i^{\max} = L_i S_i$, then the regret bound is

$$\mathsf{Reg}(T) = \widetilde{O}\left(\left(\sqrt{B}d_\star L_\star + \sqrt{Bd_\star}S_\star L_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{T} + B\sum_{i \in \mathcal{I}} L_i S_i\right) \ .$$

Proof. First, we consider the event where all learners with $d_i \geq d_\star$ are well-specified in the sense that their elliptical confidence intervals contain $\theta_\star$ at all times. This happens with probability at least $1 - M\delta$ by Lemma 5.5.1. Further, only consider outcomes where Theorem 5.6.2 and Theorem 5.6.1 hold in all epochs.[9] By a union bound, all these assumptions hold with probability at least $1 - 4M$. We now consider the decomposition in Eq. (5.16) and bound

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{s=1}^{M} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \overset{(i)}{=} \sum_{s=1}^{i_\star} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \\
&\overset{(ii)}{\leq} \sum_{s=1}^{i_\star} \left[C^{(s)} \sqrt{t^{(s)}(T) - t^{(s-1)}(T)} + 8.12 \sum_{i \in \mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta} \right] \\
&\leq \max_{s \in [i_\star]} C^{(s)} \sqrt{i_\star \sum_{s=1}^{i_\star} (t^{(s)}(T) - t^{(s-1)}(T))} + 8.12 i_\star \sum_{i \in \mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta} \\
&= \max_{s \in [i_\star]} C^{(s)} \sqrt{i_\star T} + 8.12 i_\star \sum_{i \in \mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta}
\end{aligned}$$

where (i) follows from Theorem 5.6.1 and (ii) from Theorem 5.6.2 with epoch-dependent factor $C^{(s)} \leq \widetilde{O}\left(\left(d_\star + \sqrt{d_\star} S_\star + M\right)\left(d_\star + \sqrt{d_\star} S_\star\right)\right)$ or Theorem 5.6.3 with epoch-dependent factor $C^{(s)} \leq \widetilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + M\right)\left(d_\star + \sqrt{d_\star} S_\star\right)\right) L_\star$ □

5.7 Omitted Proofs of Section 5.4

Lemma 5.7.1. *There is an absolute constant c such that the event*

$$\mathcal{G} = \left\{ \forall i \in [M],\ \forall t \in \mathbb{N} \colon |n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t)| \leq c \sqrt{\ln \frac{M \ln n_i(t)}{\delta} n_i(t)} \right\} \tag{5.17}$$

has probability at least $1 - \delta$

[9] We note that both theorems hold for arbitrary sequences of contexts and therefore also when the s-th instance of Epoch Balancing is started after a random number of rounds $t^{(s-1)}(T)$.

Proof. Consider a fixed $i \in [M]$ and write the LHS in the event definition as

$$n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t) \tag{5.18}$$

$$= \sum_{k \in T_i(t)} \left(\mu^\star - r_k - \max_{\pi' \in \Pi} \mathbb{E}\left[r_k | \pi', x_k\right] + \mathbb{E}\left[r_k | \pi_k, x_k\right] \right)$$

$$= \sum_{k \in T_i(t)} \left(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}\left[r_k | \pi', x_k\right] \right) + \sum_{k \in T_i(t)} \left(\mathbb{E}\left[r_k | \pi_k, x_k\right] - r_k \right). \tag{5.19}$$

Consider the first sum and let $\mathcal{F}_t$ be the sigma-field induced by all variables up to round t, i.e., $(\mathcal{I}_k, x_k, i_k, a_k, r_k)_{k \leq t}$. Note that i_{t+1}, the learner chosen at $t+1$ is $\mathcal{F}_t$-measurable. Hence, $X_k = \mathbf{1}\{i_k = i\}(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}\left[r_k | \pi', x_k\right]) \in [-1, +1]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from [33]. Using the terminology and definition in this article, by case Hoeffding I in Table 4, the process $S_k = \sum_{j=1}^k X_k$ is sub-ψ_N with variance process $V_k = \sum_{j=1}^k \mathbf{1}\{i_j = i\}/4$. Thus by using the boundary choice in Equation (11) of [33], we get

$$\begin{aligned} S_k &\leq 1.7\sqrt{V_k \left(\ln\ln(2V_k) + 0.72\ln(5.2/\delta)\right)} \\ &= 0.85\sqrt{n_i(k)\left(\ln\ln(n_i(k)/2) + 0.72\ln(5.2/\delta)\right)} \end{aligned}$$

for all k where $V_k \geq 1$ with probability at least $1 - \delta$. Applying the same argument to $-S_k$ gives that

$$\left| \sum_{k \in T_i(t)} \left(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}\left[r_k | \pi', x_k\right] \right) \right| \leq 3 \vee 0.85\sqrt{n_i(k)\left(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta)\right)}$$

holds with probability at least $1 - \delta$ for all t.

Consider now the second term in (5.19) and let $\mathcal{F}_t$ now be the sigma-field generated by $\sigma((\mathcal{I}_k, x_k, i_k, a_k, r_k)_{k \leq t}, \mathcal{I}_{t+1}, x_{t+1}, i_{t+1}, a_{t+1})$, i.e. all variables up to the reward at round $t+1$. Then $X_k = \mathbf{1}\{i_k = i\}(\mathbb{E}\left[r_k | \pi_k, x_k\right] - r_k) \in [-1, +1]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$ and we can apply the same concentration argument as for the first term to get with probability at least $1 - \delta$ for all t

$$\left| \sum_{k \in T_i(t)} \left(\mathbb{E}\left[r_k | \pi_k, x_k\right] - r_k \right) \right| \leq 3 \vee 0.85\sqrt{n_i(k)\left(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta)\right)}\,.$$

We now take a union bound over both concentration results and $i \in [M]$ and rebind $\delta \to \delta/M$. Then picking the absolute constant c sufficiently large gives the desired statement. ∎

Lemma 5.7.2 (Sufficient Condition for Elimination)**.** *If the psuedo-regret of learner i exceeds for any $\star \in \mathcal{W}$ the following bound in round t,*

$$\mathsf{Reg}_i(t) > R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)} R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \tag{5.20}$$

then learner i fails the misspecification test of Algorithm 12 in event $\mathcal{G}$ and is eliminated.

Proof. After dividing Equation 5.20 by $n_i(t)$, this condition implies in event $\mathcal{G}_\star$

$$\frac{\mathsf{Reg}_i(t)}{n_i(t)} > \frac{R_i(n_i(t))}{n_i(t)} + \frac{\mathsf{Reg}_\star(t)}{n_\star(t)} + 2c\sqrt{\frac{\ln(M\ln t/\delta)}{n_i(t)}} + 2c\sqrt{\frac{\ln(M\ln t/\delta)}{n_\star(t)}}$$

and by $\mathcal{G}$, this implies

$$\mu_\star - \frac{U_i(t)}{n_i(t)} > \quad \frac{R_i(n_i(t))}{n_i(t)} + \mu_\star - \frac{U_\star(t)}{n_\star(t)} + c\sqrt{\frac{\ln(M\ln t/\delta)}{n_i(t)}} + c\sqrt{\frac{\ln(M\ln t/\delta)}{n_\star(t)}} \ .$$

Rearranging terms yields

$$\frac{U_i(t)}{n_i(t)} + \frac{R_i(n_i(t))}{n_i(t)} + c\sqrt{\frac{\ln(M\ln t/\delta)}{n_i(t)}} < \frac{U_\star(t)}{n_\star(t)} - c\sqrt{\frac{\ln(M\ln t/\delta)}{n_\star(t)}} \ .$$

Hence, since $t > n_i(t)$ and $t > n_\star(t)$, the misspecification test in Algorithm 12 fails. □

Special Case with T^β Candidate Regret Bounds

We here provide the proof of our gap-independent result which we restate here for convenience:

Theorem 5.4.4. *If Algorithm 12 is used with candidate regret bounds in Equation (5.7), then its total regret is bounded with probability at least $1-\delta$ for all T as*

$$\mathsf{Reg}(T) \le \left(M + 2B^{1-\beta} d_\star^{\frac{1}{\beta}-1}\right) d_\star C T^\beta + 5 d_\star^{\frac{1}{2\beta}} c\sqrt{BT\ln\frac{M\ln T}{\delta}} + 2M,$$

where $\star \in \mathcal{W}$ is any well-specified learner and $B = |\mathcal{B}|$ is the number of misspecified learners.

Proof. We start with the general regret bound from Theorem 5.4.1 given by

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M + 2c \sum_{i\in\mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}}\ , \tag{5.21}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + 2M \le M R_\star(T) + 2M \le M d_\star C T^\beta + 2M,$$

where we only used the monotonicity of regret bounds and the definition of $R_\star$. We continue with the first part of the last term which we control as follows

$$2c \sum_{i\in\mathcal{B}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \le 2c \sqrt{B \ln \frac{M \ln T}{\delta} \sum_{i\in\mathcal{B}} n_i(t_i)} \le 2c \sqrt{BT \ln \frac{M \ln T}{\delta}}$$

where we first applied Cauchy-Schwarz inequality and then used the fact that the total number of rounds played by all base learners is at most T. Similarly, we can bound the other part of the final term in (5.21) as

$$\begin{aligned} 2c \sum_{i\in\mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} &\le 2c \sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{T \ln \frac{M \ln T}{\delta}} \\ &\le 2\sqrt{2} c d_\star^{\frac{1}{2\beta}} \sqrt{BT \ln \frac{M \ln T}{\delta}}, \end{aligned}$$

where the final step follows from Lemma 5.7.3 with

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \le 2 \sum_{i\in\mathcal{B}} \left(1 \vee \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right) \le 2B d_\star^{1/\beta}\ . \tag{5.22}$$

It only remains to bound the second term (5.21). Here again we make use of the pull-ratio bound from (5.22) to bound

$$\begin{aligned} &\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) = C d_\star \sum_{i\in\mathcal{B}} \left(\frac{n_i(t_i)}{n_\star(t_i)}\right)^{1-\beta} n_i(t_i)^\beta \\ &\le C d_\star \left(\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}\right)^{1-\beta} \left(\sum_{i\in\mathcal{B}} n_i(t_i)\right)^{\beta} \\ &\le C d_\star \left(2B d_\star^{1/\beta}\right)^{1-\beta} T^\beta \le 2 C B^{1-\beta} d_\star^{1/\beta} T^\beta, \end{aligned}$$

where the first inequality follows from Hölder's inequality. Combining all bounds for the individual terms yields the desired statement. □

Below, we prove technical results for the slightly more general candidate regret bounds that can have different exponents β. Specifically, we consider candidate regret bounds of the form

$$R_i(n) = n \wedge C d_i n^{\beta_i}, \tag{5.23}$$

where $\beta_i \in (0, 1]$, $d_i \geq 1$ and C is a term that does not depend on i or n.

Lemma 5.7.3 (Play ratio bound)**.** *If Algorithm 12 is used with candidate regret bounds of the form in Equation (5.23), then*

$$\frac{n_i(t)}{n_j(t)} \leq \begin{cases} \left(2\frac{d_j}{d_i}\right)^{\frac{1}{\beta_i}} n_j(t)^{\frac{\beta_j}{\beta_i}-1} & \text{if } n_i(t) \geq (d_i C)^{\frac{1}{1-\beta}} \\ 2 & \text{if } n_i(t) \leq (d_i C)^{\frac{1}{1-\beta}} \end{cases}$$

holds for all t and active learners $i, j \in \mathcal{I}_t$ that have been played at least once.

Proof. By Lemma 5.4.2, the regret bound of i and j are balanced at t, which means that

$$R_i(n_i(t)) \leq R_j(n_j(t)) + 1 \leq 2R_j(n_j(t)) \ .$$

When $n_i(t) \leq (d_i C)^{\frac{1}{1-\beta}}$ the regret bound R_i is still in the linear regime. The balancing condition gives in this case $n_i(t) \leq 2R_j(n_j(t)) \leq 2n_j(t)$ and hence $\frac{n_i(t)}{n_j(t)} \leq 2$. Consider now the case where R_i is in the $n_i(t)^{\beta_i}$ regime. Then the balancing condition implies

$$d_i C n_i(t)^{\beta_i} \leq 2 d_j C n_j(t)^{\beta_j}.$$

Reordering terms yields

$$\left(\frac{n_i(t)}{n_j(t)}\right)^{\beta_i} \leq 2\frac{d_j}{d_i} n_j(t)^{\beta_j - \beta_i} \ .$$

□

Gap-dependent guarantee: We now provide the full proof for our main gap-dependent guarantee which we restate her for convenience:

Theorem 5.4.6. *Assume Algorithm 12 is used with candidate regret bounds in Equation (5.7) and that the pseudo-regret of all misspcified learners $j \in \mathcal{B}$ is bounded for all t from below as $\mathsf{Reg}_j(t) \geq \Delta_j n_j(t)^\alpha$, for some constants $\Delta_j > 0$ and $\alpha > \frac{1}{2} \vee \beta$. If $0 < \beta < \frac{1}{2}$ then total regret is bounded with probability at least $1-\delta$ for all T as*

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i \in \mathcal{B}} C \left((2d_\star)^{\frac{1}{\beta} + \frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}} \right) \left[\frac{20C}{\Delta_i} \ln \frac{M \ln T}{\delta} \right]^{\frac{1}{2\alpha-1}} \right) ,$$

where $\star \in \mathcal{W}$ is any well-specified learner. If instead $\beta \geq \frac{1}{2}$, then the total regret is bounded with probability at least $1-\delta$ for all T as

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i \in \mathcal{B}} C \sqrt{\ln \frac{M \ln T}{\delta}} \left(d_\star^{\frac{1}{\beta} + \frac{1}{\alpha - \beta}} + d_\star d_i^{\frac{\beta}{\alpha-\beta}} \right) \left[\frac{20C}{\Delta_i} \right]^{\frac{\beta}{\alpha-\beta}} \right).$$

Proof. Just as for the gap-independent guarantee in Theorem 5.4.4, we start with the general regret bound from Theorem 5.4.1 given by

$$\sum_{i=1}^M R_\star(n_\star(t_i)) + \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M + 2c \sum_{i \in \mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \ , \tag{5.24}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^M R_\star(n_\star(t_i)) + 2M \leq MR_\star(T) + 2M \leq Md_\star CT^\beta + 2M,$$

where we only used the monotonicity of regret bounds and the definition of $R_\star$. All remaining terms only consider misspcified learners $i \in \mathcal{B}$. In the following, we bound the contribution from each such learner individually. We have

$$\begin{aligned}
&\frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq Cd_\star \left(\frac{n_i(t_i)}{n_\star(t_i)} \right)^{1-\beta} n_i(t_i)^\beta + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq Cd_\star Z^{1-\beta} n_i(t_i)^\beta + \left(1 + \sqrt{Z} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq Cd_\star Z^{1-\beta} n_i(t_i)^\beta + 2\sqrt{Z n_i(t_i) \ln \frac{M \ln T}{\delta}},
\end{aligned} \tag{5.25}$$

where $Z = 2 \vee \left(2\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}$. Further, using the gap-assumption, Lemma 5.7.4, which is proved below, yields an upper-bound on the number of times the learner can be played

$$n_i(T) \leq \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\right]^{\frac{1}{\alpha-\beta}} \vee \left[\frac{4c}{\Delta_i}\left(1+\sqrt{Z}\right)\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1}{\alpha-1/2}}$$
$$\leq \left[\frac{5Cd_i}{\Delta_i}Z\right]^{\frac{1}{\alpha-\beta}} \vee \left[\frac{8c}{\Delta_i}\sqrt{Z}\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1}{\alpha-1/2}} .$$

We consider now two cases.

Case I: $\beta \geq 1/2$**.** Then $n_i(T) \leq \left[\frac{5Cd_i}{\Delta_i}Z\right]^{\frac{1}{\alpha-\beta}}$ and (5.25) can be bounded as

$$Cd_\star Z^{1-\beta}n_i(t_i)^\beta + 2\sqrt{Zn_i(t_i)\ln\frac{M\ln T}{\delta}} \leq 3C\sqrt{\ln\frac{M\ln T}{\delta}}d_\star Z^{1-\beta}n_i(t_i)^\beta$$
$$\leq 3C\sqrt{\ln\frac{M\ln T}{\delta}}d_\star Z^{1-\beta}\left[\frac{5Cd_i}{\Delta_i}Z\right]^{\frac{\beta}{\alpha-\beta}} .$$

When $Z = 2$, this expression is bounded from above as $6C\sqrt{\ln\frac{M\ln T}{\delta}}d_\star\left[\frac{10Cd_i}{\Delta_i}\right]^{\frac{\beta}{\alpha-\beta}}$. When $Z > 2$, then we bound this quantity instead as

$$3C\sqrt{\ln\frac{M\ln T}{\delta}}d_\star(2d_\star)^{\frac{1-\beta}{\beta}}\left[\frac{5Cd_i}{\Delta_i}\left(\frac{2d_\star}{d_i}\right)^{1/\beta}\right]^{\frac{\beta}{\alpha-\beta}} \leq 6C\sqrt{\ln\frac{M\ln T}{\delta}}d_\star^{\frac{1}{\beta}+\frac{1}{\alpha-\beta}}\left[\frac{20C}{\Delta_i}\right]^{\frac{\beta}{\alpha-\beta}} .$$

Hence, the total regret is bounded is case as

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i\in\mathcal{B}} C\sqrt{\ln\frac{M\ln T}{\delta}}\left(d_\star^{\frac{1}{\beta}+\frac{1}{\alpha-\beta}} + d_\star d_i^{\frac{\beta}{\alpha-\beta}}\right)\left[\frac{20C}{\Delta_i}\right]^{\frac{\beta}{\alpha-\beta}}\right).$$

Case II: $\beta < 1/2$**.** To simplify the final bound, we here use the somewhat crude bound on $n_i(T)$:

$$n_i(T) \leq \left[\frac{5Cd_i}{\Delta_i}Z\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1}{\alpha-1/2}}$$

This allows us to upper-bound (5.25) by

$$3Cd_\star Z^{1-\beta}\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \le 3Cd_\star Z^{1-\beta}\sqrt{\ln\frac{M\ln T}{\delta}}\left[\frac{5Cd_i}{\Delta_i}Z\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1/2}{\alpha-1/2}}.$$

When $Z = 2$, this expression is bounded from above by $6Cd_\star\left[\frac{10Cd_i}{\Delta_i}\ln\frac{M\ln T}{\delta}\right]^{\frac{1/2}{\alpha-1/2}}$. When $Z > 2$, then we bound this quantity instead as

$$\begin{aligned}
&3Cd_\star(2d_\star)^{\frac{1-\beta}{\beta}}\sqrt{\ln\frac{M\ln T}{\delta}}\left[\frac{5Cd_i}{\Delta_i}\left(\frac{2d_\star}{d_i}\right)^{1/\beta}\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1/2}{\alpha-1/2}}\\
&\le 2C(2d_\star)^{\frac{1}{\beta}}\left[\frac{5C}{\Delta_i}(2d_\star)^{1/\beta}\ln\frac{M\ln T}{\delta}\right]^{\frac{1/2}{\alpha-1/2}}
\end{aligned}$$

Hence, the total regret is bounded is case as

$$\mathsf{Reg}(T) = O\left(Md_\star CT^\beta + \sum_{i\in\mathcal{B}} C\left((2d_\star)^{\frac{1}{\beta}+\frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}}\right)\left[\frac{20C}{\Delta_i}\ln\frac{M\ln T}{\delta}\right]^{\frac{1}{2\alpha-1}}\right).$$

□

Lemma 5.7.4 (Gap-dependent elimination bound)**.** *Assume Algorithm 12 is used with candidate regret bound of the form in Equation (5.23). If the pseudo-regret of base-learner i satisfies $\mathsf{Reg}_i(t) \ge \Delta_i n_i(t)^{\alpha_i}$ for all t for a fixed $\Delta_i > 0$ and $\alpha_i > \frac{1}{2}\vee\beta_i$, then, in event $\mathcal{G}$, learner i is played at most*

$$n_i(T) \le \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\right]^{\frac{1}{\alpha_i-\beta_i}} \vee \left[\frac{4c}{\Delta_i}\left(1+\sqrt{Z}\right)\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1}{\alpha_i-1/2}},$$

times where $Z = 2\vee\left(2\frac{d_\star}{d_i}\right)^{\frac{1}{\beta_i}} n_\star(t_i)^{\frac{\beta_\star}{\beta_i}-1}$ and $\star\in\mathcal{W}$ is any well-specified learner.

Proof. Lemma 5.7.2 yields the following sufficient condition that learner i is eliminated at round t:

$$\mathsf{Reg}_i(t) > R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)}R_\star(n_\star(t)) + 2c\left(1+\sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}. \quad (5.26)$$

We now upper-bound the RHS of this sufficient condition using Lemma 5.7.3 as

$$
\begin{aligned}
R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)} R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \\
\leq R_i(n_i(t)) + 2\frac{n_i(t)}{n_\star(t)} R_i(n_i(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \\
\leq (1+2Z)\, R_i(n_i(t)) + 2c\left(1+\sqrt{Z}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \\
\leq (1+2Z)\, C d_i n_i(t)^{\beta_i} + 2c\left(1+\sqrt{Z}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}.
\end{aligned}
$$

Using this upper-bound on the RHS of (5.26) and $\Delta_i n_i(t)^{\alpha_i}$ as a lower-bound on the LHS of (5.26), we can conclude that learner i gets eliminated if the following two conditions are met:

$$
\begin{aligned}
\frac{\Delta_i}{2} n_i(t)^{\alpha_i} &> 2c\left(1+\sqrt{Z}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \\
\frac{\Delta_i}{2} n_i(t)^{\alpha_i} &> (1+2Z)\, C d_i n_i(t)^{\beta_i}
\end{aligned}
$$

Rearranging each condition yields

$$
n_i(t) > \left[\frac{4c}{\Delta_i}\left(1+\sqrt{Z}\right)\sqrt{\ln\frac{M\ln t}{\delta}}\right]^{\frac{1}{\alpha_i - 1/2}} \quad \text{and} \quad n_i(t) > \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\right]^{\frac{1}{\alpha_i-\beta_i}}.
$$

□

Special Case with $\sqrt{T\ln T}$ Candidate Regret Bounds

Consider the regret bound for all M base learners to be of the form

$$
R_i(n) = d_i C \sqrt{n \ln_+(n/\delta))} \wedge n \tag{5.27}
$$

where $\ln_+(x) = \ln(x \vee e)$ and $d_i \geq 1$ is some parameter (not necessarily an integer dimension) and $C \geq 1$ is some term that does not depend on n or i. To prepare for proving the main regret guarantee, we first show a bound on the play ratio between two active learners:

Lemma 5.7.5. *For the choice of candidate regret bounds in Equation (5.27), the following bound*

$$\frac{n_i(t)}{n_j(t)} \leq 7\left(1 \vee \frac{d_j^2}{d_i^2}\right) \ln_+\left(4e \ln \frac{t}{\delta}\right)$$

holds for all t and active learners $i, j \in \mathcal{I}_{t+1}$ that have been played at least once.

Proof. By Lemma 5.4.2, the regret bound of i and j are balanced at t, which means that

$$R_i(n_i(t)) \leq R_j(n_j(t)) + 1 \leq 2R_j(n_j(t)) \ .$$

When R_i is still in the linear regime, this implies that $n_i(t) \leq R_j(n_j(t)) + 1 \leq n_j(t_i) + 1$ and hence $\frac{n_i(t)}{n_j(t)} \leq 2$. Consider now the case where R_i is in the $\sqrt{\cdot}$ -regime. Then the balancing condition implies

$$d_i C \sqrt{n_i(t) \ln_+ \frac{n_i(t)}{\delta})} \leq 2 d_j C \sqrt{n_j(t) \ln_+ \frac{n_j(t)}{\delta}}$$

and thus

$$\sqrt{\frac{n_i(t) \ln_+(n_i(t)/\delta)}{n_j(t) \ln_+(n_j(t)/\delta)}} \leq 2\frac{d_j}{d_i}.$$

Reordering this inequality gives:

$$\frac{n_i(t)}{n_j(t)} \leq 4\frac{d_j^2}{d_i^2}\frac{\ln_+(n_j(t)/\delta)}{\ln_+(n_i(t)/\delta)} \leq 4\frac{d_j^2}{d_i^2}\ln_+(n_j(t)/\delta) \leq 4\frac{d_j^2}{d_i^2}\ln(t/\delta) \ . \tag{5.28}$$

We now refine this crude bound by considering two cases:

Case I: If $\sqrt{n_j(t)} \leq C d_j \sqrt{\ln_+(n_j(t)/\delta)}$, then $R_j(n_j(t)_= n_j(t)$ and the balancing condition gives $n_j(t) \leq 2n_i(t)$ Plugging this in (5.28) yields

$$\frac{n_i(t)}{n_j(t)} \leq 4\frac{d_j^2}{d_i^2}\frac{\ln_+(2n_i(t)/\delta)}{\ln_+(n_i(t)/\delta)} \leq 4\frac{d_j^2}{d_i^2}\ln(2e) \leq 7\frac{d_j^2}{d_i^2}.$$

Case II: In this case, $R_j(n_j(t)) = Cd_j\sqrt{n_j(t)}$ and we use (5.28) with reversed roles of i, j to get $n_j(t) \leq 4\frac{d_i^2}{d_j^2}\ln(t/\delta)n_i(t)$. Plugging this back into the middle term of (5.28) yields

$$\frac{n_i(t)}{n_j(t)} \leq 4\frac{d_j^2}{d_i^2}\ln_+(e4d_i^2/d_j^2\ln(t/\delta)).$$

When $d_j^2/d_i^2 \geq 1$, then $\frac{n_i(t)}{n_j(t)} \leq 4\frac{d_j^2}{d_i^2}\ln_+(e4\ln(t/\delta))$ follows immediately. Otherwise,

$$\begin{aligned}\frac{n_i(t)}{n_j(t)} &\leq 4\frac{d_j^2}{d_i^2}\ln_+(e4d_i^2/d_j^2\ln(t/\delta)) \leq 4\frac{d_j^2}{d_i^2}\ln(d_i^2/d_j^2) + 4\frac{d_j^2}{d_i^2}\ln(e4\ln(t/\delta))\\ &\leq \frac{4}{e} + 4\ln(e4\ln(t/\delta)) \leq 4\ln(4\ln(t/\delta))\end{aligned}$$

□

Theorem 5.7.6. *If Algorithm 12 is used with candidate regret bounds in Equation (5.27), then its total regret is bounded with probability at least $1-\delta$ for all T as*

$$\begin{aligned}\mathsf{Reg}(T) \leq &\left(M + d_\star\sqrt{B\ln_+\left(11\ln\frac{T}{\delta}\right)}\right)d_\star C\sqrt{T\ln_+(T/\delta)} + 2M\\ &+ 8cd_\star\ln\left(\frac{11M\ln T}{\delta}\right)\sqrt{BT}\end{aligned}$$

where $\star \in \mathcal{W}$ is any well-specified learner and $B = |\mathcal{B}|$ is the number of misspecified learners.

Proof. We start with the general regret bound from Theorem 5.4.1 given by

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}}\frac{n_i(t_i)}{n_\star(t_i)}R_\star(n_\star(t_i)) + 2M + 2c\sum_{i\in\mathcal{B}}\left(1+\sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}}\ , \tag{5.29}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + 2M \leq MR_\star(T) + 2M \leq Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M,$$

where we only used the monotonicity of regret bounds and the definition of $R_\star$. We continue with the first part of the last term which we control as follows

$$2c \sum_{i\in\mathcal{B}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \leq 2c\sqrt{B \ln \frac{M \ln T}{\delta} \sum_{i\in\mathcal{B}} n_i(t_i)} \leq 2c\sqrt{BT \ln \frac{M \ln T}{\delta}}$$

where we first applied Cauchy-Schwarz inequality and then used the fact that the total number of rounds played by all base learners is at most T. Similarly, we can bound the other part of the final term in (5.29) as

$$\begin{aligned} 2c \sum_{i\in\mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} &\leq 2c\sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{T \ln \frac{M \ln T}{\delta}} \\ &\leq 6c\sqrt{B \ln_+ \left(4e \ln \frac{T}{\delta}\right)} d_\star \sqrt{T \ln \frac{M \ln T}{\delta}}, \end{aligned}$$

where the final step follows from Lemma 5.7.5 with

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \leq 7 \sum_{i\in\mathcal{B}} \left(1 \vee \frac{d_\star^2}{d_i^2}\right) \ln_+ \left(4e \ln \frac{t_i}{\delta}\right) \leq 7 d_\star^2 B \ln_+ \left(4e \ln \frac{T}{\delta}\right) \tag{5.30}$$

It only remains to bound the second term (5.29). Here again we make use of the pull-ratio bound from (5.30) to bound

$$\begin{aligned} &\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) = C d_\star \sum_{i\in\mathcal{B}} \left(\frac{n_i(t_i)}{n_\star(t_i)}\right)^{1/2} n_i(t_i)^{1/2} \sqrt{\ln_+(n_\star(t_i)/\delta)} \\ &\leq C d_\star \sqrt{\ln_+(T/\delta)} \sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{\sum_{i\in\mathcal{B}} n_i(t_i)} \leq 3C d_\star^2 \sqrt{BT \ln_+(T/\delta) \ln_+ \left(4e \ln \frac{T}{\delta}\right)}, \end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality. Combining all bounds for the individual terms yields the desired statement. □

Gap-dependent Regret Guarantee: We now prove a gap-dependent regret bound for Algorithm 12 when used with candidate regret bounds in Equation (5.27).

Lemma 5.7.7 (Gap-dependent elimination bound)**.** *Assume Algorithm 12 is used with candidate regret bound of the form in Equation (5.27). If the pseudo-regret of* *base-learner i satisfies $\mathsf{Reg}_i(t) \geq \Delta_i n_i(t)^{\alpha_i}$ for all t for a fixed $\Delta_i > 0$ and $\alpha_i > \frac{1}{2}$, then, in event $\mathcal{G}$, learner i is played at most*

$$n_i(T) \leq \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\sqrt{\ln_+(MT/\delta)}\right]^{\frac{1}{\alpha_i - 1/2}},$$

times where $Z = 7\left(1 \vee \frac{d_j^2}{d_i^2}\right)\ln_+\left(4e\ln\frac{t}{\delta}\right)$ and $\star \in \mathcal{W}$ is any well-specified learner.

Proof. This statement can be proved in full analogy to Lemma 5.7.4. □

Theorem 5.7.8. *Assume Algorithm 12 is used with candidate regret bounds in Equation (5.27) and that the pseudo-regret of all misspcified learners $j \in \mathcal{B}$ is bounded for all t from below as $\mathsf{Reg}_j(t) \geq \Delta_j n_j(t)^\alpha$ for some $\alpha > \frac{1}{2} \vee \beta$ and $\Delta_j > 0$. Then total regret is bounded with probability at least $1-\delta$ for all T as*

$$\begin{aligned}\mathsf{Reg}(T) &\leq Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M \\ &\quad + 9Cd_\star \sum_{i\in\mathcal{B}} \ln_+\left(4e\ln\frac{T}{\delta}\right)^\rho\left(\ln_+\frac{MT}{\delta}\right)^\rho\left[\frac{42d_iC}{\Delta_i}\right]^{\frac{1}{2\alpha-1}}\left(1\vee\frac{d_\star}{d_i}\right)^{2\rho}.\end{aligned} \tag{5.31}$$

for $\star \in \mathcal{W}$ is any well-specified learner and $\rho = \frac{1}{2} + \frac{1}{2\alpha - 1}$.

Proof. Just as for the gap-independent guarantee in Theorem 5.7.6, we start with the general regret bound from Theorem 5.4.1 given by

$$\sum_{i=1}^M R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}}\frac{n_i(t_i)}{n_\star(t_i)}R_\star(n_\star(t_i)) + 2M + 2c\sum_{i\in\mathcal{B}}\left(1+\sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}},$$

and bound the terms individually. We begin with

$$\sum_{i=1}^M R_\star(n_\star(t_i)) + 2M \leq MR_\star(T) + 2M \leq Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M,$$

where we only used the monotonicity of regret bounds and the definition of $R_\star$. All remaining terms only consider misspcified learners $i \in \mathcal{B}$. In the following, we bound the contribution from each such learner individually. We have

$$
\begin{aligned}
&\frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \\
&\leq Cd_\star\sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\sqrt{n_i(t_i)\ln_+(n_\star(t_i)/\delta)} + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \\
&\leq Cd_\star\sqrt{Zn_i(t_i)\ln_+(T/\delta)} + \left(1+\sqrt{Z}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \\
&\leq Cd_\star\sqrt{Zn_i(t_i)\ln_+\frac{T}{\delta}} + 2\sqrt{Z}\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \\
&\leq 3Cd_\star\sqrt{Zn_i(t_i)\ln_+\frac{MT}{\delta}}
\end{aligned}
$$

where $Z = 7\left(1 \vee \frac{d_\star^2}{d_i^2}\right)\ln_+\left(4e\ln\frac{T}{\delta}\right)$. Further, using the gap-assumption, Lemma 5.7.7 yields an upper-bound on the number of times the learner can be played

$$
\begin{aligned}
n_i(T) &\leq \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\sqrt{\ln_+(MT/\delta)}\right]^{\frac{1}{\alpha-1/2}} \leq \left[\frac{6ZCd_i}{\Delta_i}\sqrt{\ln_+(MT/\delta)}\right]^{\frac{1}{\alpha-1/2}} \\
&\leq \left[\frac{42C}{\Delta_i}\left(d_i \vee \frac{d_\star^2}{d_i}\right)\ln_+\left(4e\ln\frac{T}{\delta}\right)\sqrt{\ln_+(MT/\delta)}\right]^{\frac{1}{\alpha-1/2}}
\end{aligned}
$$

We use this upper-bound to control the term

$$
\begin{aligned}
&3Cd_\star\sqrt{Zn_i(t_i)\ln_+\frac{MT}{\delta}} \\
&\leq 9Cd_\star\left(1 \vee \frac{d_\star}{d_i}\right)\ln_+\left(4e\ln\frac{T}{\delta}\right)^{\frac{1}{2}+\frac{1}{2\alpha-1}}\left(\ln_+\frac{MT}{\delta}\right)^{\frac{1}{2}+\frac{1/2}{2\alpha-1}}\left[\frac{42C}{\Delta_i}\left(d_i \vee \frac{d_\star^2}{d_i}\right)\right]^{\frac{1}{2\alpha-1}}.
\end{aligned}
$$

Combining all bounds of individual terms yields the desired bound

$$
\begin{aligned}
\mathsf{Reg}(T) &\leq Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M + \\
&\quad 9Cd_\star\sum_{i\in\mathcal{B}}\ln_+\left(4e\ln\frac{T}{\delta}\right)^{\frac{1}{2}+\frac{1}{2\alpha-1}}\left(\ln_+\frac{MT}{\delta}\right)^{\frac{1}{2}+\frac{1/2}{2\alpha-1}}\left[\frac{42d_iC}{\Delta_i}\right]^{\frac{1}{2\alpha-1}}\left(1 \vee \frac{d_\star}{d_i}\right)^{1+\frac{2}{2\alpha-1}}
\end{aligned}
$$

□

Special Case with $\epsilon_i C_2 T + C_1\sqrt{T}$ Candidate Regret Bounds

Lemma 5.7.9. *Assume all base algorithms use regret bounds of the form (5.8) in Theorem 5.4.5. Let $i \in \mathcal{I}_{t+1}$ be an active learner and $* \in \mathcal{W}$ be a well-specified learner with $\epsilon_* \geq \epsilon_i$. Then in event $\mathcal{G}$*

$$\mathsf{Reg}_i(t) \leq 1 + 10R_*(n_*(t)) + 2\epsilon_* C_2 \left(1 + \frac{c}{C_1}\sqrt{\ln \frac{M \ln t}{\delta}}\right) n_i(t) + 8c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} + 8C_1^2 + 2C_1\sqrt{n_i(t)} + 8cC_1\sqrt{\ln \frac{M \ln t}{\delta}} .$$

Proof. First, we can assume without loss of generality that $C_2\epsilon_* \leq 1$ because the regret bound is vacuous otherwise. Since i is in the active set and $*$ is well-specified, we can apply Lemma 5.4.3 which gives

$$\mathsf{Reg}_i(t) \leq 1 + R_*(n_*(t)) + 2c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} + \frac{n_i(t)}{n_*(t)} R_*(n_*(t)) + 2c\sqrt{\frac{n_i(t)^2}{n_*(t)} \ln \frac{M \ln t}{\delta}} . \tag{5.32}$$

We now simplify the expression on the right hand side using the specific form of the regret bounds R_j. This form can be split into three phases:

$R_j(n) = n$	for $\sqrt{n} \leq \frac{C_1}{1 - C_2\epsilon_j}$	Phase I
$R_j(n) \in [C_1\sqrt{n}, 2C_1\sqrt{n}]$	for $\frac{C_1}{1 - C_2\epsilon_j} < \sqrt{n} \leq \frac{C_1}{C_2\epsilon_j}$	Phase II
$R_j(n) \in [C_2\epsilon_j n, 2C_2\epsilon_j n]$	for $\frac{C_1}{C_2\epsilon_j} < \sqrt{n}$	Phase III

We now give a regret bound for learner i based on which phase its regret bound is in.

Regret bound of i in Phase I: We first consider the case where $*$ is in Phase I. Then the balancing condition from Lemma 5.4.2 $R_i(n_i(t)) \leq 2R_*(n_*(t))$ implies that $n_i(t)/n_*(t) \leq 2$ and thus

$$\mathsf{Reg}_i(t) \leq 1 + 3R_*(n_*(t)) + 2(1 + \sqrt{2})c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}}.$$

If $*$ is in Phase II, then by the balancing condition $n_i(t) \leq 4C_1\sqrt{n_*(t)}$ which implies that $\frac{n_i(t)}{\sqrt{n_*(t)}} \leq 4C_1$. Plugging this into (5.32) yields

$$\begin{aligned}\mathsf{Reg}_i(t) &\leq 1 + R_*(n_*(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + \frac{n_i(t)}{\sqrt{n_*(t)}}2C_1 + 8cC_1\sqrt{\ln\frac{\ln t}{\delta}} \\ &\leq 1 + R_*(n_*(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + 8C_1^2 + 8cC_1\sqrt{\ln\frac{M\ln t}{\delta}}.\end{aligned}$$

If $*$ is in Phase III, then by the balancing condition $n_i(t) \leq 4C_2\epsilon_* n_*(t)$ and, hence, $\frac{n_i(t)}{n_*(t)} \leq 4C_2\epsilon_* \leq 4$. Here, we have used that $C_2\epsilon_* \leq 1$ as otherwise the regret bounds hold trivially. Plugging this into (5.32) yields

$$\mathsf{Reg}_i(t) \leq 1 + 5R_*(n_*(t)) + 6c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}.$$

Regret bound of i in Phase II: We here distinguish between two cases. If $\sqrt{n_*(t)} \leq \frac{C_1}{C_2\epsilon_*}$, then $R_*(n_*(t)) \leq 2C_1\sqrt{n_*(t)}$. Then by the balancing condition $\frac{n_i(t)}{n_*(t)} \leq 9$. Plugging this into (5.32) yields

$$\mathsf{Reg}_i(t) \leq 1 + 10R_*(n_*(t)) + 8c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}.$$

Consider now the case where $\sqrt{n_*(t)} > \frac{C_1}{C_2\epsilon_*}$ and $R_*(n_*(t)) \leq 2\epsilon_* C_2 n_*(t)$. Here, we bound (5.32) directly as

$$\begin{aligned}\mathsf{Reg}_i(t) &\leq 1 + 2\epsilon_* C_2(n_*(t) + n_i(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + 2c\frac{C_2\epsilon_*}{C_1}n_i(t)\sqrt{\ln\frac{\ln t}{\delta}} \\ &\leq 1 + 2\epsilon_* C_2\left(n_*(t) + n_i(t) + \frac{c\sqrt{\ln\frac{M\ln t}{\delta}}}{C_1}n_i(t)\right) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}.\end{aligned}$$

Regret bound of i in Phase III: First, consider the case where $\sqrt{n_*(t)} > \frac{C_1}{C_2\epsilon_*}$. Then we can directly write $\frac{n_i(t)}{n_*(t)}R_*(n_*(t)) = \epsilon_* C_2 n_i(t)$ and bound $1/\sqrt{n_*(t)} \leq \frac{C_2\epsilon_*}{C_1}$. Plugging this into (5.32) yields

$$\mathsf{Reg}_i(t) \leq 1 + R_*(n_*(t)) + \epsilon_* C_2 n_i(t) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + \frac{C_2\epsilon_*}{C_1}2c\sqrt{\ln\frac{M\ln t}{\delta}}n_i(t).$$

It remains to bound the regret when $\sqrt{n_*(t)} \leq \frac{C_1}{C_2\epsilon_*}$. Since i is in Phase III, we also have $\sqrt{n_i(t)} > \frac{C_1}{C_2\epsilon_i} \geq \frac{C_1}{C_2\epsilon_*}$. The balancing condition yields $\epsilon_i C_2 n_i(t) \leq 4C_1\sqrt{n_*(t)}$ and thus

$$\frac{n_i(t)}{\sqrt{n_*(t)}} \leq \frac{4C_1}{C_2\epsilon_i} \leq \sqrt{n_i(t)}.$$

Plugging this into (5.32) yields

$$\begin{aligned}\mathsf{Reg}_i(t) &\leq 1 + R_*(n_*(t)) + 4c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + \frac{n_i(t)}{n_*(t)}2C_1\sqrt{n_*(t)}\\ &\leq 1 + R_*(n_*(t)) + 4c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + 2C_1\sqrt{n_i(t)}.\end{aligned}$$

□

5.8 Omitted Proofs of Section 5.6

Epoch Balancing Termination (Proof of Theorem 5.6.1)

Theorem 5.6.1. *With probability at least $1-\delta$, Algorithm 14 does not terminate if all base learners are well-specified and their elliptical confidence sets contain $\theta^\star$ at all times.*

Proof. Since all learners are well-specified and their lower-confidence bounds $L_{t,i}$ satisfy $L_{t,i} \leq \mathbb{E}[r_t|a_{t,i}, x_t] \leq \mu_k^\star$, the right-hand side of the misspecification test satisfies

$$\max_{j\in\mathcal{I}}\sum_{k=1}^{t} B_{k,j} \leq \sum_{k=1}^{t}\mu_k^\star.$$

for all $t \in \mathbb{N}$ Further, with probability at least $1-\delta$, by Lemma 5.8.2, the left-hand side of the misspecification test satisfies for all $t \in \mathbb{N}$

$$\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t\ln\frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t}\mu_k^\star.$$

Thus, the misspecification test never triggers and Algorithm 14 does not terminate. □

Lemma 5.8.1. *Let $\delta \in (0,1)$ and consider the event*

$$\mathcal{G} = \left\{ \forall t \in \mathbb{N} \colon \left| \sum_{i \in \mathcal{I}} U_i(t) - \sum_{k=1}^{t} \mathbb{E}\left[r_k | a_k, x_k\right] \right| \leq c \sqrt{t \ln \frac{\ln(t)}{\delta}} \right\}.$$

where $c > 0$ is an absolute constant. Then $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$.

Proof. Let $\mathcal{F}_t = \sigma(x_1, i_1, a_1, r_1, \dots, x_{t-1}, i_{t-1}, a_{t-1}, r_{t-1}, x_{t-1}, i_{t-1}, a_{t-1})$ be the sigma-field induced by all variables up to the reward at round t. Hence, $X_k = r_k - \mathbb{E}\left[r_k | a_k, x_k\right]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from [33]. Using the terminology and definition in this article, by case Hoeffding I in Table 4, the process $S_k = \sum_{j=1}^{k} X_k$ is sub-ψ_N with variance process $V_k = k/4$. Thus by using the boundary choice in Equation (11) of [33], we get

$$\begin{aligned} S_k &\leq 1.7 \sqrt{V_k \left(\ln\ln(8V_k) + 0.72 \ln(5.2/\delta)\right)} \\ &= 0.85 \sqrt{k \left(\ln\ln(4k) + 0.72 \ln(5.2/\delta)\right)} \end{aligned}$$

for all k with probability at least $1 - \delta$. Applying the same argument to $-S_k$ gives that

$$\left| \sum_{k=1}^{t} \left(r_k - \mathbb{E}\left[r_k | a_k, x_k\right]\right) \right| \leq 0.85 \sqrt{t \left(\ln\ln(4t) + 0.72 \ln(10.4/\delta)\right)}$$

holds with probability at least $1 - \delta$ for all t. Since $\sum_{i \in \mathcal{I}} U_i(t) = \sum_{k=1}^{t} r_k$, the statement follows. Note that this concentration argument holds for all t uniformly and therefore also when t is random. □

Lemma 5.8.2 (Upper-confidence bound on optimal reward)**.** *In event $\mathcal{G}$ from Lemma 5.8.1, the following holds. If at time t all learners $i \in \mathcal{I}$ are well-specified, then the left-hand side in the misspecification test of Algorithm 14 is a lower-bound on the optimal rewards, i.e.,*

$$\sum_{i \in \mathcal{I}} [U_i(t) + R_i(n_i(t))] + c \sqrt{t \ln \frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t} \mu_k^\star.$$

Proof. By Lemma 5.8.1, in the considered event, we have

$$
\begin{aligned}
&\sum_{i\in\mathcal{I}}[U_i(t)+R_i(n_i(t))]+c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\
&\geq \sum_{i\in\mathcal{I}} R_i(n_i(t)) + \sum_{k=1}^{t}\mathbb{E}\left[r_k|a_k,x_k\right] && \text{(by Lemma 5.8.1)} \\
&\geq \sum_{i\in\mathcal{I}} \mathsf{Reg}_i(t) + \sum_{k=1}^{t}\mathbb{E}\left[r_k|a_k,x_k\right] && \text{(each learner is well-specified)} \\
&= \sum_{i\in\mathcal{I}} \left[\mathsf{Reg}_i(t) + \sum_{k\in T_i(t)}\mathbb{E}\left[r_k|a_k,x_k\right]\right] \\
&= \sum_{i\in\mathcal{I}}\sum_{k\in T_i(t)} \mu_k^\star = \sum_{k=1}^{t}\mu_k^\star. && \text{(by definition of regret)}
\end{aligned}
$$

□

Regret Bound for Epoch Balancing (Proof of Theorem 5.6.2)

Theorem 5.6.2. *Assume that Algorithm 14 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$ (see Eq. (5.15)). All base learners use expected reward range $R_i^{\max} = 1$ and $\lambda = 1$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right)$$

Proof. We apply Theorem 5.8.3 which immediately yields the desired bound

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right) \; .$$

□

Theorem 5.6.3. *Assume that Algorithm 14 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i and $R_i^{\max} = L_iS_i$ with $2z_i \leq z_{i+1}$ (see Eq. (5.15)). Denote by $\star$ the smallest index of the base learner* *so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1-2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star} S_\star) L_\star \sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i\right) .$$

Proof. We apply Theorem 5.8.3 which yields

$$\mathsf{Reg}(t) \leq \widetilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star} S_\star) L_\star \sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i \ln\ln(t)\right) .$$

□

Theorem 5.8.3 (General Regret Bound of Epoch Balancing)**.** *Assume that Algorithm 14 is run with instances of OFUL as base learners which use different dimensions $d_i, S_i, L_i, R_i^{\max}$ and regularization parameter $\lambda = 1$. Denote by $\star$ the index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1-2\delta$, the regret is bounded for all rounds t as*

$$\begin{aligned}\mathsf{Reg}(t) &\leq (|\mathcal{I}|\sqrt{z_\star} + z_\star\sqrt{\bar{M}})x(t)\sqrt{t} + 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\ &\leq \sqrt{(d_\star^2 + d_\star S_\star^2)}|\mathcal{I}|\left(R_\star^{\max}\wedge L_\star\right)\sqrt{t}(2+2c)x(t) \\ &\quad + (d_\star^2 + d_\star S_\star^2)\left(R_\star^{\max}\wedge L_\star\right)^2\sqrt{\bar{M}t}(2+2c)x(t) \\ &\quad + 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta},\end{aligned}$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing (i.e., $2z_i \leq z_{i+1}$ for all $i \in \mathcal{I}$). Here $x(t) = O(\ln\frac{tL_{\max}}{\delta} + \ln\ln(R_{\max}^{\max}t \wedge L_{\max}t)$

Proof. Since learner $i_\star$ is well-specified and its elliptical confidence set contains $\theta^\star$, it holds that

$$\sum_{k=1}^t \mu_k^\star \leq \sum_{k=1}^t \max_{a\in\mathcal{A}_k}\left[\langle\widehat{\theta}_{k,\star}, a\rangle + \beta_{k,\star}\|a\|_{\Sigma_{k,\star}^{-1}}\right] = \sum_{k=1}^t \langle\widehat{\theta}_{k,\star}, a_{k,\star}\rangle + \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}.$$

Thus, we can write the total regret up to round t as

$$\begin{aligned}\mathsf{Reg}(t) &= \sum_{k=1}^{t}\left[\mu_k^\star - \mathbb{E}\left[r_k|a_k, x_k\right]\right] = \sum_{k=1}^{t}\mu_k^\star - \sum_{k=1}^{t}\mathbb{E}\left[r_k|a_k, x_k\right]\\ &\leq \sum_{k=1}^{t}\mu_k^\star - \sum_{i\in\mathcal{I}} U_i(n_i(t)) + c\sqrt{t\ln\frac{\ln(t)}{\delta}},\end{aligned}$$

where the inequality holds in event $\mathcal{G}$ of Lemma 5.8.1. If Algorithm 14 does not stop in iteration t, then the misspecification test does not trigger for any learner, and in particular for learner $i_\star$. This implies that

$$\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t\ln\frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t} B_{k,\star}$$

Rearranging terms and plugging this inequality back into the regret bound from above yields

$$\mathsf{Reg}(t) \leq \sum_{k=1}^{t}\left[\mu_k^\star - B_{k,\star}\right] + \sum_{i\in\mathcal{I}} R_i(n_i(t)) + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \tag{5.33}$$

We bound the first term in Equation 5.33 as

$$\begin{aligned}&\sum_{k=1}^{t}\left[\mu_k^\star - B_{k,\star}\right]\\ &\overset{(i)}{\leq} \sum_{k=1}^{t}\left[R_\star^{\max} \wedge (\langle\widehat{\theta}_{k,\star}, a_{k,\star}\rangle + \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}) - (-R_\star^{\max} \vee (\langle\widehat{\theta}_{k,\star}, a_{k,\star}\rangle - \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}))\right]\\ &\leq \sum_{k=1}^{t}\left[2R_\star^{\max} \wedge 2\beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}\right] \leq 2\beta_{t,\star}\sum_{k=1}^{t}\left[\frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}\right]\\ &\overset{(ii)}{\leq} 2\beta_{t,\star}\sqrt{t\sum_{k=1}^{t}\left[\left(\frac{R_\star^{\max}}{\beta_{t,\star}}\right)^2 \wedge \frac{L^2}{\lambda_i} \wedge \|a_{k,\star}\|^2_{\Sigma_{k,\star}^{-1}}\right]}\end{aligned}$$

where (i) follows from the definition of $B_{k,i}$ and the fact that the ellipsoid confidence set of $\star$ contain the true parameter and (ii) applies the Cauchy-Schwarz inequality. We now apply a randomized version of the elliptical potential lemma which we prove in Lemma 5.9.4. This yields

$$\sum_{k=1}^{t}[\mu_k^\star - B_{\star,k}] \leq 4\beta_{t,\star}\sqrt{\frac{t}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t \vee 2)\det\Sigma_{t,\star}}{\delta\det\Sigma_{0,\star}}}$$
$$\leq 4\beta_{t,\star}\sqrt{\frac{td_\star}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t\vee 2)(d_\star\lambda_\star + tL_\star^2)}{\delta d_\star\lambda_\star}}$$

where $b_\star = \frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \frac{L_\star}{\sqrt{\lambda_\star}}$. For the second term in Equation 5.33, we apply Lemma 5.8.4 with $\alpha = \delta$ as

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \leq 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{d_i\lambda_i + tp_iL_i^2}{d_i\lambda_i}}.$$

Combining the terms for both bounds, we arrive at the regret bound

$$\begin{aligned}
\mathsf{Reg}(t) \leq\ & 4\beta_{t,\star}\sqrt{\frac{td_\star}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t\vee 2)(d_\star\lambda_\star + tL_\star^2)}{\delta d_\star\lambda_\star}} \\
&+ 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{d_i\lambda_i + tp_iL_i^2}{d_i\lambda_i}} \\
&+ 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\
\leq\ & x(t)\sqrt{\frac{z_\star t}{p_\star}} + x\sum_{i\in\mathcal{I}}\sqrt{z_ip_it} + 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}}
\end{aligned}$$

where

$$z_i = (\sigma^2 d_i + \lambda_i S_i^2)d_i(1+b_i^2) \leq 2(d_i^2 + d_iS_i^2)\left(R_i^{\max}\wedge L_i\right)^2 \qquad \text{and}$$
$$\begin{aligned}
x(t) &= 12\max_{i\in\mathcal{I}}\sqrt{\ln\left(\frac{1+tL_i^2/\lambda_i}{\delta}\right)\ln\frac{5.2\ln(2b_i^2t\vee 2)(d_i\lambda_i + tL_i^2)}{\delta d_i\lambda_i}} \\
&\leq 12\max_{i\in\mathcal{I}}\sqrt{\ln\left(\frac{1+tL_i^2}{\delta}\right)\ln\frac{10.4\ln(2\left(R_i^{\max}\wedge L_i\right)t)(1+tL_i^2)}{\delta}} \\
&\leq 12\ln\frac{10.4(1+tL_{\max}^2)\ln(2\left(R_{\max}^{\max}\wedge L_{\max}\right)t)}{\delta}.
\end{aligned}$$

We now use the definition of $p_i \propto \frac{1}{z_i}$ and bound

$$\sum_{i \in \mathcal{I}} \sqrt{z_i p_i} = \sum_{i \in \mathcal{I}} \sqrt{\frac{1}{\sum_{i \in \mathcal{I}} z_i^{-1}}} = \frac{|\mathcal{I}|}{\sqrt{\sum_{i \in \mathcal{I}} z_i^{-1}}} \leq \frac{|\mathcal{I}|}{\sqrt{z_\star^{-1}}} = |\mathcal{I}| \sqrt{z_\star}$$

where the inequality uses the fact that $\star \in \mathcal{I}$. Further

$$\sqrt{\frac{z_\star}{p_\star}} = z_\star \sqrt{\sum_{i \in \mathcal{I}} \frac{1}{z_i}} \leq z_\star \sqrt{|\mathcal{I}|}$$

holds for any z_i but if we know that $z_1 \leq 2z_2 \leq 4z_4 \dots M z_M$, then

$$\sqrt{\frac{z_\star}{p_\star}} = z_\star \sqrt{\sum_{i \in \mathcal{I}} \frac{1}{z_i}} \leq 2z_\star.$$

Thus, we can bound the total regret as

$$\begin{aligned}
\mathsf{Reg}(t) &\leq (|\mathcal{I}|\sqrt{z_\star} + z_\star \sqrt{\bar{M}}) x(t) \sqrt{t} + 8.12 \sum_{i \in \mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}| \ln(2t)}{\delta} + 2c\sqrt{t \ln \frac{\ln(t)}{\delta}} \\
&\leq \sqrt{(d_\star^2 + d_\star S_\star^2)} |\mathcal{I}| \left(R_\star^{\max} \wedge L_\star\right) \sqrt{t} (2 + 2c) x(t) \\
&\quad + \left(d_\star^2 + d_\star S_\star^2\right) \left(R_\star^{\max} \wedge L_\star\right)^2 \sqrt{\bar{M}t} (2 + 2c) x(t) \\
&\quad + 8.12 \sum_{i \in \mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}| \ln(2t)}{\delta},
\end{aligned}$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing. Note that since this bound holds in the penultimate round of Algorithm 14 and the regret in the final round can be at most 1, this bound holds for all rounds t played by Algorithm 14, including the last. □

Lemma 5.8.4 (Regret bounds are balanced)**.** *Let $\alpha \in (0, 1)$ be arbitrary but fixed. With probability at least $1 - \alpha$, the sum of regret bounds satisfy in all iterations t of Algorithm 14 the following upper-bound*

$$\sum_{i \in \mathcal{I}} R_i(n_i(t)) \leq 8.12 \sum_{i \in \mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}| \ln(2t)}{\alpha} + 2 \sum_{i \in \mathcal{I}} \beta_{t,i} \sqrt{3 d_i p_i t \left(1 + b_i^2\right) \ln \frac{\lambda_i d_i + 3t p_i L_i^2}{\lambda_i d_i}}$$

where $b_i = \frac{R_i^{\max}}{2\beta_{t,i}} \wedge \frac{L_i}{\sqrt{\lambda_i}}$.

Proof. By the choice of regret bounds we have

$$\begin{aligned}
R_i(n_i(t)) &= \sum_{k\in T_i(t)} \left[2\beta_{k,i}\|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge R_i^{\max}\right] \\
&\le R_i^{\max} n_i(t) \wedge 2\beta_{t,i} \sum_{k\in T_i(t)} \left(\|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge \frac{R_i^{\max}}{2\beta_{t,i}}\right) \\
&\le R_i^{\max} n_i(t) \wedge 2\beta_{t,i} \sqrt{n_i(t) \sum_{k\in T_i(t)} \left(\|a_{k,i}\|^2_{\Sigma_{k,i}^{-1}} \wedge \left(\frac{R_i^{\max}}{2\beta_{t,i}}\right)^2 \wedge \frac{L_i^2}{\lambda_i}\right)} \\
&\le R_i^{\max} n_i(t) \vee 2\beta_{t,i} \sqrt{d_i n_i(t)\left(1+b_i^2\right) \ln \frac{\lambda_i + n_i(t) L_i^2/d_i}{\lambda_i}}
\end{aligned}$$

where $b_i = \frac{R_i^{\max}}{2\beta_{t,i}} \wedge \frac{L_i}{\sqrt{\lambda_i}}$ and the last inequality follows from of Lemma 5.9.3. To control the the number of times each learner was chosen, we use Lemma 5.8.5. This gives with probability at least $1-\alpha$ for all iterations t simultaneously $n_i(t) \le 3tp_i \vee 8.12 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\alpha}$. This yields a regret bound of

$$R_i(n_i(t)) \le 8.12 R_i^{\max} \ln \frac{5.2|\mathcal{I}| \ln(2t)}{\alpha} \quad \vee \quad 2\beta_{t,i}\sqrt{3d_i p_i t\left(1+b_i^2\right) \ln \frac{\lambda_i + 3tp_i L_i^2/d_i}{\lambda_i}}.$$

Summing over R_i and plugging in $\beta_{t,i}$ yields

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \le 8.12 \sum_{i\in\mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\alpha} + 2\sum_{i\in\mathcal{I}} \beta_{t,i} \sqrt{3d_i p_i t\left(1+b_i^2\right) \ln \frac{\lambda_i + 3tp_i L_i^2/d_i}{\lambda_i}}$$

□

Lemma 5.8.5. *The number of times each a learner $i \in \mathcal{I}$ has been played in Algorithm 14 after t iterations is bounded with probability at least $1-\delta$ for all $t \in \mathbb{N}$ and $i \in \mathcal{I}$ as*

$$n_i(t) \le \frac{3}{2} tp_i + 4.06 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\delta} \le 3tp_i \vee 8.12 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\delta}$$

Proof. Fix an $i \in \mathcal{I}$ and consider the martingale difference sequence $X_t = \mathbf{1}\{i_t = i\} - p_i$ with variance. The process $S_t = \sum_{k=1}^t X_k$ with variance process $W_t = tp_i(1-p_i)$

satisfies the sub-ψ_P condition of [33] with constant $c = 1$ (see Bennett case in Table 3 of [33]). By Lemma 5.9.5, the bound

$$S_t \leq 1.44\sqrt{(W_t \vee m)\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)} + 0.41\frac{L^2}{\lambda}\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)$$

holds for all $t \in \mathbb{N}$ with probability at least $1 - \delta$. We set $m = tp_i$ and upper-bound the RHS further as

$$\begin{aligned} S_t &\leq 1.44\sqrt{tp_i\left(1.4 \ln\ln(2t) + \ln\frac{5.2}{\delta}\right)} + 0.41\left(1.4 \ln\ln(2t) + \ln\frac{5.2}{\delta}\right) \\ &\leq \frac{tp_i}{2} + 1.45\left(1.4 \ln\ln(2t) + \ln\frac{5.2}{\delta}\right), \end{aligned}$$

where used the AM-GM inequality in the final step. We therefore get that with probability at least $1 - \delta$, the following upper-bound in the number of times learner i was selected by time t holds for all $i \in \mathcal{I}$ and $t \in \mathbb{N}$:

$$n_i(t) \leq \frac{3}{2}tp_i + 2.9\left(1.4 \ln\ln(2t) + \ln\frac{5.2|\mathcal{I}|}{\delta}\right) \leq \frac{3}{2}tp_i + 4.06 \ln\frac{5.2|\mathcal{I}| \ln(2t)}{\delta}.$$

We can now distinguish between two cases: When $\frac{3}{2}tp_i \leq 4.06 \ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta}$, then

$$n_i(t) \leq 8.12 \ln\frac{5.2|\mathcal{I}| \ln(2t)}{\delta}$$

and otherwise $n_i(t) \leq 3tp_i$. □

5.9 Ancillary Technical Lemmas

Lemma 5.9.1 (Regret Bound for OFUL). *Assume* OFUL*(Algorithm 13) uses regularization parameter $\lambda > 0$ chooses the each action as*

$$a_t \in \underset{a \in \mathcal{A}_t}{\operatorname{argmax}} \langle \widehat{\theta}_t, a\rangle + \beta_t \|a\|_{V_t^{-1}},$$

where θ_t is a parameter estimate, $\beta_t \in \mathcal{R}$ is a confidence width and $V_t \succcurlyeq \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$ is a covariance matrix. In the event that the true parameter $\theta_\star$ was contained at *all times in the confidence ellipsoid, that is,* $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$ *for all* $t \in [T]$*, the (pseudo-)regret is bounded as*

$$\mathsf{Reg}(T) \leq 2\beta_{\max}\sqrt{dT\left(1 + \frac{L^2}{\lambda}\right)\ln\frac{d\lambda + TL^2}{d\lambda}},$$

where $\beta_{\max} = \max_{t\in[T]}\beta_t$ *is the largest confidence width during all rounds and* $L = \max_{a\in\bigcup_t \mathcal{A}_t}\|a\|_2$ *be a bound on the action norms.*

Remark 5.9.2. *This regret bound for OFUL holds for any, possibly random, sequence of confidence widths as long as the true parameter is contained in the confidence ellipsoid. It does not assume any specific form or monotonicity or* $\beta_t \geq 1$*. It also does not prescribe that the covariance matrix exactly matches* $\lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$*. This makes this regret bounds applicable to the case where* $\hat{\theta}_t$ *includes additional observations besides the ones from previous rounds played by the algorithm.*

Proof. The immediate regret at time t (defined as the difference of the expected reward of the optimal action choice $a_t^\star \in \operatorname{argmax}_{a\mathcal{A}_t}\langle\theta_\star, a\rangle$ and the action a_t taken by the algorithm) is bounded as

$$\begin{aligned}
\langle\theta_\star, a_t^\star - a_t\rangle &\overset{(i)}{\leq} \langle\widehat{\theta}_t, a_t^\star\rangle + \beta_t\|a_t^\star\|_{V_t^{-1}} - \langle\theta_\star, a_t\rangle \\
&\overset{(ii)}{\leq} \langle\widehat{\theta}_t, a_t\rangle + \beta_t\|a_t\|_{V_t^{-1}} - \langle\theta_\star, a_t\rangle \\
&\overset{(iii)}{\leq} 2\beta_t\|a_t\|_{V_t^{-1}} \overset{(iv)}{\leq} 2\beta_t\|a_t\|_{\Sigma_t^{-1}},
\end{aligned}$$

where $\Sigma_t = \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$. Step (i) follows from $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, step (ii) from the algorithm's action choice and step (iii) again from the confidence ellipsoid $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$. Finally, step (iv) follows from the assumption that $V_t \succcurlyeq \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top = \Sigma_t$.

Since L is a bound of the action norm and $\Sigma_t \succcurlyeq \lambda I$, we have $\|a_t\|_{\Sigma_t^{-1}} = \|\Sigma_t^{-1/2} a_t\|_2 \leq \frac{L}{\sqrt{\lambda}}$. Thus, we can bound the regret as

$$\begin{aligned}
\mathsf{Reg}(T) &\leq 2\sum_{t=1}^{T}\beta_t\|a_t\|_{\Sigma_t^{-1}} \\
&\leq 2\sqrt{\sum_{t=1}^{T}\beta_t^2}\sqrt{\sum_{t=1}^{T}\|a_t\|_{\Sigma_t^{-1}}^2} && \text{(Cauchy-Schwarz)} \\
&\leq 2\beta_{\max}\sqrt{T\sum_{i=1}^{T}\frac{L^2}{\lambda} \wedge \|a_t\|_{\Sigma_t^{-1}}^2}.
\end{aligned}$$

And therefore,

$$\begin{aligned}\mathsf{Reg}(T) &\le 2\beta_{\max}\sqrt{T\left(1+\frac{L^2}{\lambda}\right)\ln\frac{\det\Sigma_{T+1}}{\det\Sigma_1}} && \text{(Lemma 5.9.3 below)}\\ &\le 2\beta_{\max}\sqrt{dT\left(1+\frac{L^2}{\lambda}\right)\ln\frac{d\lambda+TL^2}{d\lambda}}.\end{aligned}$$

□

Lemma 5.9.3 (Elliptical potential)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$ and $V_t = V_0 + \sum_{i=1}^t x_i x_i^\top$ and $b > 0$ then*

$$\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \le \frac{b}{\ln(b+1)}\ln\frac{\det V_n}{\det V_0} \le (1+b)\ln\frac{\det V_n}{\det V_0}.$$

Proof Sketch. The proof is identical to the usual elliptical potential lemma [41, Lemma 19.4] where $b = 1$ except that we need to argue that for any $b > 0$

$$b \wedge u \le c\ln(u+1)$$

holds whenever $c \ge \frac{b}{\ln(1+b)}$. Since $\ln(1+\cdot)$ is strictly concave and strictly monotonically increasing, it is sufficient for us to check that this inequality holds at the critical point $u = b$ which is the case. □

Lemma 5.9.4 (Randomized elliptical potential)**.** *Let $x_1, x_2, \dots \in \mathbb{R}^d$ and $I_1, I_2, \dots \in \{0,1\}$ and $V_0 \in \mathcal{R}^{d\times d}$ be random variables so that $\mathbb{E}\left[I_k | x_1, I_1, \dots, x_{k-1}, I_{k-1}, x_k, V_0\right] = p$ for all $k \in \mathbb{N}$. Further, let $V_t = V_0 + \sum_{i=1}^t I_i x_i x_i^\top$. Then*

$$\begin{aligned}\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} &\le 1 \vee 2.9\frac{b}{p}\left(1.4\ln\ln\left(2bn \vee 2\right) + \ln\frac{5.2}{\delta}\right) + \frac{2}{p}\left(1+b\right)\ln\frac{\det V_n}{\det V_0}\\ &= \frac{4}{p}(1+b)\ln\frac{\ln(2bn\vee 2)5.2\det V_n}{\delta\det V_0}\end{aligned}$$

holds with probability at least $1-\delta$ for all n simultaneously.

Proof. We decompose the sum of squares as

$$\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} = \frac{1}{p}\sum_{t=1}^n (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) + \frac{1}{p}\sum_{t=1}^n (p - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \tag{5.34}$$

The first term can be controlled using the standard elliptical potential lemma in Lemma 5.9.3 as

$$\frac{1}{p}\sum_{t=1}^{n}(bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{p}\sum_{t=1}^{n}(b \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{p}(1+b)\ln\frac{\det V_n}{\det V_0}.$$

For the second term, we apply an empirical variance uniform concentration bound. Let $\mathcal{F}_{i-1} = \sigma(V_0, x_1, I_1, \dots, x_{i-1}, I_{i-1}, x_i)$ be the sigma-field up to before the i-th indicator. Let $Y_i = \frac{1}{p}(p - I_i)\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)$ which is a martingale difference sequence because $\mathbb{E}\left[Y_i|\mathcal{F}_{i-1}\right] = 0$ and consider the process $S_t = \sum_{i=1}^t Y_i$ with variance process

$$\begin{aligned}
W_t &= \sum_{i=1}^{t}\mathbb{E}\left[Y_i^2|\mathcal{F}_{i-1}\right] = \sum_{i=1}^{t}\frac{1}{p^2}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \mathbb{E}\left[(p-I_i)^2|\mathcal{F}_{i-1}\right] \\
&= \frac{1-p}{p}\sum_{i=1}^{t}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \leq \frac{b}{p}\sum_{i=1}^{t}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \leq \frac{tb^2}{p}.
\end{aligned}$$

Note that $Y_t \leq b$ and therefore, S_t satisfies with variance process W_t the sub-ψ_P condition of [33] with constant $c = b$ (see Bennett case in Table 3 of [33]). By Lemma 5.9.5 below, the bound

$$\begin{aligned}
S_t \leq\ & 1.44\sqrt{(W_t \vee m)\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)} \\
&+ 0.41b\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)
\end{aligned}$$

holds for all $t \in \mathbb{N}$ with probability at least $1-\delta$. We set $m = \frac{b}{p}$ and upper-bound the RHS further as

$$\begin{aligned}
&1.44\sqrt{\frac{b}{p}\left(1 \vee \sum_{i=1}^{t}\left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right)\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right)} \\
&+ 0.41b\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right) \\
&\leq \frac{1}{2}\left(1 \vee \sum_{i=1}^{t}\left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right) + 1.45\frac{b}{p}\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right),
\end{aligned}$$

where the inequality is an application of the AM-GM inequality. Thus, we have shown that with probability at least $1-\delta$, for all n, the second term in (5.34) is bounded as

$$\frac{1}{p}\sum_{t=1}^{n}(p-I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{2}\left(1 \vee \sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)\right) + Z.$$

where $Z = 1.45\frac{b}{p}\left(1.4 \ln\ln(2bn \vee 2) + \ln\frac{5.2}{\delta}\right)$. And when combining all bounds on the sum of squares term in (5.34), we get that either $\sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \leq 1$ or

$$\begin{aligned}\sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) &\leq 2Z + \frac{2}{p}(1+b)\ln\frac{\det V_n}{\det V_0} \\ &\leq \frac{4}{p}(1+b)\ln\frac{\ln(2bn \vee 2) 5.2 \det V_n}{\delta \det V_0}\end{aligned}$$

which gives the desired statement. □

Lemma 5.9.5 (Uniform empirical Bernstein bound)**.** *In the terminology of [33], let $S_t = \sum_{i=1}^{t} Y_i$ be a sub-ψ_P process with parameter $c > 0$ and variance process W_t. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\begin{aligned}S_t \leq\ & 1.44\sqrt{(W_t \vee m)\left(1.4\ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)} \\ &+ 0.41c\left(1.4\ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)\end{aligned}$$

where $m > 0$ is arbitrary but fixed.

Proof. Setting $s = 1.4$ and $\eta = 2$ in the polynomial stitched boundary in Equation (10) of [33] shows that $u_{c,\delta}(v)$ is a sub-ψ_G boundary for constant c and level δ where

$$\begin{aligned}u_{c,\delta}(v) =\ & 1.44\sqrt{(v \vee 1)\left(1.4\ln\ln(2(v \vee 1)) + \ln\frac{5.2}{\delta}\right)} \\ &+ 1.21c\left(1.4\ln\ln(2(v \vee 1)) + \ln\frac{5.2}{\delta}\right).\end{aligned}$$

By the boundary conversions in Table 1 in [33] $u_{c/3,\delta}$ is also a sub-ψ_P boundary for constant c and level δ. The desired bound then follows from Theorem 1 by [33]. □

Chapter 6

Discussion and Future Directions

In this thesis we have introduced a variety of algorithmic approaches to the problem of model selection in stochastic contextual bandits and reinforcement learning. All of our approaches allow for the combination of multiple base algorithms satisfying the condition that if they were to be deployed in their natural environments, they would satisfy a high probability regret guarantee with respect to their putative regret bounds.

In Chapter 2 we introduced the Stochastic CORRAL algorithm that successfully combines an EXP3 or CORRAL adversarial master with a wide variety of stochastic base algorithms for contextual bandits and reinforcement learning. We improve the results of the original CORRAL approach [5] that requires the base algorithms to satisfy a stability condition not often fulfilled by even the simplest stochastic bandit algorithms such as UCB and OFUL.

In Chapter 4 we devise a simple model selection strategy based on the principle of equating empirical regret bounds which we call regret balancing. In Chapters 3 and 5 we introduce two distinct stochastic master algorithms Explore-Commit-Eliminate (ECE) and Regret Bound Balancing and Elimination (RBBE) based on the principle of a statistical test to detect misspecification. ECE and RBBE recover the rates of Stochastic CORRAL under an EXP3 and a CORRAL master respectively. All of our algorithms recover meaningful model selection rates in several applications, including linear bandits and MDPs with nested function classes, linear bandits with unknown misspecification, and LinUCB applied to linear bandits with different confidence parameters. Moreover, unlike Stochastic Corral we show that when applied to the problem of model selection for linear stochastic bandits, ECE and RBBE are versatile enough to also cover cases where the context information is generated by an adversarial environment. We also present three lower bounds showing A) it is impossible to distinguish between logarithmic and square root base learners, B) knowledge of the target regret guarantee is necessary for perfect model selection and C) all algorithms achieving 'perfect' model selection regret guarantees must be performing a version of regret balancing. Several questions pertaining to the potential refinement of the aforementioned approaches remain open. We detail them below.

Leveraging shared structure in meaningful ways. Perhaps due to their general purpose nature, all the algorithms we propose in this thesis do not make use of the fine grained structure of the problem at hand. For example, Stochastic CORRAL with a CORRAL master and RBBE can recover a rate of the form $\widetilde{\mathcal{O}}(d_\star^2\sqrt{T})$ for the nested linear class problem[1], where an oracle rate of $\mathcal{O}(d_\star\sqrt{T})$ is possible when the learner has knowledge of the optimal model class. It remains an open question to show if these model selection rates are not improvable for this problem or if there exists an algorithm that successfully leverages the linearity of the contexts to achieve a model selection guarantee with the same order as the oracle rate. The same question applies to all other settings where there may exist a structural relationship between the different models to select from.

Extension to adversarial bandits. The results of this thesis apply only to the setting of *stochastic* contextual bandit problems and reinforcement learning. In contrast the original CORRAL algorithm [5] can be deployed in an adversarial bandit environment provided one of the base algorithms has a valid regret guarantee and can be shown to satisfy CORRAL's stability condition. Although this condition is often satisfied for adversarial base algorithms, it remains open to show if a simpler and more interpretable approach to model selection such as the ones used in ECE, Simple Regret Balancing and RBBE can be sucessfully combined with these types of algorithms and yield valid model selection regret guarantees. Additionally, it may be possible to show that a version of the empirical regret balancing strategy introduced in Chapter 4 can be used as an approach to adversarial bandits distinct and more interpretable than existing techniques based on mirror descent.

[1] Stochastic CORRAL with an EXP3 master and ECE achieve a rate of the form $\widetilde{\mathcal{O}}(d_\star T^{2/3})$

Bibliography

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2012.

[3] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret Bounds for Policy Iteration Using Expert Prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

[4] Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret Balancing for Bandit and RL Model Selection. *arXiv preprint arXiv:2006.05491*, 2020.

[5] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a Band of Bandit Algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

[6] Alekh Agarwal, Nan Jiang, and Sham M Kakade. Reinforcement Learning: Theory and Algorithms. Technical report, Technical Report, Department of Computer Science, University of Washington, 2019.

[7] Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling Stochastic Bandit Algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.

[8] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-Exploitation Tradeoff Using Variance Estimates in Multi-Armed Bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[9] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine learning*, 47(2):235–256, 2002.

[10] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[11] Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-Probability Regret Bounds for Bandit Online Linear Optimization. In *Conference on Learning Theory*, pages 335–342. PMLR, 2008.

[12] Aurélien F Bibaut, Antoine Chambaz, and Mark J van der Laan. Rate-Adaptive Model Selection Over a Collection of Black-Box Contextual Bandit Algorithms. *arXiv preprint arXiv:2006.03632*, 2020.

[13] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems. *CoRR*, abs/1204.5721, 2012.

[14] Sébastien Bubeck and Aleksandrs Slivkins. The Best of Both Worlds: Stochastic and Adversarial Bandits. In *Conference on Learning Theory*, pages 42–1. PMLR, 2012.

[15] Alexandra Carpentier and Rémi Munos. Bandit Theory Meets Compressed Sensing for High Dimensional Stochastic Linear Bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 190–198. PMLR, 2012.

[16] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[17] Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. Online Learning with Kernel Losses. In *International Conference on Machine Learning*, pages 971–980. PMLR, 2019.

[18] Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A Simultaneously Optimal Algorithm for Multi-Armed and Linear Contextual Bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854. PMLR, 2020.

[19] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual Bandits with Linear Payoff Functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214. PMLR, 2011.

[20] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*. PMLR, 2008.

[21] Lee H Dicker. Variance Estimation in High-Dimensional Linear Models. *Biometrika*, 101(2):269–284, 2014.

[22] Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-Regret for Learning in Markov Decision Processes with Function Approximation and Low Bellman Rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

[23] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7(Jun):1079–1105, 2006.

[24] Amir-massoud Farahmand and Csaba Szepesvári. Model Selection in Reinforcement Learning. *Machine learning*, 85(3):299–332, 2011.

[25] Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[26] Dylan Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to Misspecification in Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[27] Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model Selection for Contextual Bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752, 2019.

[28] Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-Complexity Adaptive Model Selection for Stochastic Linear Bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1396–1404. PMLR, 2021.

[29] Melody Guan and Heinrich Jiang. Nonparametric Stochastic Contextual Bandits. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[30] Zhaohan Guo and Emma Brunskill. Sample Efficient Learning with Feature Selection for Factored MDPs. In *European Workshop on Reinforcement Learning*, 2018.

[31] Assaf Hallak, Dotan Di-Castro, and Shie Mannor. Model Selection in Markovian Processes. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 374–382, 2013.

[32] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight Analyses for Non-Smooth Stochastic Gradient Descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[33] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[34] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-Optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(Apr): 1563–1600, 2010.

[35] Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction Selection in Model-Based Reinforcement Learning. In *International Conference on Machine Learning*, pages 179–188. PMLR, 2015.

[36] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual Decision Processes with Low Bellman Rank are pac-Learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[37] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-Free Exploration for Reinforcement Learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[38] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably Efficient Reinforcement Learning with Linear Function Approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[39] Weihao Kong and Gregory Valiant. Estimating Learnability in the Sublinear Data Regime. In *Advances in Neural Information Processing Systems*, pages 5455–5464, 2018.

[40] Weihao Kong, Emma Brunskill, and Gregory Valiant. Sublinear Optimal Policy Value Estimation in Contextual Bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4377–4387. PMLR, 2020.

[41] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[42] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with Good Feature Representations in Bandits and in RL with a Generative Model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[43] Jonathan Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online Model Selection for Reinforcement Learning with Function Approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2021.

[44] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A Contextual Bandit Approach to Personalized News Article Recommendation. In *International conference on World wide web*, pages 661–670, 2010.

[45] Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[46] Odalric-ambrym Maillard, Daniil Ryabko, and Rémi Munos. Selecting the State-Representation in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[47] Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning. In *International Conference on Machine Learning*, pages 543–551. PMLR, 2013.

[48] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample Complexity of Reinforcement Learning Using Linearly Combined Model Ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[49] Maillard Odalric and Rémi Munos. Adaptive Bandits: Towards the Best History-Dependent Strategy. In *International Conference on Artificial Intelligence and Statistics*, pages 570–578. PMLR, 2011.

[50] Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, and Odalric-Ambrym Maillard. Regret Bounds for Learning State Representations in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2019.

[51] Ian Osband and Benjamin Van Roy. Why is Posterior Sampling Better than Optimism for Reinforcement Learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017.

[52] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) Efficient Reinforcement Learning via Posterior Sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, 2013.

[53] Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret Bound Balancing and Elimination for Model Selection in Bandits and RL. *arXiv preprint arXiv:2012.13045*, 2020.

[54] Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model Selection in Contextual Stochastic Bandit Problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 10328–10337, 2020.

[55] Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An Analysis of Linear Models, Linear Value-Function Approximation, and Feature Selection for Reinforcement Learning. In *International Conference on Machine learning*, pages 752–759. PMLR, 2008.

[56] P. Rusmevichientong and J. N. Tsitsiklis. Linearly Parameterized Bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[57] E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments. *Marketing Science*, 36(4):500–522, 2017.

[58] Yevgeny Seldin, Csaba Szepesvari, Peter Auer, and Yasin Abbasi-Yadkori. Evaluation and Analysis of the Performance of the EXP3 Algorithm in Stochastic Environments. In *European Workshop on Reinforcement Learning*, 2013.

[59] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost Optimal Algorithms for Linear Stochastic Bandits with Heavy-Tailed Payoffs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8420–8429. 2018.

[60] Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. *arXiv preprint arXiv:1904.07272*, 2019.

[61] Alexander L Strehl and Michael L Littman. An Analysis of Model-Based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[62] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[63] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

[64] Joel Tropp et al. Freedman's Inequality for Matrix Martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[65] Nicolas Verzelen, Elisabeth Gassiat, et al. Adaptive Estimation of High-Dimensional Signal-to-Noise Ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

[66] Sofía S Villar, Jack Bowden, and James Wason. Multi-Armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[67] Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On Reward-Free Reinforcement Learning with Linear Function Approximation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17816–17826, 2020.

[68] Lin Yang and Mengdi Wang. Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[69] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning Near Optimal Policies with Low Inherent Bellman Error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.