Regret Bound Balancing and Elimination for Model Selection in Bandits and RL

Aldo Pacchiano
University of California, Berkeley
pacchiano@berkeley.edu

Christoph Dann
Google Research
cdann@cdann.net

Claudio Gentile
Google Research
cgentile@google.com

Peter Bartlett
University of California, Berkeley
peter@berkeley.edu

December 25, 2020

Abstract

We propose a simple model selection approach for algorithms in stochastic bandit and reinforcement learning problems. As opposed to prior work that (implicitly) assumes knowledge of the optimal regret, we only require that each base algorithm comes with a candidate regret bound that may or may not hold during all rounds. In each round, our approach plays a base algorithm to keep the candidate regret bounds of all remaining base algorithms balanced, and eliminates algorithms that violate their candidate bound. We prove that the total regret of this approach is bounded by the best valid candidate regret bound times a multiplicative factor. This factor is reasonably small 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. We further show that, under a suitable gap-assumption, this factor only scales with the number of base algorithms and not their complexity when the number of rounds is large enough. Finally, unlike recent efforts in model selection for linear stochastic bandits, our approach is versatile enough to also cover cases where the context information is generated by an adversarial environment, rather than a stochastic one.

Contents

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., Villar et al. [2015]), online advertising (e.g., Schwartz et al. [2017]), recommendation systems (e.g., Li et al. [2010]), 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., hyperpararameter 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 paper, 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., [Lattimore and Szepesvári, 2018, 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.

2 Related Work and our Contribution

In this paper 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., Agarwal et al. [2017], Foster et al. [2019], Chatterji et al. [2020], Pacchiano et al. [2020], Arora et al. [2020], Abbasi-Yadkori et al. [2020], Foster et al. [2020], Lee et al. [2020], Bibaut et al. [2020], Ghosh et al. [2020]).

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 arbitrary $\alpha \geq 1, \beta < 1$, where $d_\star$ depends generally on the complexity of the best model class or algorithm and other problem parameters. The original Corraling Algorithm of Agarwal et al. [2017] 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 Pacchiano et al. [2020]. The authors propose a generic model selection procedure to combine stochastic bandit algorithms with an adversarial master. As opposed to the corralling algorithm of Agarwal et al. [2017], the *Stochastic Corral* method in [Pacchiano et al., 2020] 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 [Agarwal et al., 2017]. Pacchiano et al. [2020] 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., [Lattimore and Szepesvári, 2018, 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 Agarwal et al. [2017] and Pacchiano et al. [2020] 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 Agarwal et al. [2017] and Pacchiano et al. [2020], 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 Pacchiano et al. [2020] 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. Arora et al. [2020] 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 $d_1 \le \cdots \le d_M$, 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., Foster et al. [2020] 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 Foster et al. [2020] 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 [Foster et al., 2019, Chatterji et al., 2020]. 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 Foster et al. [2019]), 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 [Abbasi-Yadkori et al., 2020] 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. Abbasi-Yadkori et al. [2020] 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 [Foster et al., 2019, Chatterji et al., 2020] the Regret Balancing approach of Abbasi-Yadkori et al. [2020] 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 [Abbasi-Yadkori et al., 2011], we are able to obtain results when model selecting among multiple base algorithms with different regret guarantees.

In Lee et al. [2020] the authors propose ECE (Explore Commit Exploit), a model selection algorithm on stochastic contextual bandit algorithms. ECE can be thought of as 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 Pacchiano et al. [2020]. Lee et al. [2020] also present gap-dependent guarantees under the same assumptions as in Arora et al. [2020] (see also Bibaut et al. [2020]): 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, Ghosh et al. [2020] 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.

2.1 Content of the paper

Building on Abbasi-Yadkori et al. [2020], we study a general regret bound balancing and elimination algorithm (Section 4) 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) that can be specialized to either the gap-independent (Section 5.1) or the gap-dependent (Section 5.2) case. Then, we specialize to relevant application examples with

[1] Technically speaking the methods in [Foster et al., 2019, Chatterji et al., 2020] do not eliminate base algorithms, but reject a statistical hypothesis on the base algorithms' model complexity.

nested model classes (Section 6) that consider linear contextual bandits or linear Markov decision processes as base learners (Section 6.2 and 6.3). We also consider therein the unknown misspecification case (Section 6.4), as well as the practically relevant problem of optimally tuning linear contextual bandit algorithms like OFUL (Section 6.5). Finally, we specifically focus on the nested linear contextual bandit setting of Section 6.2, and extend our balancing and elimination technique to the case where the context information is generated adversarially (Section Section 7). Despite we do not show this explicitly, similar extensions can be exhibited for the scenarios we consider in Section 6.3, 6.4, and 6.5.

In the next section, we introduce our basic setup and notation for stochastic contexts. For the adversarial context case, further elements of the setup with be given in Section 7. Most of our proofs are provided in the appendix.

3 Setup and Assumptions

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

$$\mathsf{Reg}_i(t) = \sum_{k \in T_i(t)} \left(\max_{\pi' \in \Pi} \mathbb{E}[r_k | \pi'(x_k), x_k] - \mathbb{E}[r_k | \pi_k(x_k), x_k] \right) , \tag{1}$$

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 $R_i \colon \mathbb{N} \to \mathbb{R}_+$, 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

[2] This assumption will actually be relaxed in Section 7.

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.,

$$0 \leq R_i(n) - R_i(n-1) \leq 1 \ , \tag{2}$$

for all number of rounds $n \in \mathbb{N}$ and base learner $i \in [M]$, with $R_i(0) = 0$.

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 subscripts) 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

$$\mu_t^\star = \max_{\pi' \in \Pi} \mathbb{E}[r | \pi'(x_t), x_t]$$

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.

4 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:

$$\mathsf{Reg}(T) = \sum_{i=1}^{M} \mathsf{Reg}_i(T) \leq \sum_{i=1}^{M} R_i(n_i(T)) \approx M \min_{i\in[M]} R_i(n_i(T)) \leq M \min_{i\in[M]} R_i(T) \ .$$

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:

$$\mathcal{G} = \left\{ \forall i \in [M],\ \forall t \in \mathbb{N} \colon |n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t)| \leq c\sqrt{n_i(t)\, \ln \frac{M \ln n_i(t)}{\delta}} \right\} .$$

We show in Lemma A.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 $\sum_{k\in T_i(t)} \mathbb{E}[r_k|\pi_k(x_k), x_k]$, while $\sum_{k\in T_i(t)} \max_{\pi'\in\Pi} \mathbb{E}[r_k|\pi'(x_k), x_k]$ 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 :

$$\mu^\star \geq \frac{U_i(t)}{n_i(t)} - c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} \ . \tag{3}$$

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 $\mu^\star$:

$$\mu^\star \leq \frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \ . \tag{4}$$

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 1. 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.

Algorithm 1: Regret Bound Balancing and Elimination Algorithm

```
$\mathcal{I}_1 \leftarrow [M]$;                                  // set of active learners
$U_i(0) = n_i(0) = 0$ for all $i \in [M]$
for round $t = 1, 2, \ldots, T$ do
    Pick the base learner as $i_t \in \mathrm{argmin}_{i \in \mathcal{I}_t} R_i(n_i(t-1))$
    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
            $\mathcal{I}_{t+1} \leftarrow \mathcal{I}_{t+1} \setminus \{i\}$
```

5 Regret Analysis

We first derive a general upper-bound on the regret of Algorithm 1 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 5.1. *With probability at least $1-\delta$, the total regret of Algorithm 1 is bounded for all rounds T as follows:*

$$\begin{aligned}\mathsf{Reg}(T) \leq \sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M \\ + 2c\sum_{i\in\mathcal{B}}\left(1+\sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right)\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}}\ ,\end{aligned} \tag{5}$$

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.

In order to prove this statement, we first show that Algorithm 1 indeed keeps all candidate regret bounds approximately equal (Lemma 5.2) 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 5.3).

Lemma 5.2 (Regret Bound Balancing)**.** *In Algorithm 1, 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 5.3. *In Algorithm 1 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*

$$\mathsf{Reg}_i(t) \le 1 + \left(\frac{n_i(t)}{n_\star(t)} + 1\right) R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}}\ , \tag{6}$$

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] $\star \in \mathcal{W}$,

$$\frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \ge \frac{U_\star(t)}{n_\star(t)} - c\sqrt{\frac{\ln(M \ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Subtracting $\mu^\star$ from both sides and rearranging terms gives

$$\mu^\star - \frac{U_i(t)}{n_i(t)} - c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \le \mu^\star - \frac{U_\star(t)}{n_\star(t)} + c\sqrt{\frac{\ln(M \ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Applying the definition of $\mathcal{G}$, we obtain an inequality in terms of pseudo-regrets:

$$\frac{\mathsf{Reg}_i(t)}{n_i(t)} - 2c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \le \frac{\mathsf{Reg}_\star(t)}{n_\star(t)} + 2c\sqrt{\frac{\ln(M \ln n_\star(t)/\delta)}{n_\star(t)}}\ .$$

Multiplying both terms by $n_i(t)$ and rearranging terms gives

$$\mathsf{Reg}_i(t) \le 2c\sqrt{\ln \frac{M \ln n_i(t)}{\delta} n_i(t)} + R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)}\mathsf{Reg}_\star(t) + 2c\sqrt{\ln \frac{M \ln n_\star(t)}{\delta}}\sqrt{\frac{n_i(t)}{n_\star(t)}}\ .$$

We now upper-bound the RHS by (i) replacing $\ln n_\star(t) \le \ln t$ 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 5.2 to replace $R_i(n_i(t))$ by $R_\star(n_\star(t)) + 1$. This yields

$$\mathsf{Reg}_i(t) \le 1 + \left(1 + \frac{n_i(t)}{n_\star(t)}\right) R_\star(n_\star(t)) + 2c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}}\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\ ,$$

which is the claimed bound. □

We are now ready to prove Theorem 5.1.

[3] Recall that, under $\mathcal{G}$, any $\star \in \mathcal{W}$ will remain active.

Presumed Bounds R_i	Regret Guarantee of Algorithm 1	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
$d_i C n^{2/3}$	$(M + B^{1/3}\sqrt{d_\star}) d_\star C T^{2/3} + d_\star^{3/4}\sqrt{BT}$	Theorem 5.4
$d_i C \sqrt{n}$	$(M + \sqrt{B} d_\star) d_\star C \sqrt{T}$	Theorem 5.4
$d_i C \sqrt{n \ln \frac{n}{\delta}}$	$(M + \sqrt{B} d_\star) d_\star C \sqrt{T \ln \frac{T}{\delta}}$	Theorem A.6
$\epsilon_i n + C\sqrt{n}$	$M(\epsilon_* T + C\sqrt{T}) + MC^2$	Theorem 5.5

Table 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.

Proof of Theorem 5.1. 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 5.3 on $\mathsf{Reg}_i(t_i)$ for all $i \in \mathcal{B}$ and the balancing condition from Lemma 5.2 on the regret-bound for $i \in \mathcal{W}$ gives the desired bound. Finally, Lemma A.1 in Appendix A shows that event $\mathcal{G}$ has probability at least $1 - \delta$. □

The general regret bound contained in Theorem 5.1 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 Section 5.1 and in Section 5.2.

5.1 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

$$R_i(n) = d_i C n^\beta \wedge n \ , \tag{7}$$

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 3, 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. *If Algorithm 1 is used with candidate regret bounds in Equation (7), then its total regret is bounded with probability at least $1 - \delta$ for all T as*

$$\mathsf{Reg}(T) \leq \left(M + 2B^{1-\beta} d_\star^{\frac{1}{\beta}-1}\right) d_\star C T^\beta + 5 d_\star^{\frac{1}{2\beta}} c \sqrt{BT \ln \frac{M \ln T}{\delta}} + 2M,$$

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 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 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 Pacchiano et al. [2020] 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 A.6 in Appendix A.

We defer the full proof of Theorem 5.4 to Appendix A, 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. The first term of the general regret bound from Theorem 5.1 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. The second term in Theorem 5.1 can be controlled as follows:

$$\begin{aligned}
&\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) \overset{(i)}{\leq} \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} C d_\star \sqrt{n_\star(t_i)} = C d_\star \sum_{i \in \mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i)} \\
&\overset{(ii)}{\leq} C d_\star \sqrt{\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{\sum_{i \in \mathcal{B}} n_i(t_i)} \overset{(iii)}{\leq} C d_\star \sqrt{2B d_\star^2} \sqrt{t_i} \leq C d_\star^2 \sqrt{2BT}\ ,
\end{aligned}$$

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 5.2,

$$d_i C\sqrt{n_i(t_i)} = R_i(n_i(t_i)) \leq R_\star(n_\star(t_i)) + 1 \overset{(iv)}{\leq} 2R_\star(n_\star(t_i)) \leq 2d_\star C\sqrt{n_\star(t_i)},$$

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 to the case where candidate regret bounds are linear in n ($\beta = 1$), then the total regret of Algorithm 1 is of order $\tilde{O}(MCd_\star T)$, 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 1 is used with contextual bandits or linear MDP algorithms with misspecified function classes (see Section 6).

Theorem 5.5. *Let the candidate regret bounds for all M base learners be of the form*

$$R_i(n) = C_1\sqrt{n} + \epsilon_i C_2 n \wedge n, \tag{8}$$

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 1 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, but it makes heavy use of a case-by-case analysis of the different regimes of candidate regret bounds provided in Lemma A.9 in Appendix A. □

5.2 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 1 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. (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 1 is triggered after playing learner i for a certain number of times. This allows us to prove the following gap-dependent regret-guarantee:

Theorem 5.6. *Assume Algorithm 1 is used with candidate regret bounds in Equation (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*

$$\mathsf{Reg}(T) = O\left(M d_\star C T^\beta + \sum_{i \in \mathcal{B}} C \left((2d_\star)^{\frac{1}{\beta} + \frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}} \right) \left[\frac{20C}{\Delta_i} \ln \frac{M \ln T}{\delta} \right]^{\frac{1}{2\alpha-1}} \right) ,$$

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

$$\mathsf{Reg}(T) = O\left(M d_\star C T^\beta + \sum_{i \in \mathcal{B}} C \sqrt{\ln \frac{M \ln T}{\delta}} \left(d_\star^{\frac{1}{\beta} + \frac{1}{\alpha - \beta}} + d_\star d_i^{\frac{\beta}{\alpha-\beta}} \right) \left[\frac{20C}{\Delta_i} \right]^{\frac{\beta}{\alpha-\beta}} \right) .$$

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 2 contains a summary of the guarantees in Theorem 5.6 for the special case where $\alpha = 1$ and $\beta = \frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$. Comparing Theorem 5.6 to Theorem 5.4 (or Table 2 to Table 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.

Presumed Bounds R_i	Gap-Dependent Regret Guarantee of Algorithm 1	Proof
$d_i C n^{1/3}$	$M d_\star C T^{1/3} + \sum_{i \in \mathcal{B}} \frac{C^2(d_\star^6 + d_\star d_i)}{\Delta_i} \ln \frac{M \ln T}{\delta}$	Theorem 5.6
$d_i C n^{2/3}$	$M d_\star C T^{2/3} + \sum_{i \in \mathcal{B}} \frac{C^3(d_\star^{4.5} + d_\star d_i^2)}{\Delta_i^2} \sqrt{\ln \frac{\ln T}{\delta}}$	Theorem 5.6
$d_i C \sqrt{n}$	$M d_\star C \sqrt{T} + \sum_{i \in \mathcal{B}} \frac{C^2(d_\star^4 + d_\star d_i)}{\Delta_i} \sqrt{\ln \frac{\ln T}{\delta}}$	Theorem 5.6
$d_i C \sqrt{n \ln \frac{n}{\delta}}$	$M d_\star C \sqrt{T \ln \frac{T}{\delta}} + \sum_{i \in \mathcal{B}} \frac{C^2(d_\star^4 + d_\star d_i)}{\Delta_i} \ln^{3/4} \frac{MT}{\delta} \ln^{3/2} \frac{\ln T}{\delta}$	Theorem A.8

Table 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.

Further, while the additive term in Table 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 2) so that the additional $\sqrt{T}$-term occurring in Table 1 can be avoided. This result is in contrast with existing approaches such as Pacchiano et al. [2020], where the $\sqrt{T}$ dependence cannot be avoided.

6 Example Applications

6.1 Brief Review of Contextual Linear Bandits and the OFUL Algorithm

One important application of the methods we presented in Section 4 and Section 5 is the setting of contextual linear bandits, which we now briefly review. To keep 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 7, 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 $a_t^\star = \operatorname{argmax}_{a \in \mathcal{A}_t} \langle a, \theta_\star \rangle$. The learner's objective is to control

Algorithm 2: OFUL [Abbasi-Yadkori et al., 2011]

```
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 $\Sigma_t$
    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 $r_t = \langle a_t, \theta_\star \rangle + \xi_t$ .
```

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 [Abbasi-Yadkori et al., 2011] shown in Algorithm 2. 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 collected so far:

$$\hat{\theta}_t := \Sigma_t^{-1} \left(\sum_{\ell=1}^{t-1} a_\ell \, r_\ell \right) \quad \text{where} \quad \Sigma_t = \lambda \mathbb{I} + \sum_{\ell=1}^{t-1} a_\ell a_\ell^\top \ . \tag{9}$$

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

$$\mathcal{C}_t := \{\theta \, : \, \|\theta - \hat{\theta}_t\|_{\Sigma_t} \le \beta_t\} \tag{10}$$

that should contain the optimal parameter $\theta_\star$. We will discuss a choice of the (possibly data-dependent) scaling factor $\beta_t \in \mathbb{R}_+$ below that ensures that this happens in all rounds with high probability. Algorithm 2 now plays any action that achieves highest expected return in what we refer to as the optimistic model

$$\tilde{\theta}_t = \operatorname*{argmax}_{\theta \in \mathcal{C}_t} \max_{a \in \mathcal{A}_t} \langle a, \theta \rangle \ . \tag{11}$$

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

$$\mathcal{E} = \{\theta_* \in \mathcal{C}_t, \quad \forall t \in \mathbb{N}\} \ . \tag{12}$$

In this event $\mathcal{E}$, the optimistic model $\tilde{\theta}_t$ indeed gives rise to an optimistic estimate of the expected reward in each round

$$\langle a_t, \tilde{\theta}_t \rangle \geq \max_{a \in \mathcal{A}_t} \langle a, \theta_\star \rangle = \langle a_t^\star, \theta_\star \rangle \ . \tag{13}$$

Abbasi-Yadkori et al. [2011] show that the following choice for β_t is sufficient to make $\mathcal{E}$ happen with high probability:

Lemma 6.1 (Theorem 1 in Abbasi-Yadkori et al. [2011])**.** *For any $\delta \in (0,1)$, let the confidence scaling be*

$$\beta_t := \sqrt{2\sigma^2 \ln\left(\frac{\det(\Sigma_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S \leq \sqrt{\sigma^2 d \ln\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S \tag{14}$$

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 2 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 C.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.

6.2 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. Foster et al., 2019, Pacchiano et al., 2020], the learner chooses among actions $\mathcal{A}_t \subseteq \mathbb{R}^{d_{\max}}$ 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

[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.

computes the least-squares estimate and confidence ellipsoid in $\mathbb{R}^{d_i}$. Based 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 A.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 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 5.6 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 6.2. *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 [Foster et al., 2019]. 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 paper this is an impossible goal to achieve since, unlike Foster et al. [2019], 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 Rusmevichientong and Tsitsiklis [2010]).*

6.3 Linear Markov Decision Processes with Nested Model Classes

We can instantiate the regret bound in Theorem 5.4 ($\beta = 1/2$) to the episodic linear MDP setting of Jin et al. [2020], again with nested feature classes of doubling dimension, as in Section 6.2. Here, each round t of Algorithm 1 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. Jin et al. [2020] 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 1 by H. By Theorem 5.4 the total regret of Algorithm 1 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 5.6 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.

6.4 Linear Bandits and MDPs with Unknown Approximation Error

Zanette et al. [2020] 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] $\tilde{O}(Hd\sqrt{T} + H\sqrt{d}\epsilon T)$ 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 Zanette et al. [2020]. Unfortunately, in most 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 1 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

[5] The $\tilde{O}$ notation is similar to the O-notation but hides poly-logarithmic dependencies.

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 5.5, the total regret of Algorithm 1 with these base learners is

$$\begin{aligned}\mathsf{Reg}(T) &= O\left(MCH(d\sqrt{T} + \sqrt{d}\epsilon_\star T)\sqrt{\ln\frac{M\ln T}{\delta}} + BC^2H^2d^2\right)\\ &= \tilde{O}\left(Hd\sqrt{T} + H\sqrt{d}\epsilon_\star T + H^2d^2\right)\end{aligned}$$

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 Zanette et al. [2020] 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, Foster et al. [2020] 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 Pacchiano et al. [2020] 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 Foster et al. [2020] are (still) unavailable. Note also that our framework is not specific to instances of the algorithm by Zanette et al. [2020] as base learners. Our model selection algorithm can, for example, also be used with approximate versions of LSVI-UCB by Jin et al. [2020] and achieve similar regret guarantees in their setting and for their notion of approximation error.

6.5 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 6.1 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 $\beta_1, \ldots, \beta_T$ 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 6.3. *Denote by $\bar{\beta}_t$ the confidence-parameter choice from Lemma 6.1 and let $\kappa \in \mathbb{R}_+$ 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 6.1 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 C.1 in Appendix C 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 1 with M instances of OFUL as base learners, 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 (with $\beta = 1/2$ therein), the regret of Algorithm 1 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.

[6] For simplicity of presentation, we set here $\lambda = 1$ and disregarded the dependence on other parameters like L, S, and σ.

7 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 from Section 6.2, but our technique for adversarial contexts can be easily adapted to all other bandit applications discussed in Section 6 and likely to episodic MDP settings with adversarial start states as well.

Let us briefly recall the setting from Section 6.2. 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 $i_\star, i_\star + 1, \dots M$ are well-specified with high probability. We focus our analysis on the event $\mathcal{E}$ where this is the case. Unlike Section 6.2 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 1, 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 4 which operates in epochs where the subroutine in Algorithm 3 is executed. We start by discussing the regret balancing subroutine in the next section before presenting the main algorithm and its regret guarantee afterwards.

7.1 The Epoch Balancing Subroutine

This subroutine in Algorithm 3 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 con-

```
Algorithm 3: EpochBalancing
Input: set of learners $\mathcal{I}$
for round $t = 1, 2, \dots$ do
    Receive context $x_t$
    foreach learner $i \in \mathcal{I}$ do
        Ask learner $i$ for a lower bound $B_{t,i}$ on the value of its proposed action
    Sample $i_t \sim p \propto \frac{1}{z_i}$ for $i \in \mathcal{I}$          (see Equation (15))
    Play learner $i_t$ and receive reward $r_t$
    Update base learner $i_t$ with $r_t$
    Test for misspecification by checking
        $\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} < \max_{i\in\mathcal{I}} \sum_{k=1}^{t} B_{k,i}$
    if above condition is triggered then
        Return ;          // At least one learner must be misspecified
```

trolled for all rounds until it terminates.

In addition to the putative bound R_i on its regret, Algorithm 3 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. 9), the covariance matrix (Eq. 9) and the ellipsoid radius (Eq. 11) of base learner i at time t, respectively. In similar spirit,

$$a_{t,i} \in \operatorname*{argmax}_{a\in\mathcal{A}_t} \langle\widehat{\theta}_{t,i}, a\rangle + \beta_{t,i}\|a_{t,i}\|_{\Sigma_{t,i}^{-1}}$$

denotes the action that base learner i would take at time t. Note that we mean here the truncated actions and covariance matrix in $\mathbb{R}^{d_i}$ and $\mathbb{R}^{d_i \times d_i}$.

At each round t, Algorithm 3 first requests these bounds from each base learner to be later used in the misspecification test. The algorithm then selects one of the base

[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 $\theta_\star$.

learners in $\mathcal{I}$ by sampling an index $i_t \sim \text{Categorical}(p)$ from a categorical distribution with probabilities

$$p_i = \frac{1/z_i}{\sum_{j \in \mathcal{I}} 1/z_j} \,, \qquad \text{where } z_i = (d_i^2 + d_i S_i^2)\left(R_i^{\max} \wedge L_i^2\right) \qquad \text{for } i \in \mathcal{I} \,. \tag{15}$$

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 3 performs a misspecification test and terminates if this test triggers. We refer to the execution of Algorithm 3 as an epoch.

Unlike the misspecification test in Algorithm 1 which considers the hypothesis that a *specific* learner i is well specified, the misspecification test in Algorithm 3 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

$$B_{t,i} \leq \mathbb{E}[r_t | a_{t,i}, x_t] \leq \max_{a \in \mathcal{A}_t} \mathbb{E}[r_t | a, x_t] = \mu_t^\star \,.$$

Hence, the right-hand side of the misspecification test in Algorithm 3 is a lower-bound on the optimal value of all rounds to far, that is, it satisfies $\max_{j \in \mathcal{I}} \sum_{k=1}^t B_{k,j} \leq \sum_{k=1}^t \mu_t^\star$. 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, $\sum_{i \in \mathcal{I}} U_i(t) \geq \sum_{k=1}^t \mathbb{E}[r_t | a_t, x_t] - c\sqrt{t \ln \frac{\ln(t)}{\delta}}$. This implies that

$$\begin{aligned}
&\sum_{i \in \mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \\
&\geq \sum_{k=1}^t \mathbb{E}[r_t | a_t, x_t] + \sum_{i \in \mathcal{I}} R_i(n_i(t)) = \sum_{i \in \mathcal{I}} \left[\sum_{k \in T_i(t)} \mathbb{E}[r_t | a_t, x_t] + R_i(n_i(t)) \right] \\
&\geq \sum_{i \in \mathcal{I}} \left[\sum_{k \in T_i(t)} \mathbb{E}[r_t | a_t, x_t] + \mathsf{Reg}_i(t) \right] = \sum_{i \in \mathcal{I}} \sum_{k \in T_i(t)} \mu_k^\star = \sum_{k=1}^k \mu_k^\star,
\end{aligned}$$

where the last inequality holds because $R_i(n_i(t)) \geq \mathsf{Reg}_i(t)$ 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:

[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.

Theorem 7.1. *With probability at least $1-\delta$, Algorithm 3 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. (15)) are sufficiently apart, i.e., $2z_i \leq z_{i+1}$ 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 7.2. *Assume that Algorithm 3 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. (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*

$$\mathsf{Reg}(t) \leq \tilde{O}\left(\left(d_\star + \sqrt{d_\star} S_\star + |\mathcal{I}|\right) (d_\star + \sqrt{d_\star} S_\star)\sqrt{t}\right)$$

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 7.3. *Assume that Algorithm 3 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. (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*

$$\mathsf{Reg}(t) \leq \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + |\mathcal{I}|\right) (d_\star + \sqrt{d_\star} S_\star) L_\star \sqrt{t} + \sum_{i \in \mathcal{I}} L_i S_i\right) .$$

The proofs of Theorem 7.3 and Theorem 7.2 are similar to the proof of Theorem 5.1 but requires a randomized version of the standard elliptical potential lemma that we prove in Lemma C.4.

7.2 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 7.2

Algorithm 4: Regret Bound Balancing and Elimination with Adversarial Contexts

```
for $s = 1, \dots, M$ do
    EpochBalancing $(\{s, s+1. \dots, M\})$ in Algorithm 3
```

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 3 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 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 4 calls Epoch Balancing (Algorithm 3) 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 7.1). 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 4 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

$$\mathsf{Reg}(T) = \sum_{s=1}^{M} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \ . \tag{16}$$

The regret incurred within each epoch can be bound using Theorem 7.2, which yields the main result of this section:

Theorem 7.4 (Model Selection for Adversarial Contexts in Stochastic Linear Bandits)**.** *Assume that Algorithm 4 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. (15)). All base learners use regularizer $\lambda = 1$. With probability at least $1 - 3(M+1)\delta$ the total regret of Algorithm 4 is bounded for all rounds $T \in \mathbb{N}$ as*

$$\mathsf{Reg}(T) = \tilde{O}\left(\left(\sqrt{B}d_\star + \sqrt{Bd_\star}S_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{T}\right) \ ,$$

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*

$$\mathsf{Reg}(T) = \tilde{O}\left(\left(\sqrt{B}d_\star L_\star + \sqrt{Bd_\star}S_\star L_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{T} + B\sum_{i\in\mathcal{I}} L_i S_i\right) .$$

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 6.1. Further, only consider outcomes where Theorem 7.2 and Theorem 7.1 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. (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 7.1 and (ii) from Theorem 7.2 with epoch-dependent factor $C^{(s)} \leq \tilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + M\right)(d_\star + \sqrt{d_\star}S_\star)\right)$ or Theorem 7.3 with epoch-dependent factor $C^{(s)} \leq \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + M\right)(d_\star + \sqrt{d_\star}S_\star)\right)L_\star$ □

8 Conclusions

We have described and analyzed a simple and general balancing and elimination technique to perform model selection in stochastic bandit and reinforcement learning tasks. We have instantiated our general principle to a number of relevant model selection scenarios with nested model classes, ranging from contextual linear bandits to linear MPDs, from mis-specified linear bandits and MDPs to hyperparameter tuning of the contextual bandit algorithm OFUL. In all these cases, we show that the total regret of our master algorithm is bounded by the best valid candidate regret bound times a multiplicative factor. Notably, this factor becomes negligible in the presence of gaps in the regret bound guarantees across

[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 $t^{(s-1)}(T)$.

the base learners, so that in such cases we essentially recover the regret bound of the best base learner in hindsight.

Our work overcomes the limitations of previous approaches by combining ideas of a statistical test for arm elimination with regret balancing for exploration. We are able to obtain gap-dependent bounds, and go beyond the $\sqrt{MT}$ dependence of corralling methods based on adversarial master algorithms. The flexibility of our approach is also witnessed by our ability to extend the linear bandit analysis to the case of adversarial contexts by means of a randomized variant of our general balancing and elimination technique.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Y. Abbasi-Yadkori, A. Pacchiano, and M. Phan. Regret balancing for bandit and rl model selection. *arXiv preprint arXiv:2006.05491*, 2020.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. *arXiv preprint arXiv:2006.09255*, 2020.

A. F. Bibaut, A. Chambaz, and M. J. van der Laan. Rate-adaptive model selection over a collection of black-box contextual bandit algorithms. *arXiv preprint arXiv:2006.03632*, 2020.

N. Chatterji, V. Muthukumar, and P. Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854, 2020.

D. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems*, 2020.

D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752, 2019.

A. Ghosh, A. Sankararaman, and K. Ramchandran. Problem-complexity adaptive model selection for stochastic linear bandits. *arXiv preprint arXiv:2006.02612*, 2020.

S. R. Howard, A. Ramdas, J. Mc Auliffe, and J. Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint*, 2018.

J. N. Lee, A. Pacchiano, V. Muthukumar, W. Kong, and E. Brunskill. Online model selection for reinforcement learning with function approximation. *arXiv preprint arXiv:2011.09750*, 2020.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc. WWW*, pages 661–670, 2010.

A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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.

S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Stat Sci*, 30(2):199–215, 2015.

A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

A Proofs for Setting with Stochastic Contexts

Lemma A.1. *There is an absolute constant c such that the event*

$$\mathcal{G} = \left\{ \forall i \in [M],\ \forall t \in \mathbb{N} \colon |n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t)| \leq c\sqrt{\ln \frac{M \ln n_i(t)}{\delta} n_i(t)} \right\} \tag{17}$$

has probability at least $1 - \delta$

Proof. Consider a fixed $i \in [M]$ and write the LHS in the event definition as

$$\begin{aligned} & n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t) && (18)\\ &= \sum_{k \in T_i(t)} \left(\mu^\star - r_k - \max_{\pi' \in \Pi} \mathbb{E}[r_k | \pi', x_k] + \mathbb{E}[r_k | \pi_k, x_k] \right) \\ &= \sum_{k \in T_i(t)} \left(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}[r_k | \pi', x_k] \right) + \sum_{k \in T_i(t)} \left(\mathbb{E}[r_k | \pi_k, x_k] - r_k \right). && (19) \end{aligned}$$

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, $X_k = \mathbf{1}\{i_k = i\}(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}[r_k | \pi', x_k]) \in [-1, +1]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from Howard et al. [2018]. 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 Howard et al. [2018], 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

$$\left| \sum_{k \in T_i(t)} \left(\mu^\star - \max_{\pi' \in \Pi} \mathbb{E}[r_k | \pi', x_k] \right) \right| \leq 3 \vee 0.85\sqrt{n_i(k) \left(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta)\right)}$$

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

Consider now the second term in (19) and let $\mathcal{F}_t$ now be the sigma-field induced by all variables up to the reward at round $t+1$, i.e., $\sigma((\mathcal{I}_k, x_k, i_k, a_k, r_k)_{k \leq t}, \mathcal{I}_{t+1}, x_{t+1}, i_{t+1}, a_{t+1})$. Then $X_k = \mathbf{1}\{i_k = i\}(\mathbb{E}[r_k | \pi_k, x_k] - r_k) \in [-1, +1]$ 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

$$\left|\sum_{k \in T_i(t)} (\mathbb{E}[r_k|\pi_k, x_k] - r_k)\right| \leq 3 \vee 0.85\sqrt{n_i(k)\left(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta)\right)} \ .$$

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 A.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,*

$$\mathsf{Reg}_i(t) > R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)} R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} \tag{20}$$

then learner i fails the misspecification test of Algorithm 1 in event $\mathcal{G}$ and is eliminated.

Proof. After dividing Equation 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 1 fails. □

A.1 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. *If Algorithm 1 is used with candidate regret bounds in Equation (7), then its total regret is bounded with probability at least $1-\delta$ for all T as*

$$\mathsf{Reg}(T) \leq \left(M + 2B^{1-\beta} d_\star^{\frac{1}{\beta}-1}\right) d_\star C T^\beta + 5 d_\star^{\frac{1}{2\beta}} c \sqrt{BT \ln \frac{M \ln T}{\delta}} + 2M,$$

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 5.1 given by

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M + 2c \sum_{i \in \mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \ , \tag{21}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + 2M \leq M R_\star(T) + 2M \leq M d_\star C T^\beta + 2M,$$

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

$$2c \sum_{i \in \mathcal{B}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \leq 2c \sqrt{B \ln \frac{M \ln T}{\delta} \sum_{i \in \mathcal{B}} n_i(t_i)} \leq 2c \sqrt{BT \ln \frac{M \ln T}{\delta}}$$

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 (21) as

$$\begin{aligned} 2c \sum_{i \in \mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} &\leq 2c \sqrt{\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{T \ln \frac{M \ln T}{\delta}} \\ &\leq 2\sqrt{2} c d_\star^{\frac{1}{2\beta}} \sqrt{BT \ln \frac{M \ln T}{\delta}}, \end{aligned}$$

where the final step follows from Lemma A.3 with

$$\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \leq 2 \sum_{i \in \mathcal{B}} \left(1 \vee \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right) \leq 2B d_\star^{1/\beta} \ . \tag{22}$$

It only remains to bound the second term (21). Here again we make use of the pull-ratio bound from (22) to bound

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) = Cd_\star \sum_{i\in\mathcal{B}} \left(\frac{n_i(t_i)}{n_\star(t_i)}\right)^{1-\beta} n_i(t_i)^\beta$$

$$\leq Cd_\star \left(\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}\right)^{1-\beta} \left(\sum_{i\in\mathcal{B}} n_i(t_i)\right)^\beta$$

$$\leq Cd_\star \left(2Bd_\star^{1/\beta}\right)^{1-\beta} T^\beta \leq 2CB^{1-\beta} d_\star^{1/\beta} T^\beta,$$

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

$$R_i(n) = n \wedge C d_i n^{\beta_i}, \tag{23}$$

where $\beta_i \in (0, 1]$, $d_i \geq 1$ and C is a term that does not depend on i or n.

Lemma A.3 (Play ratio bound)**.** *If Algorithm 1 is used with candidate regret bounds of the form in Equation* (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 5.2, 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 5.6. *Assume Algorithm 1 is used with candidate regret bounds in Equation (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*

$$\mathsf{Reg}(T) = O\left(M d_\star C T^\beta + \sum_{i \in \mathcal{B}} C \left((2d_\star)^{\frac{1}{\beta} + \frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}} \right) \left[\frac{20C}{\Delta_i} \ln \frac{M \ln T}{\delta} \right]^{\frac{1}{2\alpha-1}} \right) ,$$

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

$$\mathsf{Reg}(T) = O\left(M d_\star C T^\beta + \sum_{i \in \mathcal{B}} C \sqrt{\ln \frac{M \ln T}{\delta}} \left(d_\star^{\frac{1}{\beta} + \frac{1}{\alpha - \beta}} + d_\star d_i^{\frac{\beta}{\alpha-\beta}} \right) \left[\frac{20C}{\Delta_i} \right]^{\frac{\beta}{\alpha-\beta}} \right).$$

Proof. Just as for the gap-independent guarantee in Theorem 5.4, we start with the general regret bound from Theorem 5.1 given by

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M + 2c \sum_{i \in \mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \ , \tag{24}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + 2M \leq M R_\star(T) + 2M \leq M d_\star C T^\beta + 2M,$$

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

$$\begin{aligned}
&\frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq C d_\star \left(\frac{n_i(t_i)}{n_\star(t_i)} \right)^{1-\beta} n_i(t_i)^\beta + \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq C d_\star Z^{1-\beta} n_i(t_i)^\beta + \left(1 + \sqrt{Z} \right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \\
&\leq C d_\star Z^{1-\beta} n_i(t_i)^\beta + 2 \sqrt{Z n_i(t_i) \ln \frac{M \ln T}{\delta}},
\end{aligned} \tag{25}$$

where $Z = 2 \vee \left(2\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}$. Further, using the gap-assumption, Lemma A.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 (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 $Z = 2$, then 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 (25) by

$$3Cd_\star Z^{1-\beta}\sqrt{n_i(t_i)\ln\frac{M\ln T}{\delta}} \leq 3Cd_\star Z^{1-\beta}\sqrt{\ln\frac{M\ln T}{\delta}}\left[\frac{5Cd_i}{\Delta_i}Z\sqrt{\ln\frac{M\ln T}{\delta}}\right]^{\frac{1/2}{\alpha-1/2}} .$$

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

$$
\begin{aligned}
&3Cd_\star (2d_\star)^{\frac{1-\beta}{\beta}} \sqrt{\ln \frac{M \ln T}{\delta}} \left[\frac{5Cd_i}{\Delta_i} \left(\frac{2d_\star}{d_i}\right)^{1/\beta} \sqrt{\ln \frac{M \ln T}{\delta}}\right]^{\frac{1/2}{\alpha-1/2}} \\
&\leq 2C(2d_\star)^{\frac{1}{\beta}} \left[\frac{5C}{\Delta_i} (2d_\star)^{1/\beta} \ln \frac{M \ln T}{\delta}\right]^{\frac{1/2}{\alpha-1/2}}
\end{aligned}
$$

Hence, the total regret is bounded is case as

$$
\mathsf{Reg}(T) = O\left(Md_\star C T^\beta + \sum_{i \in \mathcal{B}} C \left((2d_\star)^{\frac{1}{\beta} + \frac{1}{\beta(2\alpha-1)}} + d_\star d_i^{\frac{1}{2\alpha-1}}\right) \left[\frac{20C}{\Delta_i} \ln \frac{M \ln T}{\delta}\right]^{\frac{1}{2\alpha-1}}\right).
$$

□

Lemma A.4 (Gap-dependent elimination bound)**.** *Assume Algorithm 1 is used with candidate regret bound of the form in Equation (23). 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 $\alpha_i > \frac{1}{2} \vee \beta_i$, then, in event $\mathcal{G}$, learner i is played at most*

$$
n_i(T) \leq \left[\frac{2Cd_i}{\Delta_i}(1 + 2Z)\right]^{\frac{1}{\alpha_i - \beta_i}} \vee \left[\frac{4c}{\Delta_i}\left(1 + \sqrt{Z}\right)\sqrt{\ln \frac{M \ln T}{\delta}}\right]^{\frac{1}{\alpha_i - 1/2}},
$$

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 A.2 yields the following sufficient condition that learner i is eliminated at round t:

$$
\mathsf{Reg}_i(t) > R_i(n_i(t)) + \frac{n_i(t)}{n_\star(t)} R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}}. \tag{26}
$$

We now upper-bound the RHS of this sufficient condition using Lemma A.3 as

$$
\begin{aligned}
R_i(n_i(t)) &+ \frac{n_i(t)}{n_\star(t)} R_\star(n_\star(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \\
&\leq R_i(n_i(t)) + 2\frac{n_i(t)}{n_\star(t)} R_i(n_i(t)) + 2c\left(1 + \sqrt{\frac{n_i(t)}{n_\star(t)}}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \\
&\leq (1 + 2Z)\, R_i(n_i(t)) + 2c\left(1 + \sqrt{Z}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \\
&\leq (1 + 2Z)\, Cd_i n_i(t)^{\beta_i} + 2c\left(1 + \sqrt{Z}\right)\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}}.
\end{aligned}
$$

Using this upper-bound on the RHS of (26) and $\Delta_i n_i(t)^{\alpha_i}$ as a lower-bound on the LHS of (26), we can conclude that learner i gets eliminated if the following two conditions are met:

$$\frac{\Delta_i}{2} n_i(t)^{\alpha_i} > 2c\left(1+\sqrt{Z}\right)\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}$$
$$\frac{\Delta_i}{2} n_i(t)^{\alpha_i} > (1+2Z)\, C d_i n_i(t)^{\beta_i}$$

Rearranging each condition yields

$$n_i(t) > \left[\frac{4c}{\Delta_i}\left(1+\sqrt{Z}\right)\sqrt{\ln\frac{M\ln t}{\delta}}\right]^{\frac{1}{\alpha_i - 1/2}} \quad \text{and} \quad n_i(t) > \left[\frac{2Cd_i}{\Delta_i}(1+2Z)\right]^{\frac{1}{\alpha_i-\beta_i}} .$$

□

A.2 Special Case with $\sqrt{T\ln T}$ Candidate Regret Bounds

Consider the regret bound for all M base learners to be of the form

$$R_i(n) = d_i C\sqrt{n\ln_+(n/\delta))} \wedge n \tag{27}$$

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 A.5. *For the choice of candidate regret bounds in Equation (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 5.2, 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

$$d_i C\sqrt{n_i(t)\ln_+\frac{n_i(t)}{\delta})} \leq 2d_j C\sqrt{n_j(t)\ln_+\frac{n_j(t)}{\delta}}$$

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:

$$\frac{n_i(t)}{n_j(t)} \leq 4\frac{d_j^2}{d_i^2}\frac{\ln_+(n_j(t)/\delta)}{\ln_+(n_i(t)/\delta)} \leq 4\frac{d_j^2}{d_i^2}\ln_+(n_j(t)/\delta) \leq 4\frac{d_j^2}{d_i^2}\ln(t/\delta)\ . \tag{28}$$

We now refine this crude bound by considering two cases:

Case I: If $\sqrt{n_j(t)} \leq Cd_j\sqrt{\ln_+(n_j(t)/\delta)}$, then $R_j(n_j(t)_= n_j(t)$ and the balancing condition gives $n_j(t) \leq 2n_i(t)$ Plugging this in (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 (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 (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 A.6. *If Algorithm 1 is used with candidate regret bounds in Equation (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 5.1 given by

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + \sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) + 2M + 2c \sum_{i\in\mathcal{B}} \left(1 + \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}}\right) \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}}\ , \tag{29}$$

and bound the terms individually. We begin with

$$\sum_{i=1}^{M} R_\star(n_\star(t_i)) + 2M \le M R_\star(T) + 2M \le M d_\star C \sqrt{T \ln_+(T/\delta)} + 2M,$$

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

$$2c \sum_{i\in\mathcal{B}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} \le 2c \sqrt{B \ln \frac{M \ln T}{\delta} \sum_{i\in\mathcal{B}} n_i(t_i)} \le 2c \sqrt{BT \ln \frac{M \ln T}{\delta}}$$

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 (29) as

$$\begin{aligned} 2c \sum_{i\in\mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{n_i(t_i) \ln \frac{M \ln T}{\delta}} &\le 2c \sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{T \ln \frac{M \ln T}{\delta}} \\ &\le 6c \sqrt{B \ln_+ \left(4e \ln \frac{T}{\delta}\right)} d_\star \sqrt{T \ln \frac{M \ln T}{\delta}}, \end{aligned}$$

where the final step follows from Lemma A.5 with

$$\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} \le 7 \sum_{i\in\mathcal{B}} \left(1 \vee \frac{d_\star^2}{d_i^2}\right) \ln_+ \left(4e \ln \frac{t_i}{\delta}\right) \le 7 d_\star^2 B \ln_+ \left(4e \ln \frac{T}{\delta}\right) \tag{30}$$

It only remains to bound the second term (29). Here again we make use of the pull-ratio bound from (30) to bound

$$\begin{aligned} &\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)} R_\star(n_\star(t_i)) = C d_\star \sum_{i\in\mathcal{B}} \left(\frac{n_i(t_i)}{n_\star(t_i)}\right)^{1/2} n_i(t_i)^{1/2} \sqrt{\ln_+(n_\star(t_i)/\delta)} \\ &\le C d_\star \sqrt{\ln_+(T/\delta)} \sqrt{\sum_{i\in\mathcal{B}} \frac{n_i(t_i)}{n_\star(t_i)}} \sqrt{\sum_{i\in\mathcal{B}} n_i(t_i)} \le 3 C d_\star^2 \sqrt{BT \ln_+(T/\delta) \ln_+ \left(4e \ln \frac{T}{\delta}\right)}, \end{aligned}$$

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 1 when used with candidate regret bounds in Equation (27).

Lemma A.7 (Gap-dependent elimination bound)**.** *Assume Algorithm 1 is used with candidate regret bound of the form in Equation (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 A.4. □

Theorem A.8. *Assume Algorithm 1 is used with candidate regret bounds in Equation (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*

$$\begin{aligned}\mathsf{Reg}(T) \leq\ & Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M \\ & + 9Cd_\star \sum_{i\in\mathcal{B}} \ln_+\left(4e\ln\frac{T}{\delta}\right)^{\frac{1}{2}+\frac{1}{2\alpha-1}} \left(\ln_+\frac{MT}{\delta}\right)^{\frac{1}{2}+\frac{1/2}{2\alpha-1}} \left[\frac{42d_iC}{\Delta_i}\right]^{\frac{1}{2\alpha-1}} \left(1 \vee \frac{d_\star}{d_i}\right)^{1+\frac{2}{2\alpha-1}}.\end{aligned} \tag{31}$$

for $\star \in \mathcal{W}$ is any well-specified learner.

Proof. Just as for the gap-independent guarantee in Theorem A.6, we start with the general regret bound from Theorem 5.1 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 A.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

$$
\begin{aligned}
\mathsf{Reg}(T) \leq{}& Md_\star C\sqrt{T\ln_+(T/\delta)} + 2M\\
&+ 9Cd_\star\sum_{i\in\mathcal{B}}\ln_+\left(4e\ln\frac{T}{\delta}\right)^{\frac{1}{2}+\frac{1}{2\alpha-1}}\left(\ln_+\frac{MT}{\delta}\right)^{\frac{1}{2}+\frac{1/2}{2\alpha-1}}\left[\frac{42d_iC}{\Delta_i}\right]^{\frac{1}{2\alpha-1}}\left(1\vee\frac{d_\star}{d_i}\right)^{1+\frac{2}{2\alpha-1}}
\end{aligned}
$$

□

A.3 Special Case with $\epsilon_i C_2 T + C_1\sqrt{T}$ Candidate Regret Bounds

Lemma A.9. *Assume all base algorithms use regret bounds of the form (8) in Theorem 5.5. 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 5.3 which gives

$$\mathsf{Reg}_i(t) \leq 1 + R_*(n_*(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}} + \frac{n_i(t)}{n_*(t)} R_*(n_*(t)) + 2c\sqrt{\frac{n_i(t)^2}{n_*(t)}\ln\frac{M\ln t}{\delta}} \ . \tag{32}$$

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:

$$\begin{aligned}
&R_j(n) = n && \text{for } \sqrt{n} \leq \frac{C_1}{1 - C_2\epsilon_j} && \text{Phase I}\\
&R_j(n) \in [C_1\sqrt{n}, 2C_1\sqrt{n}] && \text{for } \frac{C_1}{1 - C_2\epsilon_j} < \sqrt{n} \leq \frac{C_1}{C_2\epsilon_j} && \text{Phase II}\\
&R_j(n) \in [C_2\epsilon_j n, 2C_2\epsilon_j n] && \text{for } \frac{C_1}{C_2\epsilon_j} < \sqrt{n} && \text{Phase III}
\end{aligned}$$

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 5.2 $R_i(n_i(t)) \leq 2R_*(n_*(t))$ 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 (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 (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 (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 (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 (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 (32) yields

$$\begin{aligned}\mathsf{Reg}_i(t) &\leq 1 + R_*(n_*(t)) + 4c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} + \frac{n_i(t)}{n_*(t)} 2C_1\sqrt{n_*(t)} \\ &\leq 1 + R_*(n_*(t)) + 4c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} + 2C_1\sqrt{n_i(t)}.\end{aligned}$$

□

B Proofs for Setting with Adversarial Contexts

B.1 Epoch Balancing Termination (Proof of Theorem 7.1)

Theorem 7.1. *With probability at least $1-\delta$, Algorithm 3 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 B.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 3 does not terminate. □

Lemma B.1. *Let $\delta \in (0,1)$ and consider the event*

$$\mathcal{G} = \left\{ \forall t \in \mathbb{N} \colon \left| \sum_{i \in \mathcal{I}} U_i(t) - \sum_{k=1}^{t} \mathbb{E}[r_k | a_k, x_k] \right| \leq c\sqrt{t \ln \frac{\ln(t)}{\delta}} \right\}.$$

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, $X_k = r_k - \mathbb{E}[r_k | a_k, x_k]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from Howard et al. [2018]. 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 Howard et al. [2018], 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

$$\left| \sum_{k=1}^{t} (r_k - \mathbb{E}[r_k | a_k, x_k]) \right| \leq 0.85\sqrt{t \left(\ln\ln(4t) + 0.72\ln(10.4/\delta)\right)}$$

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 B.2 (Upper-confidence bound on optimal reward)**.** *In event $\mathcal{G}$ from Lemma B.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 3 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 B.1, in the considered event, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\
&\geq \sum_{i\in\mathcal{I}} R_i(n_i(t)) + \sum_{k=1}^t \mathbb{E}[r_k|a_k, x_k] && \text{(by Lemma B.1)} \\
&\geq \sum_{i\in\mathcal{I}} \mathsf{Reg}_i(t) + \sum_{k=1}^t \mathbb{E}[r_k|a_k, x_k] && \text{(each learner is well-specified)} \\
&= \sum_{i\in\mathcal{I}} \left[\mathsf{Reg}_i(t) + \sum_{k\in T_i(t)} \mathbb{E}[r_k|a_k, x_k]\right] \\
&= \sum_{i\in\mathcal{I}} \sum_{k\in T_i(t)} \mu_k^\star = \sum_{k=1}^t \mu_k^\star. && \text{(by definition of regret)}
\end{aligned}$$

□

B.2 Regret Bound for Epoch Balancing (Proof of Theorem 7.2)

Theorem 7.2. *Assume that Algorithm 3 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. (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*

$$\mathsf{Reg}(t) \leq \tilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right)$$

Proof. We apply Theorem B.3 which immediately yields the desired bound

$$\mathsf{Reg}(t) \le \tilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right) .$$

□

Theorem 7.3. *Assume that Algorithm 3 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 $2z_i \le z_{i+1}$ (see Eq. (15)). Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \ge d_\star$ 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*

$$\mathsf{Reg}(t) \le \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i\right) .$$

Proof. We apply Theorem B.3 which yields

$$\mathsf{Reg}(t) \le \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i \ln\ln(t)\right) .$$

□

Theorem B.3 (General Regret Bound of Epoch Balancing)**.** *Assume that Algorithm 3 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 $d_j \ge d_\star$ 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*

$$\begin{aligned}
\mathsf{Reg}(t) &\le (|\mathcal{I}|\sqrt{z_\star} + z_\star\sqrt{\bar{M}})x(t)\sqrt{t} + 8.12\sum_{i\in\mathcal{I}} R_i^{\max} \ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\
&\le \sqrt{(d_\star^2 + d_\star S_\star^2)}|\mathcal{I}|\left(R_\star^{\max} \wedge L_\star\right)\sqrt{t}(2+2c)x(t) \\
&\quad + (d_\star^2 + d_\star S_\star^2)\left(R_\star^{\max} \wedge L_\star\right)^2\sqrt{\bar{M}t}(2+2c)x(t) \\
&\quad + 8.12\sum_{i\in\mathcal{I}} R_i^{\max} \ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta},
\end{aligned}$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing (i.e., $2z_i \le z_{i+1}$ 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

$$\sum_{k=1}^{t} \mu_k^\star \le \sum_{k=1}^{t} \max_{a\in\mathcal{A}_k} \left[\langle \widehat{\theta}_{k,\star}, a\rangle + \beta_{k,\star}\|a\|_{\Sigma_{k,\star}^{-1}}\right] = \sum_{k=1}^{t} \langle \widehat{\theta}_{k,\star}, a_{k,\star}\rangle + \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}.$$

Thus, we can write the total regret up to round t as

$$\begin{aligned}\mathsf{Reg}(t) &= \sum_{k=1}^{t} [\mu_k^\star - \mathbb{E}[r_k|a_k, x_k]] = \sum_{k=1}^{t} \mu_k^\star - \sum_{k=1}^{t} \mathbb{E}[r_k|a_k, x_k] \\ &\le \sum_{k=1}^{t} \mu_k^\star - \sum_{i\in\mathcal{I}} U_i(n_i(t)) + c\sqrt{t\ln\frac{\ln(t)}{\delta}},\end{aligned}$$

where the inequality holds in event $\mathcal{G}$ of Lemma B.1. If Algorithm 3 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

$$\sum_{i\in\mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t\ln\frac{\ln(t)}{\delta}} \ge \sum_{k=1}^{t} B_{k,\star}$$

Rearranging terms and plugging this inequality back into the regret bound from above yields

$$\mathsf{Reg}(t) \le \sum_{k=1}^{t} [\mu_k^\star - B_{k,\star}] + \sum_{i\in\mathcal{I}} R_i(n_i(t)) + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \tag{33}$$

We bound the first term in Equation 33 as

$$\begin{aligned}&\sum_{k=1}^{t} [\mu_k^\star - B_{k,\star}] \\ &\overset{(i)}{\le} \sum_{k=1}^{t} \left[R_\star^{\max} \wedge (\langle \widehat{\theta}_{k,\star}, a_{k,\star}\rangle + \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}) - (-R_\star^{\max} \vee (\langle \widehat{\theta}_{k,\star}, a_{k,\star}\rangle - \beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}))\right] \\ &\le \sum_{k=1}^{t} \left[2R_\star^{\max} \wedge 2\beta_{k,\star}\|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}\right] \le 2\beta_{t,\star} \sum_{k=1}^{t} \left[\frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}\right] \\ &\overset{(ii)}{\le} 2\beta_{t,\star}\sqrt{t\sum_{k=1}^{t}\left[\left(\frac{R_\star^{\max}}{\beta_{t,\star}}\right)^2 \wedge \frac{L^2}{\lambda_i} \wedge \|a_{k,\star}\|^2_{\Sigma_{k,\star}^{-1}}\right]}\end{aligned}$$

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 C.4. This yields

$$\begin{aligned}\sum_{k=1}^{t}[\mu_k^\star - B_{\star,k}] &\leq 4\beta_{t,\star}\sqrt{\frac{t}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t \vee 2)\det\Sigma_{t,\star}}{\delta\det\Sigma_{0,\star}}}\\ &\leq 4\beta_{t,\star}\sqrt{\frac{td_\star}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t\vee 2)(d_\star\lambda_\star + tL_\star^2)}{\delta d_\star\lambda_\star}}\end{aligned}$$

where $b_\star = \frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \frac{L_\star}{\sqrt{\lambda_\star}}$. For the second term in Equation 33, we apply Lemma B.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

$$\sum_{i\in\mathcal{I}}\sqrt{z_ip_i} = \sum_{i\in\mathcal{I}}\sqrt{\frac{1}{\sum_{i\in\mathcal{I}} z_i^{-1}}} = \frac{|\mathcal{I}|}{\sqrt{\sum_{i\in\mathcal{I}} z_i^{-1}}} \leq \frac{|\mathcal{I}|}{\sqrt{z_\star^{-1}}} = |\mathcal{I}|\sqrt{z_\star}$$

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 $z_1 \leq 2z_2 \leq 4z_4 \dots Mz_M$, 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

$$\begin{aligned}
\mathsf{Reg}(t) &\leq (|\mathcal{I}|\sqrt{z_\star} + z_\star\sqrt{\bar{M}})x(t)\sqrt{t} + 8.12\sum_{i\in\mathcal{I}} R_i^{\max} \ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\
&\leq \sqrt{(d_\star^2 + d_\star S_\star^2)}|\mathcal{I}|\left(R_\star^{\max} \wedge L_\star\right)\sqrt{t}(2+2c)x(t) \\
&\quad + (d_\star^2 + d_\star S_\star^2)\left(R_\star^{\max} \wedge L_\star\right)^2 \sqrt{\bar{M}}t(2+2c)x(t) \\
&\quad + 8.12\sum_{i\in\mathcal{I}} R_i^{\max} \ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta},
\end{aligned}$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing. Note that since this bound holds in the penultimate round of Algorithm 3 and the regret in the final round can be at most 1, this bound holds for all rounds t played by Algorithm 3, including the last. □

Lemma B.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 3 the following upper-bound*

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \leq 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\alpha} + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{\lambda_id_i + 3tp_iL_i^2}{\lambda_id_i}}$$

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

$$
\begin{aligned}
R_i(n_i(t)) &= \sum_{k\in T_i(t)} \left[2\beta_{k,i}\|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge R_i^{\max}\right] \\
&\leq R_i^{\max} n_i(t) \wedge 2\beta_{t,i} \sum_{k\in T_i(t)} \left(\|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge \frac{R_i^{\max}}{2\beta_{t,i}}\right) \\
&\leq R_i^{\max} n_i(t) \wedge 2\beta_{t,i} \sqrt{n_i(t) \sum_{k\in T_i(t)} \left(\|a_{k,i}\|^2_{\Sigma_{k,i}^{-1}} \wedge \left(\frac{R_i^{\max}}{2\beta_{t,i}}\right)^2 \wedge \frac{L_i^2}{\lambda_i}\right)} \\
&\leq R_i^{\max} n_i(t) \vee 2\beta_{t,i} \sqrt{d_i n_i(t)\left(1+b_i^2\right) \ln \frac{\lambda_i + n_i(t) L_i^2/d_i}{\lambda_i}}
\end{aligned}
$$

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 C.3. To control the the number of times each learner was chosen, we use Lemma B.5. This gives with probability at least $1-\alpha$ for all iterations t simultaneously $n_i(t) \leq 3tp_i \vee 8.12 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\alpha}$. This yields a regret bound of

$$
R_i(n_i(t)) \leq 8.12 R_i^{\max} \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\alpha} \quad \vee \quad 2\beta_{t,i}\sqrt{3d_i p_i t\left(1+b_i^2\right)\ln \frac{\lambda_i + 3tp_i L_i^2/d_i}{\lambda_i}}.
$$

Summing over R_i and plugging in $\beta_{t,i}$ yields

$$
\sum_{i\in\mathcal{I}} R_i(n_i(t)) \leq 8.12 \sum_{i\in\mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\alpha} + 2\sum_{i\in\mathcal{I}} \beta_{t,i}\sqrt{3d_i p_i t\left(1+b_i^2\right)\ln \frac{\lambda_i + 3tp_i L_i^2/d_i}{\lambda_i}}
$$

□

Lemma B.5. *The number of times each a learner $i \in \mathcal{I}$ has been played in Algorithm 3 after t iterations is bounded with probability at least $1-\delta$ for all $t \in \mathbb{N}$ and $i \in \mathcal{I}$ as*

$$
n_i(t) \leq \frac{3}{2} tp_i + 4.06 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\delta} \leq 3tp_i \vee 8.12 \ln \frac{5.2|\mathcal{I}|\ln(2t)}{\delta}
$$

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 Howard et al. [2018] with constant $c = 1$ (see Bennett case in Table 3 of Howard et al. [2018]). By Lemma C.5, the bound

$$
\begin{aligned}
S_t \leq 1.44 &\sqrt{(W_t \vee m)\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln \frac{5.2}{\delta}\right)} \\
&+ 0.41 \frac{L^2}{\lambda}\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln \frac{5.2}{\delta}\right)
\end{aligned}
$$

holds for all $t \in \mathbb{N}$ with probability at least $1-\delta$. We set $m = tp_i$ and upper-bound the RHS further as

$$S_t \leq 1.44\sqrt{tp_i\left(1.4\ln\ln(2t) + \ln\frac{5.2}{\delta}\right)} + 0.41\left(1.4\ln\ln(2t) + \ln\frac{5.2}{\delta}\right)$$
$$\leq \frac{tp_i}{2} + 1.45\left(1.4\ln\ln(2t) + \ln\frac{5.2}{\delta}\right),$$

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 $n_i(t) \leq 3tp_i$. □

C Ancillary Technical Lemmas

Lemma C.1 (Regret Bound for OFUL). *Assume* OFUL*(Algorithm 2) uses regularization parameter $\lambda > 0$ chooses the each action as*

$$a_t \in \operatorname*{argmax}_{a\in\mathcal{A}_t}\langle\widehat{\theta}_t, a\rangle + \beta_t\|a\|_{V_t^{-1}},$$

where θ_t is a parameter estimate, $\beta_t \in \mathbb{R}$ 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 C.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

$$\begin{aligned}
\mathsf{Reg}(T) &\leq 2 \sum_{t=1}^T \beta_t \|a_t\|_{\Sigma_t^{-1}} \\
&\leq 2 \sqrt{\sum_{t=1}^T \beta_t^2} \sqrt{\sum_{t=1}^T \|a_t\|^2_{\Sigma_t^{-1}}} && \text{(Cauchy-Schwarz)} \\
&\leq 2\beta_{\max} \sqrt{T \sum_{i=1}^T \frac{L^2}{\lambda} \wedge \|a_t\|^2_{\Sigma_t^{-1}}} \\
&\leq 2\beta_{\max} \sqrt{T \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{\det \Sigma_{T+1}}{\det \Sigma_1}} && \text{(Lemma C.3 below)} \\
&\leq 2\beta_{\max} \sqrt{dT \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{d\lambda + TL^2}{d\lambda}}.
\end{aligned}$$

□

Lemma C.3 (Elliptical potential)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $V_t = V_0 + \sum_{i=1}^t x_i x_i^\top$ and $b > 0$ then*

$$\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{b}{\ln(b+1)} \ln \frac{\det V_n}{\det V_0} \leq (1+b) \ln \frac{\det V_n}{\det V_0}.$$

Proof Sketch. The proof is identical to the usual elliptical potential lemma [Lattimore and Szepesvári, 2018, Lemma 19.4] where $b = 1$ except that we need to argue that for any $b > 0$

$$b \wedge u \leq c \ln(u+1)$$

holds whenever $c \geq \frac{b}{\ln(1+b)}$. 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 C.4 (Randomized elliptical potential)**.** *Let $x_1, x_2, \dots \in \mathbb{R}^d$ and $I_1, I_2, \dots \in \{0,1\}$ and $V_0 \in \mathbb{R}^{d \times d}$ be random variables so that $\mathbb{E}[I_k | x_1, I_1, \dots, x_{k-1}, I_{k-1}, x_k, V_0] = p$ for all $k \in \mathbb{N}$. Further, let $V_t = V_0 + \sum_{i=1}^t I_i x_i x_i^\top$. Then*

$$\begin{aligned}\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} &\leq 1 \vee 2.9\frac{b}{p}\left(1.4 \ln\ln(2bn \vee 2) + \ln\frac{5.2}{\delta}\right) + \frac{2}{p}(1+b)\ln\frac{\det V_n}{\det V_0} \\ &= \frac{4}{p}(1+b)\ln\frac{\ln(2bn \vee 2) 5.2 \det V_n}{\delta \det V_0}\end{aligned}$$

holds with probability at least $1-\delta$ for all n simultaneously.

Proof. We decompose the sum of squares as

$$\sum_{t=1}^n b \wedge \|x_t\|^2_{V_{t-1}^{-1}} = \frac{1}{p}\sum_{t=1}^n (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) + \frac{1}{p}\sum_{t=1}^n (p - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \tag{34}$$

The first term can be controlled using the standard elliptical potential lemma in Lemma C.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 $\mathbb{E}[Y_i | \mathcal{F}_{i-1}] = 0$ and consider the process $S_t = \sum_{i=1}^t Y_i$ with variance process

$$\begin{aligned}W_t = \sum_{i=1}^t \mathbb{E}[Y_i^2 | \mathcal{F}_{i-1}] &= \sum_{i=1}^t \frac{1}{p^2}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \mathbb{E}[(p - I_i)^2 | \mathcal{F}_{i-1}] \\ &= \frac{1-p}{p}\sum_{i=1}^t \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \leq \frac{b}{p}\sum_{i=1}^t \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \leq \frac{tb^2}{p}.\end{aligned}$$

Note that $Y_t \leq b$ and therefore, S_t satisfies with variance process W_t the sub-ψ_P condition of Howard et al. [2018] with constant $c = b$ (see Bennett case in Table 3 of Howard et al. [2018]). By Lemma C.5 below, the bound

$$S_t \leq 1.44\sqrt{(W_t \vee m)\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)} + 0.41b\left(1.4 \ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)$$

holds for all $t \in \mathbb{N}$ with probability at least $1 - \delta$. We set $m = \frac{b}{p}$ 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 (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 $Z = 1.45\frac{b}{p}\left(1.4 \ln\ln\left(2bn \vee 2\right) + \ln\frac{5.2}{\delta}\right)$. And when combining all bounds on the sum of squares term in (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 C.5 (Uniform empirical Bernstein bound). *In the terminology of Howard et al. [2018], 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 $t \in \mathbb{N}$

$$S_t \leq 1.44\sqrt{(W_t \vee m)\left(1.4 \ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)} + 0.41c\left(1.4 \ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)$$

where $m > 0$ is arbitrary but fixed.

Proof. Setting $s = 1.4$ and $\eta = 2$ in the polynomial stitched boundary in Equation (10) of Howard et al. [2018] shows that $u_{c,\delta}(v)$ is a sub-ψ_G boundary for constant c and level δ where

$$u_{c,\delta}(v) = 1.44\sqrt{(v \vee 1)\left(1.4 \ln\ln\left(2(v \vee 1)\right) + \ln\frac{5.2}{\delta}\right)} + 1.21c\left(1.4 \ln\ln\left(2(v \vee 1)\right) + \ln\frac{5.2}{\delta}\right).$$

By the boundary conversions in Table 1 in Howard et al. [2018] $u_{c/3,\delta}$ is also a sub-ψ_P boundary for constant c and level δ. The desired bound then follows from Theorem 1 by Howard et al. [2018]. □