Best of Both Worlds Model Selection

Aldo Pacchiano
Microsoft Research, NYC
apacchiano@microsoft.com

Christoph Dann
Google, NYC
cdann@cdann.net

Claudio Gentile
Google, NYC
cgentile@google.com

Abstract

We study the problem of model selection in bandit scenarios in the presence of nested policy classes, with the goal of obtaining simultaneous adversarial and stochastic ("best of both worlds") high-probability regret guarantees. Our approach requires that each base learner comes with a candidate regret bound that may or may not hold, while our meta algorithm plays each base learner according to a schedule that keeps the base learner's candidate regret bounds balanced until they are detected to violate their guarantees. We develop careful mis-specification tests specifically designed to blend the above model selection criterion with the ability to leverage the (potentially benign) nature of the environment. We recover the model selection guarantees of the CORRAL [3] algorithm for adversarial environments, but with the additional benefit of achieving high probability regret bounds. More importantly, our model selection results also hold simultaneously in stochastic environments under gap assumptions. These are the first theoretical results that achieve best of both world (stochastic and adversarial) guarantees while performing model selection in contextual bandit scenarios.

1 Introduction

A fundamental challenge in sequential decision-making is the ability of the learning agent to adapt to the unknown properties of the environment they interact with. While adversarial environments may require some caution, we would also like to leverage situations where more benign scenarios may be disclosed as a result of this interaction.

In the literature on multi-armed bandits, this adaptation capabilities has often taken two forms:
1. Best-of-both worlds guarantees, which were pioneered by [10] and subsequently studied by a number of authors [e.g., 29, 5, 28, 30, 33, 21]. Here, the goal is to design algorithms achieving both stochastic and adversarial environment guarantees simultaneously, without knowing the type of environment in advance. Similar in spirit is the stream of literature on stochastic rewards with adversarial corruptions [23, 17, 34, 21, 31], where an adversary is assumed to corrupt the stochastic rewards observed by the algorithm, and the regret guarantees are expected to degrade gracefully with the total amount of corruption, without knowing this amount in advance.
2. Model selection guarantees, which were initiated by [3], and subject since then to intense investigations [e.g., 15, 1, 27, 4, 16, 12, 8, 14, 22, 26, 13]. Here, we assume we have access to a pool of M base bandit algorithms each operating, say, within a different class of models or under different assumptions on the environment, and the aim is to design a bandit meta-algorithm that learns to simulate the best base algorithm in hindsight, without knowing in advance which one will be best for the environment at hand. This approach has often been used in the literature to capture bandit model selection problems. In fact, a natural instantiation of this framework is when we have a sequence of nested policy classes, and the goal is to single out the best policy within this nested family, by paying as price only the complexity of the policy class the optimal policy falls into. In some sense, Item 2 is more general than Item 1, since one may attempt to achieve a best-of-both-world performance by pooling a stochastic bandit algorithm with an adversarial bandit one, and expect the meta-algorithm on top of them to eventually learn to follow one of the two. Similarly, in the setting of

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

stochastic rewards with adversarial corruptions, one can pool together stochastic algorithms operating with increasing guessed levels of corruptions, and let the meta-algorithm learn to single out the one corresponding to the true corruption level.[1]

In this paper, we combine the two items above into a bandit algorithm that exhibits both model selection and best-of-both worlds regret guarantees simultaneously. Our framework encompasses in particular a well-known linear bandit model selection scenario, where the action set $\mathcal{A}$ is a (finite but large) subset of $\mathbb{R}^{d_M}$, for some maximal dimension $d_M > 0$, and we deal with a hierarchy of possible dimensions $d_1 < \ldots < d_M$. At time t, the learner plays an action $\mathbf{a}_t \in \mathcal{A}$ and receives as reward either $r_t = \mathbf{a}_t^\top \boldsymbol{\omega}_t$, where $\boldsymbol{\omega}_t \in \mathbb{R}^{d_M}$ is an adversarially-generated reward vector (adversarial setting), or $r_t = \mathbf{a}_t^\top \boldsymbol{\omega}$ + white noise, where $\boldsymbol{\omega} \in \mathbb{R}^{d_M}$ is a fixed but unknown reward vector (stochastic setting). Associated with dimensions $d_1 < \ldots < d_M$ are a nested family of policy classes $\Pi_1 \subset \ldots \subset \Pi_M$, and M base algorithms $\mathbb{A}_1, \ldots, \mathbb{A}_M$, where the i-th algorithm $\mathbb{A}_i$ is a linear bandit algorithm that works under the assumption that only the first d_i components of $\boldsymbol{\omega}_t$ (or $\boldsymbol{\omega}$) are nonzero. Hence $\mathbb{A}_i$ operates with dimension d_i in that all policies $\pi \in \Pi_i$ are probability distributions which are projections over action set $\mathcal{A}_i$ of the set of probability distributions in Π_M. Here, $\mathcal{A}_i$ is the projection of the full-dimensional action set $\mathcal{A}$ onto its first d_i dimensions. Notice that this implies that the policy classes are nested: $\Pi_1 \subseteq \Pi_2 \ldots \subseteq \Pi_M$. If only the first $d_{i_\star}$ dimensions ($i_\star$ being unknown to the learner) of each $\boldsymbol{\omega}_t$ (adversarial setting) or $\boldsymbol{\omega}$ (stochastic setting) are nonzero, we design an algorithm whose regret bound scales as $\text{poly}(d_{i_\star})\sqrt{T}$ in the adversarial case and, simultaneously, as $\frac{\text{poly}(d_M)\log T}{\Delta}$ if the environment happens to be stochastic. Specific consideration is given to the dependence on $d_{i_\star}$. We show through a lower bounding argument that in the stochastic case a guarantee of the form $\frac{d_M \log T}{\Delta}$, that is, where d_M replaces the more desirable factor $d_{i_\star}$, is *inevitable*, if we want to insist on obtaining a $\frac{\log T}{\Delta}$-like result. Thus our bounds reflect the best achievable regret rates depending on d_M in the stochastic case and on $\text{poly}(d_{i_\star})$ in the adversarial case.

In order to achieve these best-of-both-world results, one cannot easily rely on general corralling techniques, like the one contained in [3], since the granularity offered by adversarial aggregation algorithms is no better than $\sqrt{T}$, which is not adequate for stochastic settings; hence our choice towards the model selection technique known as *regret balancing*. Yet, even in this case, the literature does not provide off-the-shelf solutions: We first need to extend regret balancing from stochastic rewards [13, 26, 1] and corrupted stochastic rewards [31] to adversarial rewards. This involves adding extra actions to the action space at a given level of the hierarchy so as to enable the base learner operating at that level to compete with higher levels. This seems to be needed because in the adversarial case, base algorithms may even incur a negative regret, thus making it hard to compare the relative performance of algorithms operating at different levels during the regret balancing operations. Specific technical hurdles arise in the linear bandit case, where mis-specification may cause low-level base learners to behave in a maliciously erratic way. In this case, mis-specification tests have to be designed with care so as to ensure that base learners that are mis-specified but not yet eliminated incur manageable regret. At the same time, these tests should also detect as soon as possible if the environment is stochastic through ad hoc regret balancing gap estimation procedures.

For our results to hold, some technical conditions are required on the base algorithms, like a notion of (high probability) stability and a notion of action space extendability, formally defined in Section 2. We show that these conditions are fulfilled by known algorithms, like an anytime variant of the GeometricHedge.P algorithm in [7], and a high-probability variant of the EXP4 algorithm in [6].

Related work. Among the references we mentioned above, those which are most relevant to our work, as directly related to model selection and best-of-both-worlds guarantees are perhaps [4, 13, 21, 31]. In both [4] and [13], model selection regret guarantees for stochastic contextual bandits are given which take the form $\text{poly}(d_{i_\star})\sqrt{T}$. Yet, no combination of best-of-both worlds and model selection results are contained in those papers. Closer in spirit are [21] and [31].

In [21], the authors consider model selection problems on top of adversarially corrupted stochastic linear bandit problems, where the total level of corruption C is unknown in advance. It is worth stressing that this is a model selection problem which is substantially different (and actually easier, see Section 4) than ours, as the selection applies to C instead of the complexity term $R(\Pi_i)$. In fact, the model selection procedure in [21] looks very different from ours, and can be roughly seen as

[1] Recent relevant papers, working with model mis-specification instead of reward corruption, include [27, 14].

a robust version of a doubling trick applied to C. Our algorithm is more general than theirs, as it applies to settings beyond linear bandits, and the same is true of our best of both worlds procedures.

In [31], the authors also consider RL settings, and more general forms of corruptions than [21]. Like ours, their model selection algorithm follows the idea of regret balancing and elimination of [1, 13]. Yet, importantly our work applies to the fully adversarial setting with $\sqrt{T}$ regret while the corruption-robust approach by [31] suffers a linear in T regret if the world is fully adversarial ($C = \Omega(T)$). To enable this sublinear regret, several important technical challenges needed to solved that are not present in the corruption setting. Most importantly, the corruption setting can, for the most part, be handled similar to the stochastic setting. For instance, since [31] measures regret always w.r.t. to the uncorrupted environment, by definition, no learner can achieve negative regret, which entails that setting does not require linking base learners as we do here – see Section 2.

On the lower bound side, relevant papers include [32, 24], where the authors investigate the Pareto frontier of model selection for (contextual) stochastic bandits. It is shown in particular that a regret upper bound of the form $\sqrt{d_{i_\star} T}$ cannot be achieved. However, these papers do not explicitly cover gap-dependent regret guarantees. A more thorough discussion of our contributions is postponed to Section 3, after introducing our main notation and assumptions.

2 Problem Setting and Assumptions

We start off by defining the adversarial scenario. An adversarial contextual bandit problem is a repeated game between a learner $\mathbb{A}$ and an environment $\mathbb{B}$. We consider the general decision-making scenario where the learner $\mathbb{A}$ has at its disposal a class of policies $\Pi_{\mathbb{A}}$ made up of functions π of the form $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ denotes the set of probability distributions over $\mathcal{A}$. At each round t, the interaction between $\mathbb{A}$ and $\mathbb{B}$ is as follows:

1. Learner $\mathbb{A}$ selects a policy $\pi_t \in \Pi_{\mathbb{A}}$
2. Simultaneously, the environment $\mathbb{B}$ selects context $x_t \in \mathcal{X}$ and reward function $r_t : \mathcal{A} \times \mathcal{X} \to [0, 1]$, and reveals x_t to $\mathbb{A}$
3. Learner $\mathbb{A}$ takes an action $a_t \sim \pi_t(\cdot \mid x_t)$ and observes reward $o_t = r_t(a_t, x_t)$.

The regret $\mathrm{Reg}_{\mathbb{A}}(\mathcal{T}, \Pi')$ of algorithm $\mathbb{A}$ in rounds $\mathcal{T} \subseteq \mathbb{N}$ against a policy class Π' is the difference between the learner's accumulated reward in $\mathcal{T}$ and the performance of the best *fixed* policy from Π':

$$\mathrm{Reg}_{\mathbb{A}}(\mathcal{T}, \Pi') = \max_{\pi \in \Pi'} \sum_{\ell \in \mathcal{T}} \left[\mathbb{E}_{a \sim \pi} \left[r_\ell(a, x_\ell) \right] - r_\ell(a_\ell, x_\ell) \right] ,$$

where $\mathbb{E}_{a \sim \pi}[r_\ell(a, x_\ell)] = \sum_{a \in \mathcal{A}} \pi(a|x_\ell) r_\ell(a, x_\ell)$ denotes the expectation over the action drawn from the given policy π which may itself be a random quantity. For ease of notation, we will omit the comparator policy class when $\Pi' = \Pi_{\mathbb{A}}$ is the policy class of $\mathbb{A}$ and replace $\mathcal{T}$ by t when $\mathcal{T} = [t] := \{1, \ldots, t\}$. Hence, $\mathrm{Reg}_{\mathbb{A}}(t)$ is the regret of $\mathbb{A}$ against its own policy class up to round t.

The stochastic scenario we consider is similar to the above, except in the way contexts x_t and reward values $r_t(a_t, x_t)$ are generated. Specifically, $\mathbb{B}$ generates contexts x_t in an i.i.d. fashion according to a fixed (but arbitrary and unknown) distribution $\mathcal{D}$ over $\mathcal{X}$, while the reward $r_t(a_t, x_t)$ is such that for all fixed $(a, x) \in \mathcal{A} \times \mathcal{X}$, we have $\mathbb{E}[r_t(a, x)] = r(a, x)$, for some fixed function $r(\cdot, \cdot)$ in some known class of reward functions. In this case, we measure performance through pseudo-regret

$$\mathrm{PseudoReg}_{\mathbb{A}}(\mathcal{T}, \Pi') = \max_{\pi \in \Pi'} \sum_{\ell \in \mathcal{T}} \left[\mathbb{E}_{x \sim \mathcal{D}, a \sim \pi} \left[r(a, x) \right] - \mathbb{E}_{x \sim \mathcal{D}, a \sim \pi_\ell} \left[r(a, x) \right] \right] ,$$

where $\mathbb{E}_{x \sim \mathcal{D}, a \sim \pi}$ is the expectation over contexts x and the action $a \sim \pi(\cdot|x)$ drawn from a policy π that may itself be a random quantity. We say an environment $\mathbb{B}$ is stochastic with gap $\Delta > 0$ if there is a policy $\pi_\star \in \Pi$ such that

$$\mathbb{E}_{x \sim \mathcal{D}, a \sim \pi_\star} \left[r(a, x) \right] \geq \max_{\pi \in \Pi \setminus \{\pi^*\}} \mathbb{E}_{x \sim \mathcal{D}, a \sim \pi} \left[r(a, x) \right] + \Delta .$$

A notable example of the above is the following linear bandit scenario. In the adversarial case ("adversarial linear bandits") the action space $\mathcal{A}$ is a subset of $\mathbb{R}^d$, for some dimension d, the context x_t is irrelevant and, upon playing action $\mathbf{a}_t \in \mathcal{A}$, the environment generates a reward which is a *linear* function of the actions, $r_t(\mathbf{a}_t, x_t) = r_t(\mathbf{a}_t) = \mathbf{a}_t^\top \boldsymbol{\omega}_t$, where $\boldsymbol{\omega}_t$ is chosen adversarially at every round

within some known class of d-dimensional vectors. In the stochastic case ("stochastic linear bandits"), the only difference is that we simply have $r_t(\mathbf{a}_t, x_t) = r_t(\mathbf{a}_t) = \mathbf{a}_t^\top \boldsymbol{\omega} + \epsilon_t$, where $\boldsymbol{\omega}$ is a fixed and unknown vector in the class of d-dimensional vectors, and ϵ_t is a sub-Gaussian noise. Moreover, we have gap Δ if there is $\pi^\star \in \Pi$ such that $\mathbb{E}_{a\sim\pi^\star}\left[\mathbf{a}^\top \boldsymbol{\omega}\right] \geq \max_{\pi\in\Pi\setminus\{\pi^\star\}} \mathbb{E}_{a\sim\pi}\left[\mathbf{a}^\top \boldsymbol{\omega}\right] + \Delta$.

Many algorithms $\mathbb{A}$ in the literature exist that enjoy a regret bound against their policy class $\Pi_{\mathbb{A}}$, holding in all environments $\mathbb{B}$ satisfying certain conditions. If this regret bound holds, we say that $\mathbb{A}$ is *adapted* to this environment. Here, we allow algorithms to also compete against policy classes Π' other than $\Pi_{\mathbb{A}}$, and focus on algorithms that come with regret bounds of the following form.

Definition 1 (Adapted)**.** *We call $\mathbb{A}$ adapted to environment $\mathbb{B}$ and policy class Π' if, with probability at least $1-\delta$, the following bound holds simultaneously for all $t \in \mathbb{N}$:*[2]

$$\mathrm{Reg}_{\mathbb{A}}(t, \Pi') = \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{t \ln \frac{t}{\delta}}\right) .$$

The term $R(\Pi_{\mathbb{A}})$ is a (known) measure of complexity of the policy class $\Pi_{\mathbb{A}}$ used by $\mathbb{A}$.

Examples of algorithms satisfying Definition 1 include a version of the Geometric Hedge algorithm in linear bandits [7] and the EXP4 algorithm [6] in contextual bandits with finite action sets that operates over policies mapping contexts to probability distributions over actions. We will discuss them in detail below. Before introducing the model selection questions addressed in this work, we present two stronger versions of the above definition, which we call *high-probability stability* and *extendability*. These will be useful for model selection among multiple learners. As we will show with examples later, these stricter conditions can be established for several common settings under the same conditions for which adaptivity from Definition 1 is guaranteed.

Additional conditions. The first condition, *high-probability stability* (or *h-stability*) generalizes adaptivity to the case where $\mathbb{A}$ only observes a certain noisy version of the reward. To state the condition formally, consider a more general interaction protocol between $\mathbb{A}$ and the environment $\mathbb{B}$, where Step 3 above is replaced by

3a. Let $b_t \sim \mathrm{Bernoulli}(\rho)$ for a fixed and known $\rho \in (0, 1]$. Learner $\mathbb{A}$ takes an action $a_t \sim \pi_t(\cdot \mid x_t)$ and observes an importance-sampled version of the reward $o_t = b_t \frac{r_t(a_t, x_t)}{\rho}$.

This encompasses the original protocol with $\rho = 1$ as a special case. An algorithm is *h-stable* if it maintains its regret guarantee up to a $1/\sqrt{\rho}$ penalty in this more general interaction protocol:

Definition 2 (h-stability)**.** *An algorithm $\mathbb{A}$ is high probability stable **(h-stable)** in an environment $\mathbb{B}$ against policy class Π' if it satisfies for any constant $\rho \in (0, 1]$ a regret bound*

$$\mathrm{Reg}_{\mathbb{A}}(t, \Pi') = \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{t}{\rho} \ln \frac{t}{\delta}}\right)$$

that holds with probability at least $1-\delta$ simultaneously for all $t \in \mathbb{N}$.

Note that $\mathbb{A}$ is adapted to $\mathbb{B}$ and Π' if it is h-stable for $\rho = 1$. Compared to the notion of stability proposed in [3], the one in Definition 2 is more demanding, in that it requires the importance-weighted regret bound to hold with high probability, rather than in expectation. Since we aim for high-probability regret bounds in adversarial and stochastic settings, this stronger notion is natural.

The second condition, *extendability*, generalizes h-stability to environments with additional actions, as specified next. Let $\mathbb{B}$ be the original environment with action set $\mathcal{A}$ and let $\bar{\mathcal{A}}_k = \mathcal{A} \cup \{a'_1, a'_2, \ldots, a'_k\}$ be the action set extended by k extra special actions $a'_1, a'_2, \ldots, a'_k$. Further, let $\bar{\Pi} = \{\pi : \mathcal{X} \to \Delta_{\bar{\mathcal{A}}_k}\}$ be an extended version of the original policy set $\Pi = \{\pi : \mathcal{X} \to \Delta_{\mathcal{A}}\}$ that contains all policies of the original policy set and the single-action policies $\mathbf{1}\{a'_i\}$ that always choose a certain special action a'_i, i.e., $\bar{\Pi} \supseteq \Pi \cup \{\mathbf{1}\{a'_1\}, \mathbf{1}\{a'_2\}, \ldots, \mathbf{1}\{a'_k\}\}$.[3] We further allow the environment on the extended action space $\bar{\mathcal{A}}_k$ to choose any values for rewards $r_t(a'_i, x_t)$, possibly depending on the entire history and all rewards $r_t(a, x_t)$ assigned to other actions $a \neq a'_i$ in that round. We denote the set of all such extended environments by $\mathcal{B}_k(\mathbb{B}, \bar{\mathcal{A}}_k)$. An algorithm $\mathbb{A}$ is now extendable if we can run it with the extended policy set $\bar{\Pi}_k$ and it competes well against Π' and policies $\mathbf{1}\{a'_i\}$ in all such extended environments.

[2] Here and throughout, the $\mathcal{O}$-notation only hides absolute constants.

[3] Policies $\pi \in \Pi$ are naturally extended from $\Delta_{\mathcal{A}}$ to $\Delta_{\bar{\mathcal{A}}_k}$ by assigning probability 0 to all special actions a'_i.

Definition 3 (Extendability). *Consider an algorithm $\mathbb{A}$ with policy set $\Pi_{\mathbb{A}} \subseteq \mathcal{X} \to \Delta_{\mathcal{A}}$ in environment $\mathbb{B}$. We call $\mathbb{A}$ k-extendable in $\mathbb{B}$ against Π' if there is an extended policy set $\bar{\Pi}_k \supseteq \Pi_{\mathbb{A}} \cup \{\mathbf{1}\{a'_1\}, \ldots \mathbf{1}\{a'_k\}\}$ such that $\mathbb{A}$ equipped with the extended policy set $\bar{\Pi}_k$ is h-stable against $\Pi' \cup \{\mathbf{1}\{a'_1\}, \ldots \mathbf{1}\{a'_k\}\}$ in all environments $\mathbb{B}' \in \mathcal{B}(\mathbb{B}, \bar{\mathcal{A}}_k)$ that extend $\mathbb{B}$ from action space $\mathcal{A}$ to action space $\bar{\mathcal{A}}_k = \mathcal{A} \cup \{a'_1, \ldots, a'_k\}$.*

One relevant example of h-stable and extendable algorithm working in the adversarial linear bandit scenario is an anytime variant of the Geometric Hedge algorithm from [7] with exploration ruled by John's ellipsoid (e.g., [9]) – see Appendix B. Another example is a high-probability variant of the EXP4 algorithm from [6], which operates with finite sets of policies.

2.1 Model selection and best-of-both-worlds regret

Our model selection for best-of-both worlds regret guarantees can be described as follows. We have a nested family of policy classes $\Pi_1 \subseteq \ldots \subseteq \Pi_M$, with (known) complexities $R(\Pi_1) \leq \ldots \leq R(\Pi_M)$. A meta-learning algorithm $\mathbb{M}$ has access to M base algorithms $\mathbb{A}_1, \ldots, \mathbb{A}_M$, algorithm $\mathbb{A}_i$ operating with policy class Π_i. These algorithms we sometimes refer to as *base learners*. Let $i_\star$ be the smallest index of the base learner that competes against the largest policy class Π_M in the following sense:

$$i_\star = \min\{i \in [M] \colon \mathbb{A}_i \text{ is } (M-i)\text{-extendable and h-stable in } \mathbb{B} \text{ against } \Pi_M\} \ .$$

The goal of model selection is to devise a meta-algorithm $\mathbb{M}$ that has access to the base learners $\mathbb{A}_1, \ldots, \mathbb{A}_M$, and which achieves with probability $1-\delta$ a regret bound of the form[4]

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\Big(\mathrm{poly}\big(M, R(\Pi_{i_\star}), \ln R(\Pi_M)\big)\sqrt{t \ln \frac{t}{\delta}}\Big) \ ,$$

holding for all t, whenever $\mathbb{A}_M$ is h-stable and the environment $\mathbb{B}$ is adversarial. Simultaneously, if $\mathbb{B}$ is stochastic with gap Δ, then we must have

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\Big(\frac{\mathrm{poly}(M, R(\Pi_M))}{\Delta} \log \frac{t}{\delta}\Big) \ .$$

Notice that the above requirement on the pseudo-regret in stochastic environments only requires a dependence on $R(\Pi_M)$, instead of $R(\Pi_{i_\star})$. This is motivated by the fact that in stochastic environments with gap Δ it is generally impossible to obtain model selection guarantees of the form $\frac{R(\Pi_{i_\star}) \log t}{\Delta}$ – see Appendix A for a proof of this claim.

3 Summary of Our Contributions and Discussion

Our contributions can be summarized as follows. (i) We introduce an algorithm, called Arbe (Algorithm 1), for high probability model selection. This is the first high-probability model selection result for adversarial contextual bandit algorithms. Arbe satisfies a high probability guarantee as long as each of the base algorithms also satisfies one. Our algorithm takes inspiration from the balancing and elimination techniques in [13], which have been designed for stochastic contextual bandit (and RL) scenarios. Yet, as mentioned above, several technical hurdles had to be overcome in the algorithm's design to make it usable with adversarial base learners. We believe the model selection rates that our algorithm achieves are minimax optimal in their dependency on $R(\Pi_{i_\star})$. Recall that, as shown by the lower bounds of [24], it is not possible to achieve a model selection rate scaling linearly with $R(\Pi_{i_\star})$ – the best one can hope for is a quadratic dependence on $R(\Pi_{i_\star})$. In the setting of adversarial linear bandits this turns into a rate with a multiplier of the form $d_{i_\star} \log(\mathcal{A})$ instead of $\sqrt{d_{i_\star} \log(\mathcal{A})}$. In particular, when $|\mathcal{A}| = \Omega(2^{d_{i_\star}})$ or[5] $|\mathcal{A}| = \infty$, this multiplier becomes $d_{i_\star}^2$. In the simpler case of EXP4 base learners, this multiplier takes the form $\log(\Pi_{i_\star})$ instead of $\sqrt{\log(\Pi_{i_\star})}$.

(ii) We introduce the first algorithm for best-of-both-worlds model selection. By leveraging the high-probability guarantees of Arbe, we develop the first best-of-both-worlds model selection algorithm

[4] Here, $\mathrm{poly}(a, b, c)$ is a polynomial function of the three arguments separately.

[5] Despite we do not explicitly work out the details here, applying our results to the infinite arm case for adversarial linear bandits can be accomplished by a standard covering argument at each dimension d_i. This turns factor $\log(\mathcal{A})$ into one of the form $d_{i_\star} \log T$.

Algorithm 1: Arbe$(\delta, s = 1, t_0 = 0)$ Adversarial Regret Balancing and Elimination

Input: initial time t_0, index s of smallest active base learner, failure probability δ

Initialize base learners $\mathbb{A}_s, \mathbb{A}_{s+1}, \dots, \mathbb{A}_M$ with extended policy classes $\tilde{\Pi}_s, \tilde{\Pi}_{s+1}, \dots, \tilde{\Pi}_M$

Set sampling probabilities $\rho_i = \frac{R(\tilde{\Pi}_i)^{-2}}{\sum_{j=s}^{M} R(\tilde{\Pi}_j)^{-2}}$ for all $i \in \{s, \dots, M\}$, and $\rho_i = 0$ for $i < s$

for *round* $t = t_0 + 1, t_0 + 2, \dots$ **do**

  Sample base learner index $b_t \sim \text{Categorical}(\rho_1, \cdots, \rho_M)$

  Get context x_t and compute $a_t^i \sim \pi_t^i(\cdot|x_t)$, the action each base learner $\mathbb{A}_i$ proposes for x_t

  Play action $a_t = a_t^{b_t}$ (resolve linked actions if necessary) and receive reward $r_t(a_t, x_t)$

  Update all base learners $\mathbb{A}_i$ with reward $\frac{\mathbf{1}\{b_t = i\} r_t(a_t, x_t)}{\rho_i}$

  Set

$$\widetilde{\text{CRew}}_i(t_0, t) = \sum_{\ell=t_0+1}^{t} \frac{\mathbf{1}\{b_\ell = i\} r_\ell(a_\ell, x_\ell)}{\rho_i}, \quad \text{D}_i(t_0, t) = \mathcal{O}\left(\sqrt{\frac{t - t_0}{\rho_i} \ln \frac{\ln t}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln t}{\delta}\right) \tag{1}$$

  Test for all $i, j \in \{s, \dots, M\}$ with $i < j$:

$$\widetilde{\text{CRew}}_j(t_0, t) > \widetilde{\text{CRew}}_i(t_0, t) + \text{D}_i(t_0, t) + \text{D}_j(t_0, t) + R(\tilde{\Pi}_i)\sqrt{\frac{t - t_0}{\rho_i} \ln \frac{t}{\delta}} \tag{2}$$

  if test triggers for $\mathbb{A}_i$ **then restart** algorithm by running Arbe$(\delta, i + 1, t)$

(Arbe-Gap + Arbe-GapExploit, Algorithm 2 + Algorithm 4) that can retain model selection rates when the environment is adversarial, and obtain logarithmic rates when the environment is stochastic with a gap. The logarithmic gap-dependent rate of this algorithm exhibits an optimal quadratic dependence w.r.t. the largest policy class $R(\Pi_M)$ (see also Item (iii) below). Our algorithm is quite complex and requires a couple of main innovations: First, a careful design of a gap identification subroutine aimed at identifying a candidate optimal policy and gap estimator (Arbe-Gap) and, second, a very precise schedule of play for exploiting this knowledge and test its truthfulness (Arbe-GapExploit).

(iii) As already mentioned, we show via a lower bound for stochastic environments that, in the presence of a gap, perfect model selection between multiple logarithmic rate learners is impossible. This can be found in Appendix A. A dependence on the complexity of the largest class $R(\Pi_M)$ is inevitable, and this dependence must be quadratic. Our algorithms (Arbe-Gap + Arbe-GapExploit) achieve exactly this dependence when the environment is stochastic and has a gap (Theorem 5).

4 Adversarial Model Selection Using Regret Balancing

We now introduce our algorithm for model selection in adversarial bandit problems with high-probability regret guarantees. The algorithm is shown in Algorithm 1, and follows the regret balancing principle. This principle has been applied successfully to model selection in bandit and RL problems with stochastic rewards [13, 26, 1] and with corrupted stochastic rewards [31]. Our work is the first to extend this approach to adversarial rewards.

Regret Balancing. In each round, we choose the index of a base learner b_t by sampling from a categorical distribution with probabilities $\rho_s, \rho_{s+1}, \dots, \rho_M$ that remain fixed throughout the epoch. The policy of learner $\mathbb{A}_{b_t}$ is then used to sample the action a_t that is passed to the environment. After receiving the reward $r_t(a_t, x_t)$, Algorithm 1 updates each base learner $\mathbb{A}_i$ with $r_t(a_t, x_t)$ importance-weighted by the probability that the learner was selected, i.e., $\frac{\mathbf{1}\{b_t = i\} r_t(a_t, x_t)}{\rho_i}$. Thus, we update all base learners. This is closer to the Corral algorithm [3] than to the regret balancing approaches for the stochastic setting, which only update the selected learner $\mathbb{A}_{b_t}$.

If the probabilities are set to $\rho_i \propto \frac{1}{R(\Pi_i)^2}$ then after t rounds, each learner $\mathbb{A}_i$ is followed roughly $\rho_i t$ times. To see why the regrets of all learners are balanced if they are h-stable against Π_M in this case, consider the following. Denote by $\text{reg}_t(a) = [\mathbb{E}_{a \sim \pi^\star}[r_t(a, x_t)] - r_t(a, x_t)]$ the regret in round t of action a where $\pi^\star \in \text{argmax}_{\pi \in \Pi_M} \sum_{t=1}^{T} \mathbb{E}_{a \sim \pi}[r_t(a, x_t)]$ is the best policy after all rounds T. The

regret in rounds where $\mathbb{A}_i$ was selected can be bounded using standard concentration arguments as

$$\sum_{t=1}^{T} \mathbf{1}\{b_t = i\} \operatorname{reg}_t(a_t) = \rho_i \sum_{t=1}^{T} \frac{\mathbf{1}\{b_t = i\}}{\rho_i} \operatorname{reg}_t(a_t^i) \approx \rho_i \sum_{t=1}^{T} \operatorname{reg}_t(a_t^i) \leq R(\tilde{\Pi}_i)\tilde{O}\left(\sqrt{\rho_i T}\right),$$

where the final inequality holds because $\mathbb{A}_i$ is h-stable. From the definition of ρ_i we have $R(\tilde{\Pi}_i)\sqrt{\rho_i} = \left(\sum_{j=s}^{M} R(\tilde{\Pi}_j)^{-2}\right)^{-1/2} \leq R(\tilde{\Pi}_s)$, which shows that the total regret in those rounds is bounded by $\tilde{O}(R(\tilde{\Pi}_s)\sqrt{T})$. Thus, if base learners are all h-stable, then the regret incurred by each of them is comparable to the regret incurred by $\mathbb{A}_s$, the learner with the smallest complexity $R(\Pi_s)$.

Eliminating Base Learners. If a learner $\mathbb{A}_i$ is not h-stable and may have linear regret, then the probabilistic schedule above which plays this learner roughly $\rho_i t$ times yields linear regret. We therefore monitor the performance of each learner and terminate the epoch whenever a learner performs significantly worse than expected, and thus cannot be h-stable in the environment. To identify such cases, we compare estimates of the rewards of all pairs of base learners as follows. For each learner $\mathbb{A}_i$, $\widetilde{\mathrm{CRew}}_i(t_0, t)$ in Equation 1 is an unbiased estimate of the learners reward sequence $\mathrm{CRew}_i(t_0, t) = \sum_{\ell=t_0+1}^{t} r_\ell(a_\ell^i, x_\ell)$ (see the appendix for details about how these estimates are computed). Further, using confidence bounds $\mathrm{D}_i(t - t_0)$ from Equation 1, we can constructs a confidence interval for $\mathrm{CRew}_i(t_0, t)$ as $\left[\widetilde{\mathrm{CRew}}_i(t_0, t) - \mathrm{D}_i(t - t_0), \widetilde{\mathrm{CRew}}_i(t_0, t) + \mathrm{D}_i(t - t_0)\right]$. If the confidence intervals for two learners with indices $i < j$ are more than the h-stable regret bound of i apart, see Equation 2, then deem all learners with index up to i not h-stable and restart the algorithm with a reduced set of base learners.

This kind of elimination condition has already been used in stochastic environments [27, 13, 31], but it requires substantially more care in an adversarial setting. In settings with stochastic rewards (even with corruption, e.g., [31]), no learner can achieve rewards that are significantly higher than the optimal policy. In contrast, in the adversarial setting, it is possible to have negative regret against any fixed policy. Thus, if we were to use the elimination condition in Equation 2 with our base learners as is, then we may eliminate an h-stable base learner i when another base learner j has negative regret against the best fixed policy $\pi_t^\star$. This could in turn lead to undesirable regret in subsequent epochs when only learners with $R(\Pi_i) \gg R(\Pi_{i_\star})$ are left.

Linking Base Learner Performances. To address this issue, we link the performance of base learners. Instead of instantiating each learner $\mathbb{A}_i$ with its original policy set Π_i, we apply it to an extended problem with $M - i$ additional actions $\tilde{a}_{i+1}, \tilde{a}_{i+2}, \ldots, \tilde{a}_M$, and an extended policy set $\tilde{\Pi}_i$ that includes all original policies, along with policies that only choose one of the additional actions $\tilde{a}_i$, that is $\tilde{\Pi}_i \supseteq \Pi_i \cup \{\mathbf{1}\{\tilde{a}_{i+1}\}, \ldots \mathbf{1}\{\tilde{a}_M\}\}$. Whenever a base learner $\mathbb{A}_i$ chooses one of the additional actions $\tilde{a}_j$, then the action proposed by $\mathbb{A}_j$ is followed. Essentially, running each base learner with such an extended action set allows it to choose to follow the actions proposed by any learner with higher index in the hierarchy. The benefit of linking base learners this way is that each base learner now not only competes against the best fixed policy in their set, but also against all learners above them in the hierarchy. Thus, if $\mathbb{A}_i$ is h-stable and extendable (see Definition 3), then it satisfies a regret bound of the form[6]

$$\sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi_\ell^j}\left[r_\ell(a, x_\ell)\right] - \mathbb{E}_{a\sim\pi_\ell^i}\left[r_\ell(a, x_\ell)\right]\right] = \tilde{\mathcal{O}}\left(R(\tilde{\Pi}_i)\sqrt{\frac{t - t_0}{\rho_i}}\right) \tag{3}$$

against any learner $\mathbb{A}_j$ with $j > i$. As a result, since the LHS of Equation 3 is approximately $\widetilde{\mathrm{CRew}}_j(t_0, t) - \widetilde{\mathrm{CRew}}_i(t_0, t) \pm \mathrm{D}_i(t_0, t) \pm \mathrm{D}_j(t_0, t)$, the test in Equation 2 cannot trigger for such $\mathbb{A}_i$, and only base learners that are not h-stable or extendable can be eliminated.

In Appendix C, we show that our regret balancing algorithm achieves the following regret bound:

Theorem 4. *Consider a run of Algorithm 1 with* Arbe$(\delta, 1, 0)$ *and M base algorithms with $1 \leq R(\widetilde{\Pi}_1) \leq \cdots \leq R(\widetilde{\Pi}_M)$ where $\tilde{\Pi}_i$ is the extended version of policy class Π_i with $(M - i)$ additional actions. Then with probability at least $1 - \mathrm{poly}(M)\delta$ the regret $\mathrm{Reg}(t, \Pi_M)$ for all rounds $t \geq i_\star$ is bounded by*

[6] We here naturally extend the domain of r_ℓ to linked actions as $r_\ell(\tilde{a}_i, x_\ell) = r_\ell(a_\ell^i, x_\ell)$ for all $i \in [M]$.

$$\mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}+M\right)R(\widetilde{\Pi}_{i_\star})\sqrt{i_\star t\ln\frac{t}{\delta}}\right)\ , \tag{4}$$

where $i_\star$ is the smallest index of the base algorithm that is h-stable.

In most settings, the complexity $\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)} \geq \sqrt{M}$ and the first term dominates. This regret recovers the expected regret guarantees of Corral [2] when used with learning rates that do not require knowledge of $i_\star$ but as stronger high-probability bounds. To the best of our knowledge, Theorem 4 is the first high-probability regret bound for adversarial model selection. To illustrate our result, consider to common problem of model selection with nested model classes of dimensions $d_i = 2^i$ for $i \in [M]$, discussed in the introduction. We can use GeometricHedge.P as base learners which are h-stable and extendable if adapted (see Appendix B) and the complexity of extended policy classes $R(\widetilde{\Pi}_i) = R(\Pi_i) + M - i \leq d_i + M$ is not much larger than those of the original policy classes. Arbe with such base learners achieves a regret bound of order $\tilde{O}(d_{i_\star}^2\sqrt{t})$ up to factors of M and log-factors which are typically small.

5 Adversarial Model Selection with Best of Both Worlds Guarantees

In this section, we present our algorithm for adversarial model selection that preserves a logarithmic regret guarantee in case it interacts with a stochastic environment with gap $\Delta > 0$. In order to achieve such a best-of-both-worlds guarantee, we will combine our adversarial regret balancing technique with the algorithmic strategy of Bubeck and Slivkins [11]. The main result for our algorithm is:

Theorem 5. *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$. Then with probability at least $1 - poly(M)\delta$, the following two conditions hold for all $t \geq M^2$ simultaneously. In any adversarial or stochastic environment $\mathbb{B}$, the regret is bounded as*

$$\mathrm{Reg}(t, \Pi_M) = \mathcal{O}\left(\left(M + \ln(t) + \frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right)\ .$$

If $\mathbb{B}$ is stochastic and there is a unique policy with gap $\Delta > 0$, then the pseudo-regret is bounded as,

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\frac{R(\Pi_M)^2}{\Delta}\ln(t)\ln\frac{t}{\delta} + \frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln^2\left(\frac{M R(\widetilde{\Pi}_M)}{\Delta\delta}\right)\right). \tag{5}$$

The algorithm has the same regret bound as Arbe up to a $\sqrt{\ln t}$ factor. However, in addition, it also maintains poly-logarithmic pseudo-regret if the environment is stochastic and the best policy exhibits a positive gap. Our pseudo-regret bound depends polynomially on $R(\Pi_M)$ in contrast to the $R(\Pi_{i_\star})$ dependency for the adversarial regret rate. This may seem undesirable but, as mentioned in Section 2.1, model selection in stochastic environments with $\mathrm{poly}(R(\Pi_{i_\star}), \ln(R(\Pi_M))\frac{\ln(t)}{\Delta}$ regret is impossible (see Appendix A), let alone while maintaining a best-of-both-worlds guarantee with $\sqrt{T}$ regret in adversarial environments. Hence, our algorithm achieves the best kind of guarantee we can hope for, up to improvements in the order of polynomial dependencies.

Our algorithm proceeds in two distinct phases. The first, Arbe-Gap shown in Algorithm 2 is designed to identify a suitable candidate for the optimal policy and to estimate its gap. If such a policy that performs significantly better than any other policy emerges, the algorithm enters the second phase, Arbe-GapExploit which hones in on this policy by playing it most of the time while monitoring its regret in case the environment turns out to be adversarial after all (in which case we simply run Arbe). We now describe both phases.

5.1 First Phase: Candidate Policy Identification and Gap Estimation

This goal of this phase is to always maintain the desired adversarial model selection regret guarantee and simultaneously identify the gap between the best policy and the rest if one exists. To achieve the first goal, we employ the regret balancing and elimination technique of Arbe, see Lines 3–11 of Algorithm 2 which are virtually identical to Algorithm 1.

Algorithm 2: Arbe-Gap$(\delta, t, \widehat{\pi}, n, (\mathbb{A}_i)_{i=s}^M)$

1 **Input:** failure probability δ, timestep t_0, focus policy $\widehat{\pi}$, number of (re)starts n, learners $(\mathbb{A}_i)_{i=s}^M$
2 Set $\Pi_{M+1} = \Pi_M \backslash \{\widehat{\pi}\}$ and $\mathbb{A}_{M+1}$ as a copy of $\mathbb{A}_M$ with policy class Π_{M+1}
3 Initialize base learners $\mathbb{A}_s, \mathbb{A}_{s+1}, \ldots, \mathbb{A}_{M+1}$ with extended policy classes $\widetilde{\Pi}_s, \widetilde{\Pi}_{s+1}, \ldots, \widetilde{\Pi}_{M+1}$
4 Set sampling probabilities $\rho_i = \frac{R(\widetilde{\Pi}_i)^{-2}}{\sum_{j=s}^{M+1} R(\widetilde{\Pi}_j)^{-2}}$ for all $i \in \{s, \ldots, M+1\}$, and $\rho_i = 0$ for $i < s$
5 **for** *round* $t = t_0 + 1, t_0 + 2, \ldots$ **do**
6 Sample base learner index $b_t \sim \text{Categorical}(\rho_1, \cdots, \rho_M)$
7 Get context x_t and compute $a_t^i \sim \pi_t^i(\cdot|x_t)$, the action each base learner $\mathbb{A}_i$ proposes for x_t
8 Play action $a_t = a_t^{b_t}$ (resolve linked actions if necessary) and receive reward $r_t(a_t, x_t)$
9 Update all base learners $\mathbb{A}_i$ with reward $\frac{\mathbf{1}\{b_t = i\} r_t(a_t, x_t)}{\rho_i}$
10 **if** *Equation 2 holds between* $\mathbb{A}_i, \mathbb{A}_j$ *for* $s \le i < j \le M+1$ **then**
11 restart algorithm by running Arbe-Gap$(\delta, t, \widehat{\pi}, n+1, (\mathbb{A}_{i+1}, \cdots, \mathbb{A}_M))$
// Gap Test: $\mathbb{A}_M$ *better than* $\mathbb{A}_{M+1}$*?*
12 Set $\mathrm{W}(t_0, t) = \Theta\left(\sqrt{\frac{R(\widetilde{\Pi}_M)^2}{\rho_M(t-t_0)} \ln \frac{n(t-t_0)}{\delta}} + \frac{\ln \frac{n \ln(t-t_0)}{\delta}}{\rho_M(t-t_0)}\right)$
13 Set $\widehat{\Delta}_t = \frac{\widetilde{\mathrm{CRew}}_M(t_0,t) - \widetilde{\mathrm{CRew}}_{M+1}(t_0,t)}{t-t_0} - \mathrm{W}(t_0, t)$
14 **if** $2W(t_0, t) \le \widehat{\Delta}_t \le R(\widetilde{\Pi}_M)^2$ **then**
15 Run Arbe-GapExploit (Appendix D) with inputs $\delta, t_0 = t, \mathbb{A}_M, \widehat{\pi}$ and $\widehat{\Delta} = \widehat{\Delta}_t$
16 Run Arbe with inputs t_0, s, δ
// New Candidate Policy Test
17 **if** *a policy* $\pi \in \Pi_M \setminus \{\widehat{\pi}\}$ *has been selected in more than* $\frac{3t}{4}$ *of all* $t \ge 9$ *rounds* **then**
18 restart algorithm by running Arbe-Gap$(\delta, t, \pi, n+1, (\mathbb{A}_i)_{i=s}^M)$

For the second goal, determining the gap, Arbe-Gap maintains a candidate $\widehat{\pi} \in \Pi_M$ for the optimal policy and estimates its gap as follows. The learner hierarchy $\mathbb{A}_1, \ldots, \mathbb{A}_M$ is augmented at the top with an additional learner $\mathbb{A}_{M+1}$ operating on the policy class $\Pi_{M+1} = \Pi_M \setminus \{\widehat{\pi}\}$. Since $\mathbb{A}_{M+1}$ is identical to $\mathbb{A}_M$ except that it does not have access to $\widehat{\pi}$, we can obtain a gap estimate for $\widehat{\pi}$ by monitoring the difference in reward estimates $\widetilde{\mathrm{CRew}}_{M+1}$ and $\widetilde{\mathrm{CRew}}_M$ of $\mathbb{A}_{M+1}$ and $\mathbb{A}_M$. In fact, $\widehat{\Delta}_t$ in Line 13 is a lower confidence bound on the difference between the best policies in $\widehat{\Pi}_M$ and $\widehat{\Pi}_M \setminus \{\widehat{\pi}\}$ and thus, the gap of $\widehat{\pi}$. We test at every round whether $\widehat{\Delta}_t$ exceeds its confidence width $2\mathrm{W}(t_0, t)$. If this test triggers then $\widehat{\pi}$ must have a positive gap Δ of order $\mathrm{W}(t_0, t)$ and further, $\widehat{\Delta}_t$ must be a multiplicative estimate of Δ, that is, $\widehat{\Delta}_t \le \Delta \le 2\widehat{\Delta}_t$. The latter holds, because $\widehat{\Delta}_t + 2\mathrm{W}(t_0, t)$ is an upper-confidence bound on Δ and the test condition implies $2\widehat{\Delta}_t \ge \widehat{\Delta}_t + 2\mathrm{W}(t_0, t) \ge \Delta$. Since in this case, we have determined that $\widehat{\pi}$ is optimal with a gap of order $\widehat{\Delta}_t$, we can move on to the second phase Arbe-GapExploit discussed later.

Assume the candidate policy $\widehat{\pi}$ is indeed optimal and exhibits a positive gap Δ. Since $\widehat{\Delta}_t$ concentrates around Δ at a rate of $\mathrm{W}(t_0, t) \approx \frac{\mathrm{poly}(R_M(\widetilde{\Pi}_M))}{\sqrt{t-t_0}}$, the condition of the gap test in Line 14 must trigger after at most $t - t_0 \lesssim \frac{\mathrm{poly}(R_M(\widetilde{\Pi}_M))}{\Delta^2}$ rounds. Finally, since Arbe-Gap always maintains the $\mathrm{poly}(R(\widetilde{\Pi}_{i_\star})\sqrt{t-t_0}$ adversarial regret rate, the total pseudo-regret incurred until the test triggers is of order $\frac{\mathrm{poly}(R_M(\widetilde{\Pi}_M))}{\Delta}$, leading to the second term in the bound in Equation 5.

With the techniques above, we can reliably detect a positive gap if the candidate policy $\widehat{\pi}$ exhibits one. It remains to identify a suitable candidate $\widehat{\pi}$ of the optimal policy $\pi^\star$ in stochastic environments. To do so, we use the following two observations: Arbe-Gap maintains a $\mathrm{poly}(R(\widetilde{\Pi}_{i_\star})\sqrt{t}$ adversarial regret rate overall and each policy but $\pi^\star$ will incur on average at least a regret of Δ per round. Hence, in order to maintain that regret, Arbe-Gap must select $\pi^\star$ in the majority of all rounds when $t = \omega(\frac{\mathrm{poly}(R(\widetilde{\Pi}_{i_\star}))}{\Delta^2})$. Otherwise the regret grows as $\Omega(t\Delta) = \omega(\mathrm{poly}(R(\widetilde{\Pi}_{i_\star}))\sqrt{t})$ violating the

adversarial rate. To leverage this observation and identify a suitable candidate policy $\widehat{\pi}$, Line 17 of Algorithm 2 always checks whether there is a policy other than the current candidate policy that has been selected in at least $3/4$ of all rounds.[7] If so, the algorithm is restarted with this policy as candidate $\widehat{\pi}$. One can show that there are at most $\mathcal{O}(\ln t)$ restarts due to candidate policy switches, only increasing the adversarial regret rate of Arbe-Gap by a factor of $\mathcal{O}(\sqrt{\ln t})$ compared to Arbe's. Our candidate policy selection approach is similar to that by Wei et al. [31] for the corrupted reward setting but they require each adapted base learner to achieve a logarithmic regret rate in the first place. Instead, our approach only requires an adversarial regret rate from the base learner.

5.2 Second Phase: Exploitation

Since each base learner only needs to satisfy a $\sqrt{T}$ regret rate even in stochastic environments, we generally cannot hope to recover logarithmic pseudo-regret by only selecting among them. For logarithmic pseudo-regret, we need to ensure that the given policy $\widehat{\pi}$ with gap estimate $\widehat{\Delta}$ is played sufficiently often. However, we also need to monitor its regret against all other policies in case the environments turns out to be adversarial. If at any point, $\widehat{\pi}$ fails to maintain a gap of order $\widehat{\Delta}$, we can conclude that the environment is adversarial. Then Arbe-GapExploit returns and we simply play an instance of Arbe (Line 16 of Algorithm 2).

We will present a brief summary of the main intuition behind our exploitation phase approach here and defer a longer discussion and the detailed pseudo-code to Appendix D. In each round, we play policy $\widehat{\pi}$ with probability approximately $1 - \frac{\text{poly}(R(\Pi_M))}{\widehat{\Delta}^2 t}$, and with the remaining probability $\frac{\text{poly}(R(\Pi_M))}{\widehat{\Delta}^2 t}$ a version of base learner $\mathbb{A}_M$ with policy class $\Pi_M \setminus \{\widehat{\pi}\}$.

If the environment is indeed stochastic, then $\widehat{\pi} = \pi^\star$ does not incur any pseudo-regret and the total regret in other $t' \approx t \cdot \frac{\text{poly}(R(\Pi_M))}{\widehat{\Delta}^2 t} \approx \frac{\text{poly}(R(\Pi_M))}{\widehat{\Delta}^2}$ rounds can be bounded as $t' \cdot \Delta + \text{Reg}_{\mathbb{A}_M}(t', \Pi_M \setminus \{\widehat{\pi}\}) \approx \frac{\text{poly}(R(\Pi_M))}{\Delta} \ln(t)$. The first term is the regret of the best policy in $\Pi_M \setminus \{\widehat{\pi}\}$ and the second term is the regret of $\mathbb{A}_M$ against that policy. Since $\mathbb{A}_M$ is h-stable on $\Pi \setminus \{\widehat{\pi}\}$, its regret against that policy is at most $\text{poly}(R(\Pi_M))\sqrt{t'}$ and we get the desired pseudo-regret.

To ensure good regret in the adversarial case, we need to detect quickly enough when $\widehat{\pi}$ does not exhibit a performance gap anymore and fall back to a fully adversarial algorithm. Similar to $\widehat{\Delta}_t$ in the first phase, we use a lower confidence bound on the average performance gap and continuously test whether it falls below $\frac{\widehat{\Delta}}{2}$. This simple approach would give $\sqrt{t}$ adversarial regret but may exhibit a $\text{poly}(R(\Pi_M))$ dependency. Fortunately, we can avoid this dependency and retain the desired model selection regret rates by extending the intuition above to also test an upper bound on the gap. For details, see Appendix D.

6 Conclusions

We have described and analyzed a novel model selection scheme for bandit algorithms that benefits from best-of-both-worlds high probability regret guarantees. Though not restricted to linear bandits, our machinery can be specifically applied to adversarial/stochastic linear bandit tasks, where model selection is performed on the unknown dimensionality of the linear reward function. This has required extending the regret balancing technique of model selection from stochastic to adversarial rewards and a very careful handling of the associated mis-specification tests. The base learners aggregated by our meta-algorithm have to satisfy an anytime high probability regret guarantee in the adversarial case, along with regret stability and action space extendability properties which we have shown are satisfied by (variants of) known algorithm in the bandits literature.

Our best-of-both world model selection regret guarantees cannot in general be improved, specifically in stochastic environments with gaps, where it is generally impossible to obtain $\log t$-like model selection bounds that only depend on the complexity of $\Pi_{i_\star}$. On the other hand, it would be nice to see in Theorem 5 a better polynomial dependence on M and $R(\Pi_M)$. Also, it might be possible to adapt Arbe-Gap to work with adversarially corrupted stochastic rewards while doing model selection on the complexity of the models. We leave this as a future research direction.

[7] Due to linking learners, there is slight ambiguity in defining the selected policy per round. We here determine the selected policy as the policy chosen by the learner that eventually picked a_t *after resolving linked actions*.

References

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

[2] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. *arXiv preprint arXiv:1612.06246*, 2016.

[3] 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.

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

[5] P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, 2016.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2003.

[7] P. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342. Omnipress, 2008.

[8] 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.

[9] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[10] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, 2012.

[11] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

[12] 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.

[13] A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit. Dynamic balancing for model selection in bandits and rl. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2276–2285. PMLR, 2021.

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

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

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

[17] A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, 2019.

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

[19] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[20] T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

[21] C. Lee, H. Luo, W. Chen-Yu, M. Zhang, and X. Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 6142–6151. PMLR, 2021.

[22] J. Lee, A. Pacchiano, V. Muthukumar, W. Kong, and E. Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2021.

[23] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

[24] T. Marinov and J. Zimmert. The pareto frontier of model selection for general contextual bandits. In *Neurips*, 2021.

[25] G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015.

[26] A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

[27] 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.

[28] Y. Seldin and G. Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, 2017.

[29] Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, 2014.

[30] C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, 2018.

[31] C.-Y. Wei, C. Dann, and J. Zimmert. A model selection approach for corruption robust reinforcement learning. *arXiv preprint arXiv:2110.03580*, 2021.

[32] Y. Zhu and R. Nowak. Pareto optimal model selection in linear bandits, 2021.

[33] J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[34] J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res*, pages 22:28–1, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is of purely theoretical nature, and we are not aware of any potential negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See in particular the definitions and desiderata in Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are provided in the appendix.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

1 Introduction **1**

2 Problem Setting and Assumptions **3**

2.1 Model selection and best-of-both-worlds regret 5

3 Summary of Our Contributions and Discussion **5**

4 Adversarial Model Selection Using Regret Balancing **6**

5 Adversarial Model Selection with Best of Both Worlds Guarantees **8**

5.1 First Phase: Candidate Policy Identification and Gap Estimation 8

5.2 Second Phase: Exploitation 10

6 Conclusions **10**

A Lower Bound on Model Selection for Stochastic Environments **15**

B Examples of h-Stability and Extendability **16**

B.1 Geometric Hedge for Adversarial Linear Bandits 16

B.1.1 h-Stability 17

B.1.2 Extendability 18

B.1.3 Removing an Individual Policy For Best of Both Worlds Regret 18

B.1.4 Proofs 18

B.2 Exp4 algorithm for finite policy classes 27

C Adversarial Regret Balancing and Elimination **28**

C.1 Proof of Lemma 21 28

C.2 Concentration Bounds on Reward Sequences 30

D Adversarial Regret Balancing and Elimination with Best of Both Worlds Regret **31**

D.1 Description of the Second Phase: Gap Exploitation 35

D.2 Analysis of the First Phase 38

D.2.1 Adversarial Guarantees 38

D.2.2 Stochastic Guarantees 41

D.3 Analysis of Arbe-GapExploit 46

D.3.1 Guarantees for stochastic environments 46

D.3.2 Exploitation Subroutine Guarantees for Adversarial Environments 49

D.4 Concentration Bounds in Stochastic Environments 53

E Additional Technical Lemmas **55**

A Lower Bound on Model Selection for Stochastic Environments

Considering a simple multi-armed bandit problem suffices to prove that in stochastic environments with gap Δ it is generally impossible to obtain model selection guarantees of the form

$$\frac{R^2(\Pi_{i_\star})\log T}{\Delta} .$$

Theorem 6. *Let $K_1, K_2 \in \mathbb{N}$ with $K_1 < K_2$ and let $\Delta \in \mathbb{R}^+$ be fixed. Further, let $c : \mathbb{N} \to \mathbb{R}$ be an arbitrary function over the real numbers. If $c(K_1) < K_2 - K_1$, there is a class of K_2 multi-armed bandit problems with one optimal arm and gaps in $[\Delta, 2\Delta]$ and a $T_0 \in \mathbb{N}$ such that the following holds. For any algorithm $\mathbb{A}$ and for all $T \geq T_0$, there is a problem instance such that*

$$\mathbb{E}[\mathrm{Reg}_{\mathbb{A}}(T)] > \begin{cases} \frac{c(K_1)}{\Delta}\ln T & \textit{if the optimal arm } a^\star \in [K_1] \\ \frac{c(K_2)}{\Delta}\ln T & \textit{otherwise} . \end{cases}$$

Remark 7. *The proof below actually shows the following stronger version*

$$\mathbb{E}[\mathrm{Reg}_{\mathbb{A}}(T)] > \begin{cases} \frac{c(K_1)}{\Delta}\ln T & \textit{if the optimal arm } a^\star \in [K_1] \\ \frac{\Delta}{16}T^{1-\frac{c(K_1)}{K_2-K_1}} & \textit{otherwise} . \end{cases}$$

This shows that if we aim to obtain an instance-dependent logarithmic regret bound that only scales with K_1 when $a^\star \in [K_1]$, then we cannot recover sublinear regret for $K_2 \gg K_1$ when $a^\star \notin [K_1]$.

Proof. We will show the statement for $K_1 = 1$ but it can be trivially generalized to $K_1 > 1$. The rewards of all arms and in all instances are drawn from a Gaussian distribution with variance 1 but different means. We denote by μ_i^j the mean reward of arm j in instance i. We identify each bandit instance in the family by its mean rewards μ_i. They are given by

$$\mu_i = [\Delta, 0, \ldots, 0] \qquad \text{for } i = 1 \tag{6}$$

$$\mu_i = [\Delta, 0, \ldots, \underbrace{2\Delta}_{i\text{'th pos}}, 0, \ldots, 0] \qquad \text{for } i > 1 \tag{7}$$

Thus, arm i is optimal in instance i. Consider any algorithm $\mathbb{A}$ and assume it violates the lower bound for the case where $a^\star \in [K_1]$, i.e., in the first problem instance,

$$\mathbb{E}\left[\mathrm{Reg}_{\mathbb{A}}(T, \mu_1)\right] \leq \frac{c(K_1)\log(T)}{\Delta}$$

holds for some $T \geq T_0$ (we will specify T_0 later). Otherwise, the statement is already true. We will now show that this algorithm has to satisfy the regret lower bound in one of the other problem instances for T. By definition of μ_1, we can write the expected regret as

$$\mathbb{E}\,\mathrm{Reg}_{\mathbb{A}}(T, \mu_1) \geq 2\Delta \sum_{i=K_1+1}^{K_2} \mathbb{E}_1\left[T_i(T)\right] ,$$

where $T_i(T)$ is the (random) number of times $\mathbb{A}$ has pulled arm i up to time T. $\mathbb{E}_j$ and $\mathbb{P}_j$ denote the expectation and the probability distribution induced by algorithm $\mathbb{A}$ in instance j, respectively. Consider now a problem instance $\widehat{i} \in \mathrm{argmin}_{K_1+1 \leq i \leq K_2} \mathbb{E}_1\left[T_i(T)\right]$. As a consequence of the previous two observations,

$$2\Delta(K_2 - K_1)\mathbb{E}_1\left[T_{\widehat{i}}(T)\right] \leq \frac{c(K_1)\log(T)}{\Delta} ,$$

hence

$$\mathbb{E}_1\left[T_{\widehat{i}}(T)\right] \leq \frac{c(K_1)\log(T)}{2(K_2 - K_1)\Delta^2} .$$

By the divergence decomposition [20, Lemma 15.1],

$$\mathrm{KL}\left(\mathbb{P}_1, \mathbb{P}_{\widehat{i}}\right) = \mathbb{E}_1\left[T_{\widehat{i}}(T)\right]\frac{(2\Delta)^2}{2} \leq \frac{c(K_1)\log(T)}{K_2 - K_1} .$$

Define the event $\mathcal{E} = \{T_1(T) \leq \frac{T}{2}\}$. Notice that,

$$\frac{T}{2}\Delta\mathbb{P}_1(\mathcal{E}) \leq \mathbb{E}_1 \operatorname{Reg}_{\mathbb{A}}(T, \mu_1) \leq \frac{c(K_1)\log(T)}{\Delta}$$

which implies,

$$\mathbb{P}_1(\mathcal{E}) \leq \frac{2c(K_1)\log(T)}{\Delta^2 T} . \tag{8}$$

By the Bretagnolle-Huber inequality,

$$\mathbb{P}_1(\mathcal{E}) + \mathbb{P}_{\widehat{i}}(\mathcal{E}^c) \geq \frac{1}{2}\exp\left(-\frac{c(K_1)\log(T)}{K_2 - K_1}\right) = \frac{1}{2}\left(\frac{1}{T}\right)^{\frac{c(K_1)}{K_2-K_1}} ,$$

and by combining the last two inequalities, we can lower bound $\mathbb{P}_{\widehat{i}}(\mathcal{E}^c)$ as

$$\mathbb{P}_{\widehat{i}}(\mathcal{E}^c) \geq \frac{1}{2}\left(\frac{1}{T}\right)^{\frac{c(K_1)}{K_2-K_1}} - \frac{2c(K_1)\log(T)}{\Delta^2 T} .$$

Since $\mathbb{E}_{\widehat{i}} \operatorname{Reg}_{\mathbb{A}}(T, \mu_{\widehat{i}}) \geq \frac{T\Delta}{2}\mathbb{P}_{\widehat{i}}(\mathcal{E}^c)$ this implies,

$$\mathbb{E}_{\widehat{i}} \operatorname{Reg}_{\mathbb{A}}(T, \mu_{\widehat{i}}) \geq \frac{T\Delta}{2}\left(\frac{1}{2}\left(\frac{1}{T}\right)^{\frac{c(K_1)}{K_2-K_1}} - \frac{2c(K_1)\log(T)}{\Delta T}\right) = \frac{\Delta}{4}T^{1-\frac{c(K_1)}{K_2-K_1}} - c(K_1)\log(T) .$$

By setting T_0 sufficiently large as a function of Δ, K_1, K_2 and c, we can conclude that

$$\mathbb{E}_{\widehat{i}} \operatorname{Reg}_{\mathbb{A}}(T, \mu_{\widehat{i}}) \geq \frac{\Delta}{4}T^{1-\frac{c(K_1)}{K_2-K_1}} - c(K_1)\log(T) \geq \frac{\Delta}{8}T^{1-\frac{c(K_1)}{K_2-K_1}} > \frac{c(K_2)\log(T)}{\Delta} .$$

Specifically, it suffices to set T_0 as the smallest value of T that satisfies the last two inequalities in the display. This shows that the lower bound holds at time T for problem instance $\widehat{i}$. □

B Examples of h-Stability and Extendability

This appendix shows examples of h-stability and extendabilty. The first example, contained in Section B.1, is a variant of the Geometric Hedge algorithm from [7]. We sketch a second example in Section B.2, where we deal with a high probability variant of the Exp4 algorithm from [6].

B.1 Geometric Hedge for Adversarial Linear Bandits

Let us start by introducing a weighted version of the Geometric Hedge Algorithm that satisfies the h-stability condition of Definition 2. We consider the setting where the action $\mathcal{A} \subset \mathbb{R}^d$ is finite (but potentially large). We denote by $\mathbf{a}_t$ and $\boldsymbol{\omega}_t$ the learner's action selection and the adversary's reward vector at time t, respectively. The associated reward $r_t = \mathbf{a}_t^\top \boldsymbol{\omega}_t$ is assumed to lie in the interval $[-1, 1]$. Moreover, let $\delta \in (0, 1)$ be a probability parameter and set for brevity $\delta' = \frac{\delta}{|\mathcal{A}|}$.

The algorithm, an anytime variant of the Geometric Hedge algorithm from [7] with John's ellipsoid exploration (e.g., [9]) is detailed in Algorithm 3. We call the algorithm Anytime Weighted Geometric Hedge. The algorithm takes in input the set of actions $\mathcal{A}$, the failure probability δ, and a weighting probability $\rho \in (0, 1]$ which will play the role of an importance weight. As is standard, the algorithm maintains over time a distribution p_t over $\mathcal{A}$, which is itself a mixture of an exponential weight distribution q_t and an exploration distribution p_E. The exploration distribution p_E is defined beforehand to be the John's ellipsoid distribution associated with $\mathcal{A}$ (see [9] for details). As in Geometric Hedge [7], the algorithm builds a covariance matrix $\boldsymbol{\Sigma}_t$ by computing the expectation of $\mathbf{a}\mathbf{a}^\top$ where $\mathbf{a}$ is drawn according to the current distribution p_t over actions.[8] Then, the algorithm samples a Bernoulli random variable b_t, and computes an (importance-weighted) unbiased estimator $\widehat{\boldsymbol{\omega}}_t$ of $\boldsymbol{\omega}_t$, which is plugged into the (biased) reward estimator $\widetilde{r}_t(\mathbf{a})$ that the algorithm associates with every action $\mathbf{a} \in \mathcal{A}$. The factors $\widetilde{r}_t(\mathbf{a})$ are those that determine the exponential update of distribution p_t.

[8] In this section, we denote by $\mathcal{F}_{t-1}$ the σ-algebra generated by all past random variables up to, but excluding, the random draw of $\mathbf{a}_t$ (so that p_t is $\mathcal{F}_{t-1}$-measurable).

Algorithm 3: Anytime Weighted Geometric Hedge.

Input: Action set $\mathcal{A}$, failure probability δ, weighting probability $\rho \in (0, 1]$.
Initialize $w_1(\mathbf{a}) = 1$, $W_1 = |\mathcal{A}|$ and $q_1(\mathbf{a}) = \frac{1}{|\mathcal{A}|}$ for all $\mathbf{a} \in \mathcal{A}$
for $t = 1, 2, \cdots$ **do**
  Compute sampling distribution

$$p_t(\mathbf{a}) = (1 - \gamma_t) q_t(\mathbf{a}) + \gamma_t p_E(\mathbf{a}) \qquad \text{where } q_t(\mathbf{a}) = \frac{w_t(\mathbf{a})}{W_t} \tag{9}$$

  Adversary generates reward vector $\boldsymbol{\omega}_t$
  Sample action $\mathbf{a}_t \sim p_t$
  Observe and gather reward $r_t = \mathbf{a}_t^\top \boldsymbol{\omega}_t$
  Build covariance matrix $\boldsymbol{\Sigma}_t = \mathbb{E}_{\mathbf{a} \sim p_t} \left[\mathbf{a}\mathbf{a}^\top | \mathcal{F}_{t-1} \right]$
  Sample $b_t \sim \mathrm{Ber}(\rho)$
  Compute unbiased reward vector estimator $\widehat{\boldsymbol{\omega}}_t = b_t \frac{r_t \boldsymbol{\Sigma}_t^{-1} \mathbf{a}_t}{\rho}$
  Compute the reward upper bounds

$$\widetilde{r}_t(\mathbf{a}) = \mathbf{a}^\top \widehat{\boldsymbol{\omega}}_t + 2\mathbf{a}^\top \boldsymbol{\Sigma}_t^{-1} \mathbf{a} \sqrt{\frac{\ln(12t^2/\delta')}{\rho\, d\, t}} \quad \forall \mathbf{a} \in \mathcal{A}$$

  Update distribution

$$w_{t+1}(\mathbf{a}) = \exp\left(\eta_{t+1} \sum_{\ell=1}^{t} \widetilde{r}_\ell(\mathbf{a}) \right) \quad \forall \mathbf{a} \in \mathcal{A} \tag{10}$$

  Update normalization factor $W_{t+1} = \sum_{\mathbf{a} \in \mathcal{A}} w_{t+1}(\mathbf{a})$

Remark 8. *For simplicity, the algorithm is formulated for the case where the action space $\mathcal{A}$ is finite. When $\mathcal{A}$ is infinite, we can still formulate an algorithm that applies to a ϵ-cover of $\mathcal{A}$ that is restarted at exponentially increasing time-steps $t_0 = 1, 2, 4, 8, \ldots,$. At the beginning of each epoch, the covering level ϵ is set to $\mathcal{O}(1/t_0)$ so as to obtain an anytime algorithm. A very similar bound to the one in Theorem 10 is obtained, where $|\mathcal{A}|$ therein is replaced by $|\mathcal{A}_{1/t}|$, $\mathcal{A}_{1/t}$ being a $1/t$-cover of $\mathcal{A}$ w.r.t. the infinity norm.*

Remark 9. *Notice that we are using the fact that $\boldsymbol{\Sigma}_t$ is invertible for all t. This is no loss of generality, as we can always assume that $\mathcal{A}$ spans the whole d-dimensional space (if this is not the case, we can project each $\mathbf{a}$ onto the space spanned by $\mathcal{A}$ and reduce to this case). Combined with the fact that, for all t, distribution p_t assigns a nonzero probability to each action, this implies that the expectation $\boldsymbol{\Sigma}_t = \mathbb{E}_{\mathbf{a} \sim p_t} \left[\mathbf{a}\mathbf{a}^\top | \mathcal{F}_{t-1} \right]$ must be full rank.*

B.1.1 h-Stability

Setting the mixture factor γ_t and the learning rate η_t appropriately, we can prove the following regret guarantee for Algorithm 3.

Theorem 10. *Let the Anytime Weighted Geometric Hedge Algorithm be run with*

$$\eta_t = \mathcal{O}\left(\frac{\rho \gamma_t}{d + \sqrt{\frac{d}{t}} \sqrt{\rho \ln |\mathcal{A}| \ln \frac{t}{\delta}}} \right), \qquad \gamma_t = \min\left\{ \sqrt{\frac{d \ln |\mathcal{A}| \ln \frac{t}{\delta}}{\rho\, t}}, \frac{1}{2} \right\},$$

and $\rho \in (0, 1]$. Then with probability at least $1 - \delta(3 + 2d^2)$, simultaneously for all t, the regret $\mathrm{Reg}(t)$ after t rounds can be bounded as

$$\mathrm{Reg}(t) = \mathcal{O}\left(\sqrt{\frac{d\, t\, \ln |\mathcal{A}|}{\rho} \ln \frac{t}{\delta}} + \rho \ln |\mathcal{A}| \ln \frac{t}{\delta} \ln t + \left(\frac{\ln |\mathcal{A}|}{d} \ln \frac{t}{\delta} \right)^{1/4} t^{1/4} \left(\frac{1}{\rho} \right)^{3/4} + \frac{1}{\rho} \ln \frac{t}{\delta} \right).$$

In the above, the big-oh notation only hides absolute constants.

Hence we have the following corollary.

Corollary 11. *Let the complexity $R(\Pi)$ of the policy space Π be defined as $R(\Pi) = \sqrt{d \log |\mathcal{A}|}$. Then with the same assumptions and setting as in Theorem 10, the Anytime Weighted Geometric Hedge algorithm is h-stable in that, for $t \to \infty$ and constant ρ independent of t, its regret $\mathrm{Reg}(t)$ satisfies*

$$\mathrm{Reg}(t) = \mathcal{O}\left(R(\Pi)\sqrt{\frac{t}{\rho}\ln\frac{t}{\delta}}\right) ,$$

with probability at least $1-\delta$, where the big-oh hides terms in t which are lower order than $\sqrt{t \log t}$ as $t \to \infty$.

B.1.2 Extendability

The extendability of the Anytime Weighted Geometric Hedge algorithm is easily obtained by simply observing that, since $\mathcal{A} \subseteq \mathbb{R}^d$, we can extend the dimensionality d to $d+k$. Then we extend each $\mathbf{a} = (a_1, \ldots, a_d) \in \mathcal{A}$ to $\mathbf{a}' = (a_1, \ldots, a_d, \underbrace{0, \ldots, 0}_{k \text{ zeros}})$, and add k new actions $\mathbf{b}_i$ of the form

$$\mathbf{b}_i = (\underbrace{0, \ldots, 0}_{d \text{ zeros}}, \underbrace{0, \ldots, 1, \ldots, 0}_{\text{position } i})$$

for $i = 1, \ldots, k$. Denote the extended policy space by $\mathcal{A}'$. Consistent with this extension, any policy π in the original policy space Π has to be interpreted as a probability distribution in $\Delta_{\mathcal{A}'}$, whose last k components are zero, while the k extra indicator policies $\mathbf{1}\{b_1\}, \ldots, \mathbf{1}\{b_k\}$ are degenerate probability distributions in $\Delta_{\mathcal{A}'}$, where $\mathbf{1}\{b_i\}$ places all its probability mass on the $(d+i)$-th component. Finally the adversary reward vector $\boldsymbol{\omega}_t \in \mathbb{R}^d$ turns into the $(d+k)$-dimensional vector $\boldsymbol{\omega}'_t \in \mathbb{R}^{d+k}$, where the first d components of the two vectors are the same, and the $(d+i)$-th component of $\boldsymbol{\omega}'_t$ is simply the regret generated for the i-th extra action $\mathbf{b}_i$.

B.1.3 Removing an Individual Policy For Best of Both Worlds Regret

For the best of both worlds regret in Section 5 we adopt the following view. Here, each policy corresponds to a single action $a \in \mathcal{A}$, i.e., the policy space of the learner is $\Pi = \{\mathbf{1}\{a\} : a \in \mathcal{A}\}$. In this case, the random draw of $\mathbf{a}_t \sim p_t$ is interpreted as a random choice of *policy*. Note that this view does not impact the regret guarantee and is consistent with approach for extendability above. This view is necessary for a positive gap to be possible and removing a certain policy can be easily implemented by removing the corresponding action.

B.1.4 Proofs

We work under the assumption all the rewards r_t have values between -1 and 1:

Assumption B.1 (Boundedness)**.** *The true rewards r_t are bounded in that $\forall \mathbf{a} \in \mathcal{A}$ and $\forall t \in \mathbb{N}$ we have $|\mathbf{a}^\top \boldsymbol{\omega}_t| \leq 1$.*

For all $\mathbf{a} \in \mathcal{A}$, define

$$\widehat{r}_t(\mathbf{a}) = \mathbf{a}^\top \widehat{\boldsymbol{\omega}}_t \quad \text{where} \quad \widehat{\boldsymbol{\omega}}_t = b_t \frac{r_t \left(\boldsymbol{\Sigma}_t\right)^{-1} \mathbf{a}_t}{\rho} .$$

We will use the notation $\mathbb{E}[\cdot|\mathcal{F}_t]$ to denote the conditional expectation where the sigma algebra $\mathcal{F}_t$ is generated by the random variables $(\boldsymbol{\omega}_1, b_1, \mathbf{a}_1, \cdots, \boldsymbol{\omega}_{t-1}, b_{t-1}, \mathbf{a}_{t-1}, \boldsymbol{\omega}_t, b_t, \mathbf{a}_t)$. Let $\mathcal{F}_t^-$ be the sigma algebra generated by $(\boldsymbol{\omega}_1, b_1, \mathbf{a}_1, \cdots, \boldsymbol{\omega}_{t-1}, b_{t-1}, \mathbf{a}_{t-1}, \boldsymbol{\omega}_t)$. Observe that $\widehat{\boldsymbol{\omega}}_t$, $\widehat{r}_t(\cdot)$ and $\widetilde{r}_t(\cdot)$ are $\mathcal{F}_t$ measurable, $\boldsymbol{\Sigma}_t$ is $\mathcal{F}_{t-1}$ measurable, $\mathbb{E}\left[\widehat{\boldsymbol{\omega}}_t|\mathcal{F}_t^-\right] = \boldsymbol{\omega}_t$ and $\mathbb{E}\left[b_t|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[b_t|\mathcal{F}_t^-\right] = \rho$. When considering $\mathbb{E}\left[\cdot|\mathcal{F}_t^-\right]$, the expectation is over $\mathbf{a}_t, b_t$ holding $\boldsymbol{\omega}_t$ fixed. Every time we consider an expectation of the form $\mathbb{E}_{\mathbf{a}\sim p_t}\left[\widehat{r}_t(\mathbf{a})|\mathcal{F}_t^-\right]$, $\mathbb{E}_{\mathbf{a}\sim p_t}\left[\widetilde{r}_t(\mathbf{a})|\mathcal{F}_t^-\right]$, $\mathbb{E}_{\mathbf{a}\sim p_t}\left[\widehat{r}_t(\mathbf{a})|\mathcal{F}_t\right]$ or $\mathbb{E}_{\mathbf{a}\sim p_t}\left[\widetilde{r}_t(\mathbf{a})|\mathcal{F}_t\right]$ the random variable $\mathbf{a}$ is a sample from p_t conditionally independent from $\mathbf{a}_t$ given $\mathcal{F}_t$ or $\mathcal{F}_t^-$. Moreover, for notational simplicity, whenever possible, we will omit the absolute multiplicative constants, and instead resort to a big-oh notation.

Lemma 12. *Let*

$$\sup_{\mathbf{a},\mathbf{b}\in\mathcal{A}} \mathbf{a}^\top \boldsymbol{\Sigma}_t^{-1}\mathbf{b} \le \frac{c(d)}{\gamma_t}\ , \tag{11}$$

for some function $c(\cdot)$ whose value will be detailed later on. Then for any fixed $\mathbf{a}\in\mathcal{A}$ and $t\in\mathbb{N}$ the following holds:

1. $|\widehat{r}_t(\mathbf{a})| \le \frac{c(d)}{\rho\gamma_t}$;
2. $\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a} \le \frac{c(d)}{\gamma_t}$;
3. $\mathbb{E}_{\widetilde{\mathbf{a}}\sim p_t}\left[\widetilde{\mathbf{a}}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\widetilde{\mathbf{a}}\middle|\mathcal{F}_{t-1}\right] = d$ *and* $\mathbb{E}_{\widetilde{\mathbf{a}}\sim p_t}\left[\widetilde{\mathbf{a}}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\widetilde{\mathbf{a}}\middle|\mathcal{F}_t^-\right] = d$;
4. $\mathbb{E}\left[\widehat{r}_t^2(\mathbf{a})|\mathcal{F}_t^-\right] \le \frac{\mathbf{a}^\top(\boldsymbol{\Sigma}_t)^{-1}\mathbf{a}}{\rho}$.

Proof. Item 1 simply follows by recalling that $\widehat{r}_t(\mathbf{a}) = b_t\frac{r_t\,\mathbf{a}^\top(\boldsymbol{\Sigma}_t)^{-1}\mathbf{a}_t}{\rho}$ with $|r_t|\le 1$. The condition from Equation 11 then implies the result. Item 2 follows from the same condition. Item 3 follows by observing that

$$\mathbb{E}_{\widetilde{\mathbf{a}}\sim p_t}[\widetilde{\mathbf{a}}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\widetilde{\mathbf{a}}|\mathcal{F}_{t-1}] = \mathbb{E}_{\widetilde{\mathbf{a}}\sim p_t}[\mathrm{tr}(\widetilde{\mathbf{a}}\widetilde{\mathbf{a}}^\top\boldsymbol{\Sigma}_t^{-1})|\mathcal{F}_{t-1}] = \mathrm{tr}\left(\mathbb{E}_{\widetilde{\mathbf{a}}\sim p_t}[\widetilde{\mathbf{a}}\widetilde{\mathbf{a}}^\top]\boldsymbol{\Sigma}_t^{-1}\right) = d\ .$$

Item 4 follows from the definition of $\widehat{r}_t(\mathbf{a})$. In fact, for any fixed $\mathbf{a}$, we can write

$$\begin{aligned}
\mathbb{E}\left[\widehat{r}_t^2(\mathbf{a})\,|\,\mathcal{F}_t^-\right] &= \frac{1}{\rho^2}\mathbb{E}_{b_t\sim\mathrm{Ber}(\rho),\mathbf{a}_t\sim p_t}[b_t^2 r_t^2\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}_t\mathbf{a}_t^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}\,|\,\mathcal{F}_t^-] \\
&= \frac{1}{\rho}\mathbb{E}_{\mathbf{a}_t\sim p_t}[r_t^2\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}_t\mathbf{a}_t^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}\,|\,\mathcal{F}_t^-] \\
&\overset{(i)}{\le} \frac{1}{\rho}\mathbb{E}_{\mathbf{a}_t\sim p_t}[\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}_t\mathbf{a}_t^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}\,|\,\mathcal{F}_t^-] \\
&= \frac{1}{\rho}\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbb{E}_t\left[\mathbf{a}_t\mathbf{a}_t^\top\right]\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a} \\
&= \frac{1}{\rho}\mathbf{a}^\top\left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{a}\ ,
\end{aligned}$$

where (i) holds because $|r_t|\le 1$. □

This allows us to prove the following version of Lemma 5 in [7],

Lemma 13. *Let $\{\alpha_\ell\}_{\ell=1}^\infty$ be a sequence of deterministic nonnegative weights satisfying $\alpha_\ell\le 1$ for all $\ell\in\mathbb{N}$. Let $\delta' = \frac{\delta}{|\mathcal{A}|}$. Then with probability at least $1-\delta$, simultaneously for all $\mathbf{a}\in\mathcal{A}$ and all $t\in\mathbb{N}$,*

$$\sum_{\ell=1}^t \alpha_\ell\widetilde{r}_\ell(\mathbf{a}) \ge \sum_{\ell=1}^t \alpha_\ell\,\mathbf{a}^\top\boldsymbol{\omega}_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho}\ln\frac{t}{\delta'}} + B_t\ln\frac{t}{\delta'}\right)\ , \tag{12}$$

where $B_t = \max_{\ell\le t}\frac{c(d)\alpha_\ell}{\rho\gamma_\ell} + \alpha_\ell$.

Proof. Fix $\mathbf{a}\in\mathcal{A}$, and recall the definition of $\widetilde{r}_\ell(\mathbf{a})$ in Algorithm 3. Define $M_t(\mathbf{a}) = \alpha_t\,\mathbf{a}^\top\boldsymbol{\omega}_t - \alpha_t\widehat{r}_t(\mathbf{a})$, and notice that $\{M_t(\mathbf{a})\}_{t=1,2,\ldots}$ is a martingale difference sequence. Using Lemma 12 (Item 1), along with Assumption B.1, we see that

$$|M_t(\mathbf{a})| \le \frac{c(d)\alpha_t}{\rho\gamma_t} + \alpha_t\ .$$

Let $V_t(\mathbf{a}) = \sum_{\ell=1}^t \mathrm{Var}[M_\ell(\mathbf{a})\,|\,\mathcal{F}_\ell^-]$ be the sum of conditional variances of variables $M_\ell(\mathbf{a})$. Using Lemma 50 we see that with probability at least $1-\delta'$, simultaneously for all t,

$$\sum_{\ell=1}^t \alpha_\ell\widehat{r}_\ell(\mathbf{a}) \ge \sum_{\ell=1}^t \alpha_\ell\,\mathbf{a}^\top\boldsymbol{\omega}_\ell - \mathcal{O}\left(\sqrt{V_t\ln\frac{t}{\delta'}} + B_t\ln\frac{t}{\delta'}\right)\ . \tag{13}$$

Since $\mathrm{Var}[M_t(\mathbf{a}) \,|\, \mathcal{F}_\ell^-] \le \mathbb{E}[M_t^2(\mathbf{a}) \,|\, \mathcal{F}_\ell^-] \le \alpha_t^2 \mathbb{E}_t\left[\widehat{r}_t^2(\mathbf{a})\right]$, by Lemma 12 (Item 4) we can write

$$\begin{aligned}
\sqrt{V_t(\mathbf{a})} &\le \sqrt{\sum_{\ell=1}^t \frac{\alpha_\ell^2 \mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\rho}} \\
&\le \sqrt{\left(\frac{1}{\sqrt{dt}} \sum_{\ell=1}^t \frac{\alpha_\ell^2 \mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho}}\right) \sqrt{\frac{dt}{\rho}}} \\
&\le \frac{1}{2}\left(\frac{1}{\sqrt{dt}} \sum_{\ell=1}^t \frac{\alpha_\ell^2 \mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho}} + \sqrt{\frac{dt}{\rho}}\right) ,
\end{aligned}$$

the last inequality being the arithmetic-geometric inequality $\sqrt{ab} \le \frac{1}{2}(a+b)$. Substituting back into Eq. (13) gives

$$\sum_{\ell=1}^t \alpha_\ell \widehat{r}_\ell(\mathbf{a}) \ge \sum_{\ell=1}^t \alpha_\ell \, \mathbf{a}^\top \boldsymbol{\omega}_\ell - \mathcal{O}\left(\left(\frac{\sum_{\ell=1}^t \alpha_\ell^2 \mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d t}} + \sqrt{\frac{dt}{\rho}}\right) \sqrt{\ln \frac{t}{\delta'}} + B_t \ln \frac{t}{\delta'}\right)$$

with probability at least $1-\delta'$ for all $t \in \mathbb{N}$. Since the function $g(t) = \frac{\ln \frac{t}{\delta'}}{t}$ is decreasing for all $t \ge 1$ we conclude that $\frac{\ln \frac{\ell}{\delta'}}{d\ell}$ is a decreasing function of ℓ. Using this last fact together with the condition $\alpha_\ell \le 1$ we see that

$$\alpha_\ell \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d \ell}} \sqrt{\ln \frac{\ell}{\delta'}} \ge \alpha_\ell \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}} \ge \alpha_\ell^2 \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}} ,$$

and therefore

$$\begin{aligned}
\sum_{\ell=1}^t \alpha_\ell \widehat{r}_\ell(\mathbf{a}) + \alpha_\ell \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d \ell}} \sqrt{\ln \frac{\ell}{\delta'}} &\ge \sum_{\ell=1}^t \alpha_\ell \widehat{r}_\ell(\mathbf{a}) + \alpha_\ell \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_\ell\right)^{-1} \mathbf{a}}{\sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}} \\
&\ge \sum_{\ell=1}^t \alpha_\ell \mathbf{a}^\top \boldsymbol{\omega}_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + B_t \ln \frac{t}{\delta'}\right)
\end{aligned}$$

with probability at least $1-\delta'$ for all $t \in \mathbb{N}$. The result follows by taking a union bound over all $\mathbf{a} \in \mathcal{A}$. □

In particular when all weights $\alpha_\ell = 1$ Lemma 13 implies the following.

Corollary 14. *With the same notation as in Lemma 13, with probability at least $1-\delta$ simultaneously for all $\mathbf{a} \in \mathcal{A}$ and all $t \in \mathcal{N}$,*

$$\sum_{\ell=1}^t \widetilde{r}_\ell(\mathbf{a}) \ge \sum_{\ell=1}^t \mathbf{a}^\top \boldsymbol{\omega}_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + B_t \ln \frac{t}{\delta'}\right) ,$$

where $B_t = \max_{\ell \le t} \frac{c(d)}{\rho \gamma_\ell} + 1$.

We now proceed to upper bound $|\widetilde{r}_t(\mathbf{a})|$. This will inform our choice for learning rate η_t.

Lemma 15. *Let $\delta' = \frac{\delta}{|\mathcal{A}|}$. For all $\mathbf{a} \in \mathcal{A}$, $|\widetilde{r}_t(\mathbf{a})| = \mathcal{O}\left(\frac{c(d)}{\rho \gamma_t} + \left(\frac{c(d)}{\gamma_t \sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}}\right)\right)$.*

Proof. For each $\mathbf{a} \in \mathcal{A}$, we can write

$$\begin{aligned}
|\widetilde{r}_t(\mathbf{a})| &= \mathcal{O}\left(|\widehat{r}_t(\mathbf{a})| + \frac{\mathbf{a}^\top \left(\mathbf{\Sigma}_t\right)^{-1} \mathbf{a}}{\sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}}\right) \\
&= \mathcal{O}\left(\frac{c(d)}{\rho \gamma_t} + \left(\frac{c(d)}{\gamma_t \sqrt{\rho d t}} \sqrt{\ln \frac{t}{\delta'}}\right)\right) ,
\end{aligned}$$

the last inequality holding as a consequence of Lemma 12. □

For the analysis of exponential weights, we will insure that $|\eta_t \widetilde{r}_t(\mathbf{a})| \le 1$ for all $\mathbf{a}$ and t. This imposes the restriction of the following form

$$\eta_t = \mathcal{O}\left(\frac{\rho}{\frac{c(d)}{\gamma_t} + \left(\frac{c(d)}{\gamma_t\sqrt{dt}}\sqrt{\rho \ln \frac{t}{\delta'}}\right)}\right) = \mathcal{O}\left(\frac{\rho\gamma_t}{c(d) + \frac{c(d)}{\sqrt{dt}}\sqrt{\rho \ln \frac{t}{\delta'}}}\right) . \tag{14}$$

We are now ready to tackle the anytime high probability regret guarantees for Algorithm 3.

Lemma 16. *Let the condition in Eq. (14) hold with a nonincreasing sequence of learning rates η_t. Then for all $\bar{\mathbf{a}} \in \mathcal{A}$ and $t \in \mathbb{N}$*

$$\sum_{\ell=1}^{t} \widetilde{r}_\ell(\bar{\mathbf{a}}) \le 1 + \frac{\ln(\mathcal{A})}{\eta_t} + \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell}\left(\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\widetilde{r}_\ell(\mathbf{a}) + \eta_\ell\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 \middle| \mathcal{F}_\ell\right] - \gamma_\ell \mathbb{E}_{\mathbf{a}\sim p_{\mathrm{E}}(\mathbf{a})}\left[\widetilde{r}_\ell(\mathbf{a}) \middle| \mathcal{F}_\ell\right]\right) .$$

Proof. Recall that

$$w_\ell(\mathbf{a}) = \exp\left(\eta_\ell \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a})\right)$$

and $W_\ell = \sum_{\mathbf{a}\in\mathcal{A}} w_\ell(\mathbf{a})$. Let us also define

$$w_\ell^-(\mathbf{a}) = \exp\left(\eta_{\ell-1} \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a})\right)$$

and, $W_\ell^- = \sum_{\mathbf{a}\in\mathcal{A}} w_\ell^-(\mathbf{a})$. Moreover, set for brevity $A_\ell = \ln\left(\frac{W_{\ell+1}^-}{W_\ell}\right)$. We can write

$$\begin{aligned}
\exp(A_\ell) &= \frac{W_{\ell+1}^-}{W_\ell} \\
&= \frac{\sum_{\mathbf{a}\in\mathcal{A}} \exp\left(\eta_\ell \sum_{\ell'=1}^{\ell} \widetilde{r}_{\ell'}(\mathbf{a})\right)}{\sum_{\mathbf{a}\in\mathcal{A}} \exp\left(\eta_\ell \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a})\right)} \\
&= \sum_{\mathbf{a}\in\mathcal{A}} q_\ell(\mathbf{a}) \exp\left(\eta_\ell \widetilde{r}_\ell(\mathbf{a})\right) \\
&\le 1 + \sum_{\mathbf{a}\in\mathcal{A}} q_\ell(\mathbf{a})\eta_\ell \widetilde{r}_\ell(\mathbf{a}) + q_\ell(\mathbf{a})\eta_\ell^2\left(\widetilde{r}_\ell(\mathbf{a})\right)^2
\end{aligned}$$

the last inequality holding because $e^x \le 1 + x + x^2$ whenever $|x| \le 1$. Taking logs and using the fact that $\ln(1+x) \le x$ yields

$$\begin{aligned}
A_\ell &\le \eta_\ell \sum_{\mathbf{a}\in\mathcal{A}} q_\ell(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) + q_\ell(\mathbf{a})\eta_\ell\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 \\
&\overset{(i)}{\le} \eta_\ell \sum_{\mathbf{a}\in\mathcal{A}} q_\ell(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) + \frac{p_\ell(\mathbf{a})}{1-\gamma_\ell}\eta_\ell\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 \\
&\overset{(ii)}{=} \frac{\eta_\ell}{1-\gamma_\ell} \sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) + p_\ell(\mathbf{a})\eta_\ell\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 - \gamma_\ell p_{\mathrm{E}}(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) ,
\end{aligned}$$

where (i) follows because $q_\ell(\mathbf{a}) \le \frac{p_\ell(\mathbf{a})}{1-\gamma_\ell}$ and (ii) because $p_\ell = (1-\gamma_\ell)q_\ell(\mathbf{a}) + \gamma_\ell p_{\mathrm{E}}(\mathbf{a})$. Hence

$$\begin{aligned}
\sum_{\ell=1}^{t} \frac{A_\ell}{\eta_\ell} &= \sum_{\ell=1}^{t} \frac{1}{\eta_\ell} \ln\left(\frac{W_{\ell+1}^-}{W_\ell}\right) \\
&\le \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) + p_\ell(\mathbf{a})\eta_\ell\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 - \gamma_\ell p_{\mathrm{E}}(\mathbf{a})\widetilde{r}_\ell(\mathbf{a}) .
\end{aligned} \tag{15}$$

Now, define the following potential

$$\Phi_\ell(\eta) = \frac{1}{\eta} \ln \left(\frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}\in\mathcal{A}} \exp \left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) \right) \right) .$$

Notice that by de L'Hopital's rule, this implies $\lim_{\eta\to 0} \Phi_\ell(\eta) = \Phi_\ell(0) = 1$ for all ℓ. Let $\eta_0 = 0$. We have

$$\begin{aligned} 1 + \sum_{\ell=1}^{t} \frac{1}{\eta_\ell} \ln \left(\frac{W_{\ell+1}^-}{W_\ell} \right) &= \Phi_1(\eta_0) + \sum_{\ell=1}^{t} \Phi_{\ell+1}(\eta_\ell) - \Phi_\ell(\eta_\ell) \\ &= \left(\sum_{\ell=1}^{t} \left(\Phi_\ell(\eta_{\ell-1}) - \Phi_\ell(\eta_\ell) \right) \right) + \Phi_{t+1}(\eta_t) . \end{aligned}$$

Next, we now show that for all ℓ the function $\Phi_\ell(\eta)$ is an increasing function of η. To this effect, let $p_\ell^\eta(\mathbf{a}) = \frac{\exp\left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a})\right)}{\sum_{\mathbf{a}'\in\mathcal{A}} \exp\left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}')\right)}$. Observe that the following relationship holds,

$$\begin{aligned} \Phi'_\ell(\eta) &= \frac{-1}{\eta^2} \ln \left(\frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}\in\mathcal{A}} \exp \left(\eta \sum_{\ell'}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) \right) \right) + \frac{1}{\eta} \frac{\sum_{\mathbf{a}\in\mathcal{A}} \left[\sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) \right] \exp \left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) \right)}{\sum_{\mathbf{a}\in\mathcal{A}} \exp \left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) \right)} \\ &= \frac{1}{\eta^2} \sum_{\mathbf{a}\in\mathcal{A}} p_\ell^\eta(\mathbf{a}) \left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}) - \ln \left(\frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}'\in\mathcal{A}} \exp \left(\eta \sum_{\ell'=1}^{\ell-1} \widetilde{r}_{\ell'}(\mathbf{a}') \right) \right) \right) \\ &= \frac{1}{\eta^2} \mathrm{KL} \left(p_\ell^\eta, \mathrm{Uniform}(\mathcal{A}) \right) \\ &\geq 0 , \end{aligned}$$

where KL(·,·) denotes the Kullback Leibler divergence between the two distributions at arguments. Since we are assuming $\eta_\ell \leq \eta_{\ell-1}$, this implies that $\Phi_\ell(\eta_{\ell-1}) \geq \Phi_\ell(\eta_\ell)$. Thus,

$$1 + \sum_{\ell=1}^{t} \frac{1}{\eta_\ell} \ln \left(\frac{W_{\ell+1}^-}{W_\ell} \right) = \left(\sum_{\ell=1}^{t} \left(\Phi_\ell(\eta_{\ell-1}) - \Phi_\ell(\eta_\ell) \right) \right) + \Phi_{t+1}(\eta_t) \geq \Phi_{t+1}(\eta_t). \tag{16}$$

Combining (15) with (16) gives

$$\begin{aligned} \Phi_{t+1}(\eta_t) &\leq 1 + \sum_{\ell=1}^{t} \frac{1}{\eta_\ell} \ln \left(\frac{W_{\ell+1}^-}{W_\ell} \right) \\ &\leq 1 + \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) + p_\ell(\mathbf{a}) \eta_\ell \left(\widetilde{r}_\ell(\mathbf{a}) \right)^2 - \gamma_\ell p_{\mathrm{E}}(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) . \end{aligned}$$

Since for any $\bar{\mathbf{a}} \in \mathcal{A}$ the potential $\Phi_{t+1}(\eta_t)$ satisfies,

$$\Phi_{t+1}(\eta_t) = \frac{1}{\eta_t} \ln \left(\frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}\in\mathcal{A}} \exp \left(\eta_t \sum_{\ell=1}^{t} \widetilde{r}_\ell(\mathbf{a}) \right) \right) \geq \sum_{\ell=1}^{t} \widetilde{r}_\ell(\bar{\mathbf{a}}) - \frac{\ln(\mathcal{A})}{\eta_t}$$

we have

$$\sum_{\ell=1}^{t} \widetilde{r}_\ell(\bar{\mathbf{a}}) \leq 1 + \frac{\ln(\mathcal{A})}{\eta_t} + \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \sum_{\mathbf{a}\in\mathcal{A}} \left(p_\ell(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) + p_\ell(\mathbf{a}) \eta_\ell \left(\widetilde{r}_\ell(\mathbf{a}) \right)^2 - \gamma_\ell p_{\mathrm{E}}(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) \right)$$

The claimed result now follows by simply observing that

$$\sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) + p_\ell(\mathbf{a}) \eta_\ell \left(\widetilde{r}_\ell(\mathbf{a}) \right)^2 = \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widetilde{r}_\ell(\mathbf{a}) + \eta_\ell \left(\widetilde{r}_\ell(\mathbf{a}) \right)^2 \middle| \mathcal{F}_\ell \right] .$$

□

In the sequel, we shall impose the restriction

$$\gamma_t \in (0, 1/2] \,, \tag{17}$$

holding for all t, so that $\frac{1}{1-\gamma_t} \leq 2$ and $\frac{\gamma_t}{1-\gamma_t} \leq 1$.

To get a high probability anytime bound starting from Lemma 16, we are required to prove high probability bounds for each of the terms $\mathbf{I}, \mathbf{II}, \mathbf{III}$ and $\mathbf{IV}$ defined below:

$$\mathbf{I} = \sum_{\ell=1}^{t} \widetilde{r}_\ell(\bar{\mathbf{a}}).$$

$$\mathbf{II} = \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widetilde{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right].$$

$$\mathbf{III} = \sum_{\ell=1}^{t} \frac{\eta_\ell}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\left(\widetilde{r}_\ell(\mathbf{a})\right)^2 | \mathcal{F}_\ell\right]$$

$$\mathbf{IV} = -\sum_{\ell=1}^{t} \frac{\gamma_\ell}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_E} \left[\widetilde{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right]$$

We proceed by (upper or lower) bounding the four terms above in turn.

Bounding term I. By Corollary 14 with probability at least $1-\delta$ for all $\bar{\mathbf{a}} \in \mathcal{A}$ and all $t \in \mathbb{N}$ simultaneously,

$$\mathbf{I} \geq \sum_{\ell=1}^{t} \bar{\mathbf{a}}^\top \boldsymbol{\omega}_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} - \left(\max_{\ell \leq t} \frac{c(d)}{\rho\gamma_\ell} + 1\right) \ln \frac{t}{\delta'}\right) .$$

Let us denote the event where this bound holds by $\mathcal{E}_{\mathbf{I}}$. The preceding discussion implies $\mathbb{P}\left(\mathcal{E}_{\mathbf{I}}\right) \geq 1-\delta$.

Bounding term II. Recalling the definition of $\widetilde{r}_\ell(\mathbf{a})$, we can write

$$\begin{aligned}
\mathbf{II} &= \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widetilde{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right] \\
&= \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widehat{r}_\ell(\mathbf{a}) + \mathcal{O}\left(\frac{\mathbf{a}^\top \Sigma_\ell^{-1} \mathbf{a}}{\sqrt{\rho d \ell}} \sqrt{\ln\left(\frac{\ell}{\delta'}\right)}\right) \Big| \mathcal{F}_\ell\right] \\
&\overset{(i)}{=} \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widehat{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right] + \mathcal{O}\left(\frac{1}{(1-\gamma_\ell)} \sqrt{\frac{d}{\rho\ell}} \sqrt{\ln\left(\frac{\ell}{\delta'}\right)}\right) \\
&\overset{(ii)}{\leq} \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widehat{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right] + \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln\left(\frac{t}{\delta'}\right)}\right) ,
\end{aligned} \tag{18}$$

where (i) follows from Item 3 in Lemma 12, and in (ii) we have used $\frac{1}{1-\gamma_\ell} \leq 2$, $\ln\left(\frac{\ell}{\delta'}\right) \leq \ln\left(\frac{t}{\delta'}\right)$, along with $\sum_{\ell=1}^{t} \sqrt{\frac{d}{\ell}} \leq 2\sqrt{dt}$.

We are left to prove a high probability upper bound for $\sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widehat{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right]$ which we achieve through the following Lemma.

Lemma 17. *With probability at least $1-\delta$ for all $t \in \mathbb{N}$,*

$$\sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell} \left[\widehat{r}_\ell(\mathbf{a}) | \mathcal{F}_\ell\right] \leq \sum_{\ell=1}^{t} \frac{r_\ell}{1-\gamma_\ell} + \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta}} + \max_{\ell \leq t} \left(\frac{c(d)}{\rho\gamma_\ell} + 1\right) \ln \frac{t}{\delta}\right) .$$

Proof. Let $\bar{\mathbf{a}}_\ell = \mathbb{E}_{\mathbf{a}\sim p_\ell}[\mathbf{a}|\mathcal{F}_\ell]$ where the samples $\mathbf{a} \sim p_\ell$ are conditionally independent from $\mathbf{a}_\ell$. Observe that

$$\sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell}\mathbb{E}_{\mathbf{a}\sim p_\ell}[\widehat{r}_\ell(\mathbf{a})|\mathcal{F}_\ell] = \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell}\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\widehat{\boldsymbol{\omega}}_\ell^\top \mathbf{a}|\mathcal{F}_\ell\right] = \sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell}\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell \ .$$

The proof of this lemma follows closely the proof of Lemma 6 in [7]. Consider the martingale difference sequence $Y_\ell = \frac{\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell - r_\ell}{1-\gamma_\ell}$ with respect to the filtration $\{\mathcal{F}_\ell^-\}_{\ell=1}^\infty$, where we recall that $\widehat{\boldsymbol{\omega}}_\ell = b_\ell \frac{r_\ell (\boldsymbol{\Sigma}_\ell)^{-1}\mathbf{a}_\ell}{\rho}$. The process $\{Y_\ell\}_{\ell=1}^\infty$ is a martingale difference sequence w.r.t. the filtration $\{\mathcal{F}_\ell^-\}_{\ell=1}^\infty$ since $\mathbb{E}\left[\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell|\mathcal{F}_\ell^-\right] = \boldsymbol{\omega}_\ell^\top \bar{\mathbf{a}}_\ell = \mathbb{E}[r_\ell|\mathcal{F}_\ell^-]$, and therefore $\mathbb{E}\left[\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell - r_\ell|\mathcal{F}_\ell^-\right] = 0$.

The conditional variance of Y_ℓ can be bounded as follows:

$$\begin{aligned}
\mathrm{Var}\left[Y_\ell|\mathcal{F}_\ell^-\right] &= \mathbb{E}\left[(Y_\ell)^2|\mathcal{F}_\ell^-\right] \\
&= \frac{\mathbb{E}\left[\left(\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell - r_\ell\right)^2|\mathcal{F}_\ell^-\right]}{(1-\gamma_\ell)^2} \\
&\overset{(i)}{\leq} 4\mathbb{E}\left[\left(\widehat{\boldsymbol{\omega}}_\ell^\top \bar{\mathbf{a}}_\ell\right)^2\middle|\mathcal{F}_\ell^-\right] \\
&\overset{(ii)}{\leq} \frac{4\bar{\mathbf{a}}_\ell^\top \Sigma_\ell^{-1}\bar{\mathbf{a}}_\ell}{\rho} \\
&\overset{(iii)}{\leq} \frac{4\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\mathbf{a}^\top \Sigma_\ell^{-1}\mathbf{a}|\mathcal{F}_{\ell-1}\right]}{\rho} \\
&\overset{(iv)}{=} \frac{4d}{\rho} \ ,
\end{aligned}$$

where: (i) holds because $\mathbb{E}\left[\left(\widehat{\omega}_\ell^\top \bar{\mathbf{a}}_\ell - r_\ell\right)^2|\mathcal{F}_\ell^-\right] \leq \mathbb{E}\left[\left(\widehat{\omega}_\ell^\top \bar{\mathbf{a}}_\ell\right)^2|\mathcal{F}_\ell^-\right]$ and $\frac{1}{1-\gamma_\ell} \leq 2$; (ii) is a consequence of Item 4 of Lemma 12 (treating $\bar{\mathbf{a}}_\ell$ as a fixed vector); (iii) holds by Jensen's inequality; (iv) holds by Item 3 of Lemma 12. Thus, $\mathrm{Var}\left[Y_\ell|\mathcal{F}_\ell^-\right] \leq \frac{4d}{\rho}$

As a consequence, $\sum_{\ell=1}^t \mathrm{Var}\left[Y_\ell|\mathcal{F}_\ell^-\right] \leq \frac{4td}{\rho}$. Furthermore, $|Y_\ell| \leq \frac{2c(d)}{\rho\gamma_\ell} + 2$ by Item 1 in Lemma 12 and because $\frac{1}{1-\gamma_\ell} \leq 2$.

We are in a position to apply Lemma 51 (setting therein $V_t = \frac{4td}{\rho}$ and $B_t = \frac{2c(d)}{\rho\gamma_t} + 2$) to the martingale differences sequence $\{Y_\ell\}_{\ell=1}^\infty$. Rearranging terms this gives the claimed bound. □

Denote by $\mathcal{E}_{\mathbf{II}}$ the event where the bound of Lemma 17 holds. By the previous result we have $\mathbb{P}(\mathcal{E}_{\mathbf{II}}) \geq 1-\delta$. Lemma 17 along with $|r_\ell| \leq 1$ for all ℓ (see Assumption B.1) together imply the following.

Corollary 18. *If $\mathcal{E}_{\mathbf{II}}$ holds then*

$$\sum_{\ell=1}^{t} \frac{1}{1-\gamma_\ell}\mathbb{E}_{\mathbf{a}\sim p_\ell}[\widehat{r}_\ell(\mathbf{a})|\mathcal{F}_\ell] \leq \sum_{\ell=1}^{t} r_\ell + 2\gamma_\ell + \mathcal{O}\left(\sqrt{\frac{dt}{\rho}\ln\frac{t}{\delta}} + \max_{\ell\leq t}\left(\frac{c(d)}{\rho\gamma_\ell}+1\right)\ln\frac{t}{\delta}\right) \ .$$

Proof. Since $\frac{1}{1-\gamma_\ell} - 1 = \frac{\gamma_\ell}{1-\gamma_\ell}$ we immediately see that

$$\sum_{\ell=1}^{t} \frac{r_\ell}{1-\gamma_\ell} = \sum_{\ell=1}^{t} r_\ell + \frac{\gamma_\ell r_\ell}{1-\gamma_\ell}$$

And from $|r_\ell| \leq 1$ and $\frac{1}{1-\gamma_\ell} \leq 2$,

$$\sum_{\ell=1}^{t} \frac{r_\ell}{1-\gamma_\ell} = \sum_{\ell=1}^{t} r_\ell + 2\gamma_\ell \ ,$$

thereby concluding the proof. □

Finally, (18) and Corollary 18 imply that, in case $\mathcal{E}_{\mathbf{II}}$ holds,

$$\mathbf{II} \le \sum_{\ell=1}^{t} r_\ell + 2\gamma_\ell + \mathcal{O}\left(\sqrt{\frac{dt}{\rho}\ln\frac{t}{\delta}} + \max_{\ell\le t}\left(\frac{c(d)}{\rho\gamma_\ell}+1\right)\ln\frac{t}{\delta} + \sqrt{\frac{dt}{\rho}\ln\left(\frac{t}{\delta'}\right)}\right) .$$

Bounding term III. By definition of $\widetilde{r}_\ell(\mathbf{a})$, the fact that $\frac{1}{1-\gamma_\ell} \le 2$, along with the inequality $(a+b)^2 \le 2a^2 + 2b^2$, we can write

$$\begin{aligned}
\sum_{\ell=1}^{t} \frac{\eta_\ell}{1-\gamma_\ell}\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\widetilde{r}_\ell^2(\mathbf{a})|\mathcal{F}_\ell\right] &\le \sum_{\ell=1}^{t} 4\eta_\ell \mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\left(\widehat{r}_\ell(\mathbf{a})\right)^2 + \frac{4\ln\left(\frac{12\ell^2}{\delta'}\right)}{\rho d\ell}\left(\mathbf{a}^\top\Sigma_\ell^{-1}\mathbf{a}\right)^2 \Big|\mathcal{F}_\ell\right] \\
&= \underbrace{\sum_{\ell=1}^{t} 4\eta_\ell \mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\left(\widehat{r}_\ell(\mathbf{a})\right)^2 \Big|\mathcal{F}_\ell\right]}_{\mathbf{A}} \\
&\quad \underbrace{+\mathcal{O}\left(\sum_{\ell=1}^{t}\eta_\ell \mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\frac{\left(\mathbf{a}^\top\Sigma_\ell^{-1}\mathbf{a}\right)^2\ln\left(\frac{\ell}{\delta'}\right)}{\rho d\ell}\Big|\mathcal{F}_\ell\right]\right)}_{\mathbf{B}} .
\end{aligned}$$

We proceed to upper bound $\mathbf{A}$ and $\mathbf{B}$ separately. Let us start from term $\mathbf{B}$. We have

$$\mathbf{B} \overset{(i)}{=} \mathcal{O}\left(\sum_{\ell=1}^{t}\frac{\eta_\ell c(d)\ln\left(\frac{\ell}{\delta'}\right)}{\rho\ell\gamma_\ell}\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\mathbf{a}^\top\Sigma_\ell^{-1}\mathbf{a}\right]\right) \overset{(ii)}{=} \sum_{\ell=1}^{t}\mathcal{O}\left(\frac{\eta_\ell c(d) d\ln\left(\frac{\ell}{\delta'}\right)}{\rho\ell\gamma_\ell}\right) , \tag{19}$$

where (i) follows from Item 2. of Lemma 12, and (ii) follows from Item 3. of the same lemma. Notice that this upper bound holds deterministically. Let us now turn to handling term $\mathbf{A}$ now. We use a similar argument as Lemma 8 in [7].

Lemma 19. *With probability at least $1-\delta$ simultaneously for all $t \in \mathbb{N}$,*

$$\mathbf{A} = \mathcal{O}\left(\sum_{\ell=1}^{t}\frac{\eta_\ell d}{\rho} + \sqrt{\ln\left(\frac{t}{\delta}\right)\sum_{\ell=1}^{t}\frac{\eta_\ell^2 c(d) d}{\gamma_\ell\rho^3}} + \max_{\ell\le t}\frac{c(d)\eta_\ell}{\rho^2\gamma_\ell}\ln\frac{t}{\delta}\right) . \tag{20}$$

Proof. Recalling that $\widehat{\boldsymbol{\omega}}_\ell = b_\ell\frac{r_\ell(\boldsymbol{\Sigma}_\ell)^{-1}\mathbf{a}_\ell}{\rho}$, we first observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\left(\widehat{r}_\ell(\mathbf{a})\right)^2\Big|\mathcal{F}_\ell\right] &= \sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a})\widehat{\boldsymbol{\omega}}_\ell^\top\mathbf{a}\mathbf{a}^\top\widehat{\boldsymbol{\omega}}_\ell \\
&= \widehat{\boldsymbol{\omega}}_\ell^\top\left(\sum_{\mathbf{a}\in\mathcal{A}} p_\ell(\mathbf{a})\mathbf{a}\mathbf{a}^\top\right)\widehat{\boldsymbol{\omega}}_\ell \\
&= \frac{r_\ell^2 b_\ell^2}{\rho^2}\mathbf{a}_\ell^\top\Sigma_\ell^{-1}\Sigma_\ell\Sigma_\ell^{-1}\mathbf{a}_\ell \\
&= \frac{r_\ell^2 b_\ell}{\rho^2}\mathbf{a}_\ell^\top\Sigma_\ell^{-1}\Sigma_\ell\Sigma_\ell^{-1}\mathbf{a}_\ell \\
&\le \frac{b_\ell\mathbf{a}_\ell^\top\Sigma_\ell^{-1}\mathbf{a}_\ell}{\rho^2} .
\end{aligned}$$

Summing over ℓ and multiplying by $4\eta_\ell$ yields

$$\mathbf{A} \le \sum_{\ell=1}^{t}\frac{4\eta_\ell}{\rho^2}b_\ell\mathbf{a}_\ell^\top\Sigma_\ell^{-1}\mathbf{a}_\ell .$$

Now, Item 2 of Lemma 12 implies the magnitude of each of the terms $\frac{4\eta_\ell}{\rho^2}b_\ell\mathbf{a}_\ell^\top\Sigma_\ell^{-1}\mathbf{a}_\ell$ is at most $4\frac{c(d)\eta_\ell}{\rho^2\gamma_\ell}$. Moreover, Item 3 of Lemma 12 implies that for each term the conditional expectation

$\mathbb{E}\left[\frac{4\eta_\ell}{\rho^2} b_\ell \mathbf{a}_\ell^\top \Sigma_\ell^{-1} \mathbf{a}_\ell \middle| \mathcal{F}_\ell^-\right]$ equals $\frac{4\eta_\ell d}{\rho}$. As for the conditional variance, we can write

$$\begin{aligned}
\mathrm{Var}\left[\frac{4\eta_\ell}{\rho^2} b_\ell \mathbf{a}_\ell^\top \Sigma_\ell^{-1} \mathbf{a}_\ell - \frac{4\eta_\ell d}{\rho} \middle| \mathcal{F}_\ell^-\right] &\le \frac{16\eta_\ell^2}{\rho^4} \mathbb{E}\left[b_\ell^2 \left(\mathbf{a}_\ell^\top \Sigma_\ell^{-1} \mathbf{a}_\ell\right)^2 \middle| \mathcal{F}_\ell^-\right] \\
&\overset{(i)}{\le} \frac{16\eta_\ell^2 c(d)}{\gamma_\ell \rho^4} \mathbb{E}\left[b_\ell \mathbf{a}_\ell^\top \Sigma_\ell^{-1} \mathbf{a}_\ell | \mathcal{F}_\ell^-\right] \\
&\overset{(ii)}{=} \frac{16\eta_\ell^2 c(d) d}{\gamma_\ell \rho^3}
\end{aligned}$$

where (i) follows from Item 2 in Lemma 12, and (ii) is from Item 3. An application of Lemma 51 concludes the proof □

We denote by $\mathcal{E}_{\mathbf{III}}$ the event where the bound of Lemma 19 holds. By the previous result, $\mathbb{P}\left(\mathcal{E}_{\mathbf{III}}\right) \ge 1 - \delta$. Thus if $\mathcal{E}_{\mathbf{III}}$ holds, Equations 19 and 20 imply,

$$\begin{aligned}
\mathbf{III} &= \sum_{\ell=1}^t \frac{\eta_\ell}{1-\gamma_\ell} \mathbb{E}_{\mathbf{a}\sim p_\ell}\left[\widetilde{r}_\ell^2(\mathbf{a})\right] \\
&= \mathcal{O}\left(\sum_{\ell=1}^t \frac{\eta_\ell d}{\rho} + \sqrt{\ln\left(\frac{t}{\delta}\right) \sum_{\ell=1}^t \frac{\eta_\ell^2 c(d) d}{\gamma_\ell \rho^3}} + \max_{\ell \le t} \frac{c(d)\eta_\ell}{\rho^2 \gamma_\ell} \ln \frac{t}{\delta} + \sum_{\ell=1}^t \frac{\eta_\ell c(d) d \ln\left(\frac{\ell}{\delta'}\right)}{\rho \ell \gamma_\ell}\right) .
\end{aligned}$$

Bounding Term IV. Define $\mathrm{supp}(p_E) = \{\mathbf{a} \in \mathcal{A} : p_E(\mathbf{a}) > 0\}$, and recall that

$$\mathbf{IV} = -\sum_{\ell=1}^t \frac{\gamma_\ell}{1-\gamma_\ell} \sum_{\mathbf{a}\in\mathcal{A}} p_{\mathrm{E}}(\mathbf{a}) \widetilde{r}_\ell(\mathbf{a}) .$$

Let $\mathbf{a}$ be any action in $\mathrm{supp}(p_E)$, and set in Lemma 13 $\alpha_\ell = \frac{\gamma_\ell}{1-\gamma_\ell}$. This implies that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{\ell=1}^t \frac{\gamma_\ell}{1-\gamma_\ell} \widetilde{r}_\ell(\mathbf{a}) &\ge \sum_{\ell=1}^t \frac{\gamma_\ell}{1-\gamma_\ell} \mathbf{a}^\top \omega_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + \left(\frac{c(d)}{\rho} + 1\right) \ln \frac{t}{\delta'}\right) \\
&\ge -2 \sum_{\ell=1}^t \gamma_\ell - \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + \left(\frac{c(d)}{\rho} + 1\right) \ln \frac{t}{\delta'}\right)
\end{aligned}$$

the last inequality following from $|\mathbf{a}^\top \boldsymbol{\omega}_\ell| \le 1$ and $\frac{1}{1-\gamma_\ell} \le 2$.

A simple union bound along with the fact that $\sum_{\mathbf{a}\in\mathrm{supp}(p_E)} p_E(\mathbf{a}) = 1$ implies that with probability at least $1 - |\mathrm{supp}(p_E)|\delta$

$$\mathbf{IV} \le 2 \sum_{\ell=1}^t \gamma_\ell + \mathcal{O}\left(\sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + \left(\frac{c(d)}{\rho} + 1\right) \ln \frac{t}{\delta'}\right) .$$

Similar to before, we denote by $\mathcal{E}_{\mathbf{IV}}$ the event where this bound holds. By the previous result $\mathbb{P}\left(\mathcal{E}_{\mathbf{IV}}\right) \ge 1 - |\mathrm{supp}(p_E)|\delta$.

Putting it all together. We plug the bounds so obtained on $\mathbf{I} - \mathbf{IV}$ back into Lemma 16, collect common terms, and overapproximate. We obtain that, when $\mathcal{E}_{\mathbf{I}} \cap \mathcal{E}_{\mathbf{II}} \cap \mathcal{E}_{\mathbf{III}} \cap \mathcal{E}_{\mathbf{IV}}$ holds,

$$\begin{aligned}
\sum_{\ell=1}^t \left(\bar{\mathbf{a}}^\top \boldsymbol{\omega}_\ell - r_\ell\right) = \mathcal{O}\Bigg(&\frac{\ln(\mathcal{A})}{\eta_t} + \sqrt{\frac{dt}{\rho} \ln \frac{t}{\delta'}} + \max_{\ell \le t} \left(\frac{c(d)}{\rho\gamma_\ell} + 1\right) \ln \frac{t}{\delta'} + \sum_{\ell=1}^t \gamma_\ell + \sum_{\ell=1}^t \frac{\eta_\ell c(d) d \ln\left(\frac{\ell}{\delta'}\right)}{\rho \ell \gamma_\ell} \\
&+ \sum_{\ell=1}^t \frac{\eta_\ell d}{\rho} + \sqrt{\ln\left(\frac{t}{\delta}\right) \sum_{\ell=1}^t \frac{\eta_\ell^2 c(d) d}{\gamma_\ell \rho^3}} + \max_{\ell \le t} \frac{c(d)\eta_\ell}{\rho^2 \gamma_\ell} \ln \frac{t}{\delta}\Bigg) ,
\end{aligned}$$

with $\delta' = \frac{\delta}{|\mathcal{A}|}$.

Now, recall the restriction on η_t as in (14). In order to fulfill this requirement, we set

$$\eta_t = \mathcal{O}\left(\frac{\rho\gamma_t}{c(d) + \frac{c(d)}{\sqrt{dt}}\sqrt{\rho \ln|\mathcal{A}| \ln\frac{t}{\delta}}}\right) .$$

This gives

$$\sum_{\ell=1}^{t}\left(\bar{\mathbf{a}}^\top \boldsymbol{\omega}_\ell - r_\ell\right) = \mathcal{O}\Bigg(\frac{c(d)\,\ln|\mathcal{A}|}{\rho\gamma_t}\left(1 + 2\sqrt{\frac{\rho\ln|\mathcal{A}|}{dt}\ln\frac{t}{\delta}}\right) + \max_{\ell\le t}\left(\frac{c(d)}{\rho\gamma_\ell} + 1\right)\ln|\mathcal{A}|\ln\frac{t}{\delta}$$
$$+ \sum_{\ell=1}^{t}\gamma_\ell + \sum_{\ell=1}^{t}\frac{\ln|\mathcal{A}|\ln\left(\frac{\ell}{\delta}\right)}{\ell} + \sum_{\ell=1}^{t}\frac{\gamma_\ell d}{c(d)} + \sqrt{\frac{dt\ln|\mathcal{A}|}{\rho}\ln\frac{12t^2}{\delta}}$$
$$+ \sqrt{\ln\frac{t}{\delta}\sum_{\ell=1}^{t}\frac{\gamma_\ell d}{c(d)\rho}} + \frac{1}{\rho}\ln\frac{t}{\delta}\Bigg) .$$

We now set γ_ℓ so as to satisfy (17):

$$\gamma_\ell = \min\left\{\sqrt{\frac{c(d)\ln|\mathcal{A}|\ln\frac{\ell}{\delta}}{\rho\,\ell}}, \frac{1}{2}\right\} .$$

Under the assumption that $c(d) \ge d$ (see below) this gets

$$\sum_{\ell=1}^{t}\left(\bar{\mathbf{a}}^\top \boldsymbol{\omega}_\ell - r_\ell\right) = \mathcal{O}\Bigg(\sqrt{\frac{c(d)t\ln|\mathcal{A}|}{\rho}\ln\frac{t}{\delta}} + \rho\ln|\mathcal{A}|\ln\frac{t}{\delta}\ln t + \sqrt{\frac{dt\ln|\mathcal{A}|}{\rho}\ln\frac{t}{\delta}}$$
$$+ \left(\frac{\ln|\mathcal{A}|}{c(d)}\ln\frac{t}{\delta}\right)^{1/4} t^{1/4}\left(\frac{1}{\rho}\right)^{3/4} + \frac{1}{\rho}\ln\frac{t}{\delta}\Bigg) .$$

Now, from [9] (Ch. 5 therein), it is known that with John's exploration, the smallest eigenvalue of $\mathbf{\Sigma}_t$ is at least $\frac{\gamma_t}{d}$, so that the function $c(d)$ in Lemma 12 is $\le d$. Moreover, the support of John's exploration distribution has size at most $d(d+1)/2 + 1 \le 2d^2$. Combining with the last displayed equation, and taking a final union bound so as to make the events $\mathcal{E}_{\mathbf{I}}, \mathcal{E}_{\mathbf{II}}, \mathcal{E}_{\mathbf{III}}$, and $\mathcal{E}_{\mathbf{IV}}$ jointly hold concludes the proof of Theorem 10.

B.2 Exp4 algorithm for finite policy classes

Consider now the case of the Exp4 algorithm from [6]. The algorithm operates with a finite set of policies Π. An anytime high probability regret guarantee for a biased version of Exp4 can be derived by following a similar pattern as in Section B.1, but it can also be derived, e.g., by modifying the high-probability analysis for Exp3 contained in [25]. The proof is omitted.

Corollary 20. *Let the complexity $R(\Pi)$ of the policy space Π be defined as $R(\Pi) = \sqrt{|\mathcal{A}|\log|\Pi|}$. Then a version of the Exp4 algorithm from [6] exists that is h-stable in that, for $t \to \infty$ and constant ρ independent of t, its regret $\mathrm{Reg}(t)$ satisfies*

$$\mathrm{Reg}(t) = \mathcal{O}\left(R(\Pi)\sqrt{\frac{t}{\rho}\ln\frac{t}{\delta}}\right) ,$$

with probability at least $1-\delta$, where the big-oh hides terms in t which are lower order than $\sqrt{t\log t}$ as $t \to \infty$.

Regarding extendability and the ability to handle policy removals, this is fairly immediate for Exp4, and we omit the trivial details.

C Adversarial Regret Balancing and Elimination

In this section, we provide the proof of Arbe's regret bound for adversarial environments. For convenience, we restate the main theorem here:

Theorem 4. *Consider a run of Algorithm 1 with* Arbe$(\delta, 1, 0)$ *and M base algorithms with $1 \leq R(\widetilde{\Pi}_1) \leq \cdots \leq R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M-i)$ additional actions. Then with probability at least $1 - \mathrm{poly}(M)\delta$ the regret $\mathrm{Reg}(t, \Pi_M)$ for all rounds $t \geq i_\star$ is bounded by*

$$\mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star} + M\right) R(\widetilde{\Pi}_{i_\star})\sqrt{i_\star t \ln \frac{t}{\delta}}\right) , \tag{4}$$

where $i_\star$ is the smallest index of the base algorithm that is h-stable.

The proof of this regret bound relies on the following regret bound of Arbe in between restarts:

Lemma 21 (Regret per Epoch of Arbe)**.** *Consider a run of Algorithm 1 with* Arbe(δ, s, t_0) *let $T \in \mathbb{N} \cup \{\infty\}$ be the round when the algorithm restarts ($T = \infty$ if there is no restart). Then the regret against Π_M is bounded with probability at least $1 - poly(M)\delta$ for all $t \in [T]$ simultaneously as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_M) = \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right).$$

Further, if $s \geq i^\star$, then the algorithm does not restart, i.e., $T = \infty$.

With this result, Theorem 4 can be proven quickly:

Proof of Theorem 4. Denote by $t_0 = 0$ and t_i the round of the i-th restart and ∞ if it does not exist for $i \in [i_\star]$. By Lemma 21, there can be at most $i_\star - 1$ restarts and thus $t_{i_\star} = \infty$. The total regret of Arbe$(\delta, 1, 0)$ can be decomposed into the regret between two restarts

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \sum_{i=1}^{i_\star} \mathrm{Reg}([t_{i-1}+1, \min\{t_i, t\}], \Pi_M) .$$

We can now plug in the bound from Lemma 21 for each term on the RHS which gives

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \sum_{s=1}^{i_\star} \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{\bar{t}_s \ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right) . \tag{21}$$

where $\bar{t}_s = \max\{\min\{t_s, t\} - t_{s-1}, 0\}$. We can bound this further using Jensen's inequality and the fact that $\sum_{s=1}^{i_\star} \bar{t}_s = t$ as

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star} + M\right) R(\widetilde{\Pi}_{i_\star})\sqrt{i_\star t \ln\frac{t}{\delta}} + Mi_\star \ln\frac{\ln t}{\delta}\right) \tag{22}$$

When $t \geq i_\star$ the last term is dominated by the others and hence, the proof is complete. □

C.1 Proof of Lemma 21

We first show that there are at most $i_\star$ restarts. This is because, due to extendability and h-stability of learners above $i_\star$, the elimination test can never trigger for them. The following lemma makes this argument formal:

Lemma 22. *With probability at least $1 - poly(M)\delta$, the elimination test in Equation 2 never triggers for i, j with $i_\star \leq i < j \leq M$.*

Proof. Let $i, j \in [i_\star, M] = \{i_\star, i_\star + 1, \ldots, M\}$ and $t \in \mathbb{N}$ and denote by t_0 the round of the last restart before t. By definition of $i_\star$ in Section 2.1, base learner i is h-stable and extendable. We consider the event where the regret bound of i holds and where the statements in Lemma 23 hold.

This happens for all $i \geq i_\star$ with probability at least $1 - \text{poly}(M)\delta$. Since $i \geq i_\star$ is h-stable and extendable, we have

$$\max_{\pi \in \widetilde{\Pi}_i} \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi}[r_\ell(a, x_\ell)] - r_\ell(a^i_\ell, x_\ell) \leq cR(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t}{\delta}}$$

where c is an absolute constant. Since the extended policy class $\widetilde{\Pi}_i$ includes an action $\widetilde{a}_j$ that always follows base learner $\mathbb{A}_j$, i.e., $r_\ell(\widetilde{a}_j, x_\ell) = r_\ell(a^j_\ell, x_\ell)$, we have in particular

$$\sum_{\ell=t_0+1}^{t} \Big[r_\ell(a^j_\ell, x_\ell) - r_\ell(a^i_\ell, x_\ell)\Big] \leq cR(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t}{\delta}} \ .$$

Using now Equation 25 from Lemma 23, we have

$$\begin{aligned}
\widetilde{\text{CRew}}_j(t_0, t) - \widetilde{\text{CRew}}_i(t_0, t) &\leq \text{D}_j(t_0, t) + \text{D}_i(t_0, t) + \sum_{\ell=t_0+1}^{t} \Big[r_\ell(a^j_\ell, x_\ell) - r_\ell(a^i_\ell, x_\ell)\Big] \\
&\leq \text{D}_j(t_0, t) + \text{D}_i(t_0, t) + cR(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t}{\delta}}
\end{aligned}$$

and thus, the test in Equation 2 does not trigger for i and j in round t. □

We are now ready to do the proof of Lemma 21.

Proof of Lemma 21. By Lemma 22 there are at most $i_\star$ restarts and $s \leq i_\star$ at all times. The regret of Arbe in rounds $[t_0 + 1, t]$ against any policy $\pi' \in \Pi_M$ can be written as

$$\sum_{\ell=t_0+1}^{t} [\mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - r_\ell(a_\ell, x_\ell)] = \sum_{i=s}^{M} \sum_{\ell=t_0+1}^{t} [\rho_i \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = i\} r_\ell(a_\ell, x_\ell)] \ .$$

For $i \geq i_\star$, we can bound the summand on the RHS directly using h-stability and extendability of i as

$$\begin{aligned}
&\sum_{\ell=t_0+1}^{t} [\rho_i \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = i\} r_\ell(a_\ell, x_\ell)] \\
&\quad = \rho_i \sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \frac{\mathbf{1}\{b_\ell = i\} r_\ell(a^i_\ell, x_\ell)}{\rho_i}\right] \\
&\quad \leq \rho_i \sum_{\ell=t_0+1}^{t} \big[\mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - r_\ell(a^i_\ell, x_\ell)\big] + \rho_i \text{D}_i(t_0, t) && \text{(Lemma 23)} \\
&\quad \leq cR(\widetilde{\Pi}_i)\sqrt{(t-t_0)\rho_i \ln \frac{t}{\delta}} + \rho_i \text{D}_i(t_0, t) && \text{(h-stability of } \mathbb{A}_i) \\
&\quad = \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{(t-t_0)\rho_i \ln \frac{t}{\delta}} + \sqrt{(t-t_0)\rho_i \ln \frac{\ln t}{\delta}} + \ln \frac{\ln t}{\delta}\right) && \text{(definition of } \text{D}_i) \\
&\quad = \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{(t-t_0)\rho_i \ln \frac{t}{\delta}} + \ln \frac{\ln t}{\delta}\right) && \text{(since } R(\widetilde{\Pi}_i) \geq 1) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{t-t_0}{\sum_{j=s}^{M} R(\widetilde{\Pi}_j)^{-2}} \ln \frac{t}{\delta}} + \ln \frac{\ln t}{\delta}\right) && \text{(definition of } \rho_i) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{t-t_0}{R(\widetilde{\Pi}_{i_\star})^{-2}} \ln \frac{t}{\delta}} + \ln \frac{\ln t}{\delta}\right) && \text{(since } s \leq i_\star) \\
&\quad = \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0) \ln \frac{t}{\delta}} + \ln \frac{\ln t}{\delta}\right) \ . && (23)
\end{aligned}$$

For $i < i_\star$, we cannot rely on h-stability and extendability and instead have to use the fact that the misspecification test in Equation 2 did not trigger until the last round where there can be at most a regret of 1. This allows us to bound

$$
\begin{aligned}
&\sum_{\ell=t_0+1}^{t} \left[\rho_i \mathbb{E}_{a\sim\pi'}[r_\ell(a,x_\ell)] - \mathbf{1}\{b_\ell = i\}\, r_\ell(a_\ell, x_\ell)\right] \\
&= \rho_i \sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi'}[r_\ell(a,x_\ell)] - \frac{\mathbf{1}\{b_\ell = i\}\, r_\ell(a_\ell^i, x_\ell)}{\rho_i}\right] \\
&\le \rho_i \sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi'}[r_\ell(a,x_\ell)] - \frac{\mathbf{1}\{b_\ell = i_\star\}\, r_\ell(a_\ell^{i_\star}, x_\ell)}{\rho_{i_\star}}\right] \\
&\qquad + \rho_i \mathrm{D}_i(t_0,t) + \rho_i \mathrm{D}_{i_\star}(t_0,t) + R(\widetilde{\Pi}_i)\sqrt{\rho_i (t-t_0)\ln\frac{t}{\delta}} + 1 && \text{(misspecification test failed)} \\
&\le \rho_i \sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi'}[r_\ell(a,x_\ell)] - r_\ell(a_\ell^{i_\star}, x_\ell)\right] && \text{(Lemma 23)} \\
&\qquad + 2\rho_i \mathrm{D}_{i_\star}(t_0,t) + R(\widetilde{\Pi}_i)\sqrt{\rho_i (t-t_0)\ln\frac{t}{\delta}} + 1 && (\mathrm{D}_{i_\star}(t_0,t) \ge \mathrm{D}_i(t_0,t)) \\
&\le \rho_i c R(\widetilde{\Pi}_{i_\star})\sqrt{\frac{t-t_0}{\rho_{i_\star}}\ln\frac{t}{\delta}} + 2\rho_i \mathrm{D}_{i_\star}(t_0,t) + R(\widetilde{\Pi}_i)\sqrt{\rho_i (t-t_0)\ln\frac{t}{\delta}} + 1 && \text{(h-stability of } \mathbb{A}_{i_\star}) \\
&= \mathcal{O}\left(\rho_i R(\widetilde{\Pi}_{i_\star})\sqrt{\frac{t-t_0}{\rho_{i_\star}}\ln\frac{t}{\delta}} + \sqrt{\rho_i}\sqrt{\frac{\rho_i}{\rho_{i_\star}}(t-t_0)\ln\frac{t}{\delta}} + R(\widetilde{\Pi}_i)\sqrt{\rho_i}\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) \\
& && \text{(definition of } \mathrm{D}_{i_\star}) \\
&= \mathcal{O}\left(\left(\frac{\rho_i R(\widetilde{\Pi}_{i_\star})}{\sqrt{\rho_{i_\star}}} + R(\widetilde{\Pi}_i)\sqrt{\rho_i}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) \\
&= \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_i)} + R(\widetilde{\Pi}_i)\right)\sqrt{\rho_i (t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) \\
&= \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_i)}\sqrt{\rho_i (t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) && (R(\widetilde{\Pi}_i) \le R(\widetilde{\Pi}_{i_\star}))
\end{aligned}
$$

Finally, combining both bounds yields

$$
\begin{aligned}
\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_M) &= \mathcal{O}\left(\left(\sum_{i=i_\star}^{M} R(\widetilde{\Pi}_{i_\star}) + \sum_{i=s}^{i_\star-1}\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_i)}\sqrt{\rho_i}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right) \\
&= \mathcal{O}\left(\left(M R(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right).
\end{aligned}
$$

□

C.2 Concentration Bounds on Reward Sequences

Lemma 23. *With probability at least $1 - poly(M)\delta$, the following inequalities hold for all base learners $i \in [M]$ for all time steps $t \in \mathbb{N}$ where $\mathbb{A}_i$ was not eliminated yet*

$$
\left|\sum_{\ell=t_0+1}^{t}\left[\mathbb{E}_{a\sim\pi_\ell^i}[r_\ell(a,x_\ell)] - r_\ell(a_\ell^i, x_\ell)\right]\right| = \mathcal{O}\left(\sqrt{(t-t_0)\ln\frac{\ln(t-t_0)}{\delta}}\right) \tag{24}
$$

$$
\left|\sum_{\ell=t_0+1}^{t}\left[\frac{\mathbf{1}\{b_\ell = i\}\, r_\ell(a_\ell,x_\ell)}{\rho_i} - r_\ell(a_\ell^i, x_\ell)\right]\right| = \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i}\ln\frac{\ln(t-t_0)}{\delta}\right) \tag{25}
$$

where t_0 is the time of the last restart of Algorithm 1 before t and ρ_i is the probability with which learner i is chosen in round t.

Proof. There can be at most M restarts of Algorithm 1. Thus, we can prove the concentration bounds for a single restart and a single base learner i and obtain the statement for all restarts with a union bound over M^2.

Consider first Equation 24 and let

$$\mathcal{F}_\ell = \sigma\left(\{r_j, x_j, \pi_1^j, \{a_j^i\}_{i\in[M]}, b_j\}_{j\in[t_0+1,\ell-1]} \cup \{r_\ell, x_\ell, \pi_\ell^i, \{a_\ell^k\}_{k\in[M]\setminus\{i\}}\}\right)$$

be the sigma field of all previous reward functions, contexts and actions as well as the context and action in the current round. Further let τ be the stopping time w.r.t. $\{\mathcal{F}_\ell\}$ of when the algorithm restarts, and denote

$$X_\ell = \mathbf{1}\{\ell \geq \tau\}\left(\mathbb{E}_{a\sim\pi_\ell^i}[r_\ell(a, x_\ell)] - r_\ell(a_\ell^i, x_\ell)\right) .$$

The sequence $\{X_\ell\}_{\ell>t_0}$ is a martingale difference sequence w.r.t. $\{\mathcal{F}_\ell\}_{\ell>t_0}$ and $X_\ell \in [-1, 1]$ almost surely for all ℓ. Then by Lemma 46 (setting $m = 1$ and $a_t = -1, b_t = 1$) with probability at least $1-\delta$, we have for all $t > t_0$

$$\sum_{\ell=t_0+1}^{t} X_\ell \leq 1.44\sqrt{(t-t_0)\left(1.4\ln\ln\left(4\left(t-t_0\right)\right) + \ln\frac{5.2}{\delta}\right)} = \mathcal{O}\left(\sqrt{(t-t_0)\ln\frac{\ln(t-t_0)}{\delta}}\right)$$

We can apply the same argument to $-\sum_{\ell=t_0+1}^{t} X_\ell$ which proves Equation 24.

Consider now Equation 25 and let

$$\mathcal{F}_\ell = \sigma\left(\{r_j, x_j, \pi_1^j, \{a_j^i\}_{i\in[M]}, b_j\}_{j\in[t_0+1,\ell-1]} \cup \{r_\ell, x_\ell, \pi_\ell^i, \{a_\ell^k\}_{k\in[M]}\}\right) .$$

Again, let τ be the stopping time w.r.t. $\{\mathcal{F}_\ell\}_\ell$ of when the algorithm restarts and denote

$$X_\ell = \mathbf{1}\{\ell \geq \tau\}\left[\frac{\mathbf{1}\{b_\ell = i\}\, r_\ell(a_\ell^i, x_\ell)}{\rho_i} - r_\ell(a_\ell^i, x_\ell)\right]$$

which is a martingale difference sequence w.r.t. $\{\mathcal{F}_\ell\}_{\ell>t_0}$. We have $X_\ell \leq \frac{1}{\rho_i}$ almost surely and $\mathbb{E}_[X_\ell^2|\mathcal{F}_\ell] \leq \frac{1}{\rho_i}$. By Lemma 47 (with $m = 1/\rho_i$), this implies that with probability at least $1-\delta$ for all $t > t_0$

$$\sum_{\ell=t_0+1}^{t} X_\ell \leq \underbrace{1.44\sqrt{\frac{t-t_0}{\rho_i}\left(1.4\ln\ln\left(4(t-t_0)\right) + \ln\frac{5.2}{\delta}\right)} + \frac{0.41}{\rho_i}\left(1.4\ln\ln\left(4(t-t_0)\right) + \ln\frac{5.2}{\delta}\right)}_{=:\mathrm{D}_i(t_0,t)} .$$

We define the RHS as the precise definition of $\mathrm{D}_i(t_0, t)$ used in Algorithm 1. Note that

$$\mathrm{D}_i(t_0, t) = \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i}\ln\frac{\ln(t-t_0)}{\delta}\right) \tag{26}$$

are required. Finally, we can apply the exact same argument to $-\sum_{\ell=t_0+1}^{t} X_\ell$ which finishes the proof. □

D Adversarial Regret Balancing and Elimination with Best of Both Worlds Regret

In this section, we provide the proofs of the main regret bound for Arbe-Gap in Theorem 5 and describe the Arbe-GapExploit subroutine in detail. We restate the theorem her for convenience:

Theorem 5. *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$. Then with probability at least $1 - poly(M)\delta$, the following two conditions hold for all $t \geq M^2$ simultaneously. In any adversarial or stochastic environment $\mathbb{B}$, the regret is bounded as*

$$\mathrm{Reg}(t, \Pi_M) = \mathcal{O}\left(\left(M + \ln(t) + \frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right) R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t) + i_\star)\ln\frac{t}{\delta}}\right) .$$

If $\mathbb{B}$ is stochastic and there is a unique policy with gap $\Delta > 0$, then the pseudo-regret is bounded as,

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\frac{R(\Pi_M)^2}{\Delta}\ln(t)\ln\frac{t}{\delta} + \frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln^2\left(\frac{M R(\widetilde{\Pi}_M)}{\Delta\delta}\right)\right). \tag{5}$$

We break the proof of this statement into several parts based on the phases of the algorithm. The first phase end when Arbe-Gap calls Arbe-GapExploit and the second phase are all rounds played by Arbe-GapExploit. Finally, in case Arbe-GapExploit was called but terminated at some point, we have a final phase where we simply execute Arbe. For the regret in this final phase, we can directly use the guarantees of Arbe. The behavior in the first two phases is analyzed below. The following lemma characterizes the regret and pseudo-regret in the first phase until Arbe-GapExploit is called. It also ensures that if the environment is stochastic, the inputs of Arbe-GapExploit are correct, i.e., $\widehat{\pi}$ is the optimal policy and $\widehat{\Delta}$ is an accurate estimate of its gap.

Lemma 24 (Guarantee for First Phase)**.** *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$. Further, let $t_{\mathrm{gap}} \in \mathbb{N} \cup \{\infty\}$ be the round where the Arbe-GapExploit subroutine is called. Then with probability at least $1 - poly(M)\delta$, the following conditions hold for all rounds $t \in [2i_\star, t_{\mathrm{gap}}]$. The regret is bounded as*

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\left(M + \frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right) R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t) + i_\star)\ln\frac{t}{\delta}}\right).$$

If $\mathbb{B}$ is stochastic and there is a unique policy $\pi_\star$ with gap $\Delta > 0$, then the gap estimator $\widehat{\Delta}$ and policy $\widehat{\pi}$ passed onto Arbe-GapExploit satisfy $\frac{\Delta}{2} \leq \widehat{\Delta} \leq \Delta$ and $\widehat{\pi} = \pi_\star$. Further, the pseudo-regret is bounded as

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln\left(\frac{M R(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right).$$

The proof of this result can be found in Appendix D.2. To characterize the regret and pseudo-regret of the second phase, we use the following main properties of the Arbe-GapExploit routine in Algorithm 4. It ensures that the regret and pseudoregret are well controlled and that the routine never terminates if the environment was indeed stochastic with a gap.

Lemma 25 (Guarantee for Second Phase)**.** *Let $\mathbb{A}$ be an h-stable learner with policy class $\Pi_{\mathbb{A}}$. Then the regret of Algorithm 4 against $\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}$ is bounded with probability at least $1 - \mathcal{O}(\delta)$ for all rounds $t > t_0$ that the algorithm has not terminated yet as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0 + 1, t], \Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t - t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

Further, if the environment is stochastic with an optimal policy $\pi_\star$ that has a gap Δ compared to the best policy in $\Pi_{\mathbb{A}}$ and the inputs satisfy $\widehat{\pi} = \pi_\star$ and $\widehat{\Delta} \leq \Delta \leq 2\widehat{\Delta}$, then with probability at least $1 - \mathcal{O}(\delta)$ the pseudo-regret of Algorithm 4 is bounded in all rounds $t > t_0$ as

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0 + 1, t], \{\widehat{\pi}\} \cup \Pi_{\mathbb{A}}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right)\right)$$

and the algorithm never terminates.

The proof of this statement is available in Appendix D.3. While combining Lemma 24 and Lemma 25 gives the desired pseudo-regret guarantee in Theorem 5 above for stochastic environments fairly directly, the bound on the regret in Theorem 5 requires more work. Arbe-GapExploit guarantees only guarantees that the regret is of order $\tilde{\mathcal{O}}(R(\Pi_M \setminus \{\widehat{\pi}\})^2/\widehat{\Delta} + \sqrt{t})$ while we would like a bound that does not scale with $R(\Pi_M \setminus \{\widehat{\pi}\}$ in our final result. To achieve that, we will use the following lemma which states that the length of the initial phase t_{gap} has to be sufficiently large as a function of the gap estimate $\widehat{\Delta}$. This will allow us to absorb the $R(\Pi_M \setminus \{\widehat{\pi}\})^2/\widehat{\Delta}$ term into the regret of the first phase.

Lemma 26. *Consider a run of Algorithm 2 with inputs $t_0 = 0, n = 1$, arbitrary policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$ with $1 \le R(\widetilde{\Pi}_1) \le R(\widetilde{\Pi}_2) \le \dots \le R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M+1-i)$ additional actions. Let t_{gap} be the round where ArbeGap-Exploit was called with gap estimate $\widehat{\Delta}$. Then with probability at least $1 - poly(M)\delta$*

$$\widehat{\Delta} = \Omega\left(\frac{R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_{i_\star})}\sqrt{\frac{\ln\frac{t_{gap}}{\delta}}{t_{gap}}}\right).$$

Proof. ArbeGap calls ArbeGap-Exploit as soon as $2W(t_0, t) \le \widehat{\Delta}_t$. Hence, when the call happened in round t_{gap}, we must have $2W(t_0, t_{gap}) \le \widehat{\Delta}$ or, plugging in the definition of $W(t_0, t_{gap})$ with an appropriate absolute constant c

$$\frac{cR(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{n(k)}{\delta}}{k}} + \frac{c}{\rho_M}\frac{\ln\frac{n\ln(k)}{\delta}}{k} \le \widehat{\Delta},$$

where $k = t_{gap} - t_0$. We can further lower-bound the LHS as

$$\begin{aligned}
\frac{cR(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{nk}{\delta}}{k}} + \frac{c}{\rho_M}\frac{\ln\frac{n\ln(k)}{\delta}}{k} &\ge \frac{cR(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{k}{\delta}}{k}} \\
&\ge cR(\widetilde{\Pi}_M)^2\sqrt{\sum_{i=s}^{M} R(\widetilde{\Pi}_i)^{-2}}\sqrt{\frac{\ln\frac{k}{\delta}}{k}} \\
&\ge c\frac{R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_{i_\star})^2}\sqrt{\frac{\ln\frac{k}{\delta}}{k}}
\end{aligned}$$

where the last equation holds because $i_\star$ is never eliminated with high probability Lemma 28. Finally, since the function on the RHS is monotonically decreasing in k, we can further lower-bound this quantity by replacing k with $t_{gap} \ge k$. Hence, we have $\widehat{\Delta} \ge c\frac{R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_{i_\star})^2}\sqrt{\frac{\ln\frac{t_{gap}}{\delta}}{t_{gap}}}$. Reordering terms gives the desires statement. □

We are now ready to state the proof of Theorem 5:

Proof of Theorem 5. We first consider stochastic environments with a gap and apply Lemma 24. We denote by t_{gap} the round where Arbe-Gap calls Arbe-GapExploit. For all $t \le t_{gap}$, the pseudo-regret is bounded as

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right).$$

Further, Arbe-GapExploit can only be called with $\Pi_{\mathbb{A}} = \Pi_M \setminus \{\pi_\star\}$, $\widehat{\pi} = \pi_\star$ and $\widehat{\Delta}$ that satisfies $\widehat{\Delta} \le \Delta \le 2\widehat{\Delta}$. This allows us to apply Lemma 25 to bound the pseudo-regret of any round played by Arbe-GapExploit as

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_{gap}+1, t], \Pi_M) = \mathcal{O}\left(\frac{R(\Pi_M)^2}{\Delta}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_M)}{\Delta\delta}\right)\right).$$

It further tells us that, with high probability, Arbe-GapExploit will never return. Hence, we can bound the pseudo-regret of both phases to get a bound on the total pseudo-regret after any number of rounds

$$\begin{aligned}\mathrm{PseudoReg}_{\mathbb{M}}(t,\Pi_M) &= \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t_{gap}}{\delta}\right.\\ &\qquad\left.+\frac{R(\Pi_M)^2}{\Delta}\left(\ln(t)\ln\frac{t}{\delta}+\ln\frac{R(\Pi_M)}{\Delta\delta}\right)\right)\\ &= \mathcal{O}\left(\frac{R(\Pi_M)^2}{\Delta}\ln(t)\ln\frac{t}{\delta}+\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln^2\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\right)\end{aligned}$$

where we upper-bounded t_{gap} using a crude upper-bound $\mathcal{O}\left(\frac{R(\widetilde{\Pi}_M)^6 M^5}{\Delta^4\delta^2}\right)$ of the bound in Lemma 34 which gives that $\ln(t_{gap}) = \mathcal{O}(\ln(MR(\widetilde{\Pi}_M)/\Delta\delta))$.

We now move on to the regret bound in any environment. Again, Lemma 24 gives us a regret bound that holds with high probability for any round $t \le t_{gap}$ of

$$\mathrm{Reg}_{\mathbb{M}}(t,\Pi_M) = \mathcal{O}\left(\left(M+\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right).$$

If t falls into a round that is played by the routine Arbe-Gap, then Lemma 25 the regret after t_{gap} and before t is bounded as

$$\mathrm{Reg}_{\mathbb{M}}([t_{gap}+1,t],\Pi_M) = \mathcal{O}\left(\frac{R(\Pi_M)^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta}+\ln\frac{R(\Pi_M)}{\widehat{\Delta}\delta}\right)+\sqrt{(t-t_{gap})\ln(t)\ln\frac{t}{\delta}}\right).$$

Using Lemma 26, we can further bound $\frac{1}{\widehat{\Delta}} = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_M)^2}\sqrt{\frac{t_{gap}}{\ln(1/\delta)}}\right)$ and plugging this into the bound above gives

$$\begin{aligned}&\mathrm{Reg}_{\mathbb{M}}([t_{gap}+1,t],\Pi_M)\\ &= \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{t_{gap}}\left(\ln^{3/2}(t)\ln^{1/2}\frac{t}{\delta}+\ln(t_{gap}R(\Pi_M))\right)+\sqrt{(t-t_{gap})\ln(t)\ln\frac{t}{\delta}}\right)\\ &= \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{t}\ln^{3/2}(t)\ln^{1/2}\frac{t}{\delta}+\sqrt{t\ln(t)\ln\frac{t}{\delta}}\right).\end{aligned}$$

Here, we also used the fact that $t_{gap} = \Omega(R(\Pi_M))$ since the test in Line 14 of Algorithm 2 can only trigger when $W(t_0,t) \le R(\Pi_M)^2$ which is only possible after at least $\Omega(R(\Pi_M))$ rounds. Finally, if t falls into a round after Arbe-GapExploit returned (in round t_{adv}, then the regret since the return can be bounded using the regret bound of Arbe in Theorem 4 as

$$\mathrm{Reg}_{\mathbb{M}}([t_{adv}+1,t],\Pi_M) = \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}+M\right)R(\widetilde{\Pi}_{i_\star})\sqrt{i_\star t\ln\frac{t}{\delta}}\right),$$

Combining the bounds from all three possible phases gives the following bound that holds for all $t > M^2$ as

$$\begin{aligned}\mathrm{Reg}_{\mathbb{M}}(t,\Pi_M) &= \mathcal{O}\left(\left(M+\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right.\\ &\qquad\left.+R(\widetilde{\Pi}_{i_\star})\sqrt{t}\ln^{3/2}(t)\ln^{1/2}\frac{t}{\delta}\right)\\ &= \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}+M+\ln(t)\right)R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right)\end{aligned}$$

□

Algorithm 4: Arbe-GapExploit

Input: current round t_0, learner $\mathbb{A}$ with policy class $\Pi_{\mathbb{A}}$, candidate policy $\widehat{\pi}$, gap estimate $\widehat{\Delta}$, failure probability δ

Initialize $k_0 = \Theta\left(\frac{R^2(\Pi_{\mathbb{A}})}{\widehat{\Delta}^2} \ln \frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right)$

for *epoch* $e = 0, 1, 2 \ldots$ **do**

  Set next epoch length $k_{e+1} = 2k_e$ and final round $t_{e+1} = t_e + k_e$

  Set learner probability $\rho^e = \Theta\left(\frac{R^2(\Pi_{\mathbb{A}})}{k_e \widehat{\Delta}^2} \ln \frac{k_e}{\delta_e}\right)$ where $\delta_e = \frac{\delta}{(e+1)^2}$

  Restart $\mathbb{A}$ with failure probability δ_e

  for *round* $t = t_e + 1, t_e + 2, \ldots, t_{e+1}$ **do**

    Set π_t^1 as the current policy of $\mathbb{A}$ and $\pi_t^0 = \widehat{\pi}$

    Sample $b_t \sim \text{Bernoulli}(\rho^e)$

    Get context x_t and compute $a_t^i \sim \pi_t^i(\cdot|x_t)$ for $i \in \{0, 1\}$

    Play action $a_t = a_t^{b_t}$ and receive reward $r_t(a_t, x_t)$

    Update learner $\mathbb{A}$ with reward $\frac{b_t r_t(a_t, x_t)}{\rho^e}$

    Set $V(t) = \Theta\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{\ln \frac{t-t_e}{\delta_e}}{\rho^e(t-t_e)}} + \frac{\ln \frac{\ln(t-t_e)}{\delta_e}}{t-t_e}\right)$

    if $\frac{\widetilde{\text{CRew}}_0(t_e+1,t) - \widetilde{\text{CRew}}_1(t_e+1,t)}{t-t_e} < \widehat{\Delta} - V(t)$ **then**

      return `// environment is adversarial`

    if $\frac{\widetilde{\text{CRew}}_0(t_e+1,t) - \widetilde{\text{CRew}}_1(t_e+1,t)}{t-t_e} > 4\widehat{\Delta} + V(t)$ **then**

      return `// environment is adversarial`

D.1 Description of the Second Phase: Gap Exploitation

We present the Arbe-GapExploit algorithm of the Gap Exploitation phase as a general procedure that takes a focus policy $\widehat{\pi}$, a policy class $\Pi_{\mathbb{A}}$ and a gap estimate $\widehat{\Delta}$ and is tasked with testing the hypothesis '$\widehat{\pi}$ is the optimal policy, and has a gap of order $\Theta(\widehat{\Delta})$', incurring in small regret while doing so.

We use Arbe-GapExploit with input policy class $\Pi_{\mathbb{A}} = \Pi_M$. If it ever returns, the learner can conclude the environment is adversarial and thus, start playing Arbe with the policy classes $\Pi_s, \ldots, \Pi_M$ not yet eliminated by the misspecification tests during Arbe-Gap. We develop results for the more general case when the input algorithm and policy class equal $\mathbb{A}$ and $\Pi_{\mathbb{A}}$. In case the environment is stochastic with gap Δ and $\widehat{\Delta} = \Theta(\Delta)$, we show Arbe-GapExploit has a pseudo regret of order $\mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln(t) \ln \frac{t}{\delta} + \ln \frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right)\right)$ at time t (see Lemma 38). Similarly we show that Arbe-GapExploit has an adversarial regret rate of order $\mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t) \ln \frac{t}{\delta} + \ln \frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right)$ at time t (see Lemma 41). The adversarial rate consists of a poly-logarithmic factor with an upfront multiplier of order $\frac{R^2(\Pi_{\mathbb{A}})}{\widehat{\Delta}}$ plus a factor scaling with $\sqrt{t-t_0}$. Since in our case $\widehat{\Delta} = \Omega\left(\frac{R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_{i_\star})}\sqrt{\frac{M \ln \frac{t_{gap}}{\delta}}{t_{gap}}}\right)$ (see Lemma 26) and $\Pi_{\mathbb{A}} = \Pi_M$, the adversarial regret has an upper bound of the form $\widetilde{\mathcal{O}}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{t_{\text{gap}}} + \sqrt{t - t_{\text{gap}}}\right)$ (where $\widetilde{\mathcal{O}}$ hides polylogarithmic factors) thus satisfying a model selection guarantee.

In Arbe-GapExploit we divide time into epochs indexed from $e = 0, 1, \ldots$ of length $k_e = k_0 \cdot 2^e$ where $k_0 = \Theta\left(\frac{R^2(\widetilde{\Pi}_{\mathbb{A}}) \log \frac{R(\widetilde{\Pi}_{\mathbb{A}})}{\widehat{\Delta}\delta}}{\widehat{\Delta}^2}\right)$. During epoch e, learner $\mathbb{A}$ is sampled with probability $\rho^e = \Theta\left(\frac{R^2(\widetilde{\Pi}_{\mathbb{A}}) \log\left(\frac{k_e}{\delta_e}\right)}{k_e \widehat{\Delta}^2}\right)$. We define k_0 and ρ^e so that for all epochs $1 - \rho^e \geq \frac{1}{2}$. Thus

for all $t \in \{t_e+1, \cdots, t_{e+1}\}$, it holds that $\left|\widetilde{\mathrm{CRew}}_0(t_e+1,t) - \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi}}\left[r_\ell(a,x_\ell)\right]\right| = \mathcal{O}\left(\sqrt{(t-t_e)\ln\frac{t-t_e}{\delta_e}}\right)$. By the h-stability of $\mathbb{A}$, and using a concentration argument, we prove that $\widetilde{\mathrm{CRew}}_1(t_e+1,t)$ can be used to estimate $\max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi}\left[r_\ell(a,x_\ell)\right]$ up to $V(t) = \Theta\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{\ln\frac{t-t_e}{\delta_e}}{\rho^e(t-t_e)}} + \frac{\ln\frac{\ln(t-t_e)}{\delta_e}}{t-t_e}\right)$ accuracy.

When the environment is stochastic, $\widehat{\pi}$ is the optimal policy and $\Delta/2 \le \widehat{\Delta} \le \Delta$ it follows that

$$(t-t_e)\widehat{\Delta} \le \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi},x\sim\mathcal{D}}\left[r(a,x)\right] - \max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(a,x)\right] \le 4\widehat{\Delta}\,.$$

Therefore when the condition in line 14 or 16 of Algorithm 4 trigger, we would have found evidence that

$$\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi},x\sim\mathcal{D}}\left[r(a,x)\right] - \max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(a,x)\right] < (t-t_e)\widehat{\Delta}$$

or

$$\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi},x\sim\mathcal{D}}\left[r(a,x)\right] - \max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(a,x)\right] > 4(t-t_e)\widehat{\Delta}$$

thus indicating the environment cannot be stochastic.

The observations above imply that in case the environment is stochastic the tests of lines 14 and 16 in Algorithm 4 do not trigger. Let us jump to the task of analyzing the regret of Arbe-GapExploit in stochastic environments. We will assume t lies in epoch e. By the h-stability of $\mathbb{A}$, the sum of its pseudo-rewards in a stochastic environment satisfies

$$\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi^1_\ell,x\sim\mathcal{D}}\left[r(a,x)\right] + \underbrace{\mathcal{O}\left(R(\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\})\sqrt{\frac{(t-t_e)}{\rho^e}\ln\frac{t-t_e}{\delta_e}}\right)}_{\mathbf{I}} \ge \max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(a,x)\right]\,.$$

In this case, pseudo-regret is only incurred when $b_t = 1$. From t_e+1 to t the variable b_t equals 1 an average of $\rho^e(t-t_e)$ times. Let us use the notation $\pi'_\star = \max_{\pi\in\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}} \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(x,a)\right]$. The regret collected during these rounds can be upper bounded by

$$\mathcal{O}\left(\underbrace{\Delta\cdot\rho^e(t-t_e)}_{\text{Regret of } \pi'_\star \text{ w.r.t. } \widehat{\pi}} + \underbrace{\rho^e\times\mathbf{I}}_{\text{Regret of } \mathbb{A} \text{ w.r.t. } \pi'_\star}\right) = \mathcal{O}\left(\Delta\rho^e(t-t_e) + R(\Pi_{\mathbb{A}}\setminus\{\widehat{\pi}\}))\sqrt{\rho^e(t-t_e)\ln\frac{t-t_e}{\delta_e}}\right).$$

Substituting in the value of ρ^e and using the fact that $\widehat{\Delta} = \Theta(\Delta)$ when the environment is stochastic allow us to write

$$\begin{aligned}\Delta\cdot\rho^e(t-t_e) + \rho^e\cdot\mathbf{I} &= \mathcal{O}\left(\frac{R^2(\Pi_{\mathbb{A}})}{k_e\Delta}(t-t_e)\ln\left(\frac{k_e}{\delta_e}\right) + \frac{R^2(\Pi_{\mathbb{A}})}{\Delta}\sqrt{\frac{t-t_e}{k_e}\ln\left(\frac{t-t_e}{\delta_e}\right)\ln\left(\frac{k_e}{\delta_e}\right)}\right)\\ &= \mathcal{O}\left(\frac{R^2(\Pi_{\mathbb{A}})}{\Delta}\ln\left(\frac{k_e}{\delta_e}\right)\right)\,.\end{aligned}$$

Summing over all epochs $e' \le e$, and using $\sum_{e'=0}^{e}\ln\left(\frac{k_{e'}}{\delta_{e'}}\right) = \mathcal{O}\left(\ln(t)\ln\left(\frac{t}{\delta}\right) + \ln\left(\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right)\right)$ we conclude that

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta} + \ln(t)\ln\frac{t}{\delta}\right)\right)$$

and thus the proof sketch of Lemma 38.

We now bound the adversarial regret of Arbe-GapExploit during the timesteps before Lines 14 or 16 of Algorithm 4 trigger. In this case, the h-stability of $\mathbb{A}$ implies,

$$\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi_\ell^1}\left[r_\ell(a,x_\ell)\right] + \underbrace{\mathcal{O}\left(R(\Pi_{\mathbb{A}}\backslash\{\widehat{\pi}\})\sqrt{\frac{(t-t_e)}{\rho^e}\ln\frac{t-t_e}{\delta_e}}\right)}_{\mathbf{II}(t,t_e)} \geq \max_{\pi\in\Pi_{\mathbb{A}}\backslash\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi}\left[r_\ell(a,x_\ell)\right]. \tag{27}$$

While Lines 14 or 16 of Algorithm 4 have not triggered, we can certify that with high probability

$$\widehat{\Delta}(t-t_e) - V_e(t)(t-t_e) \leq \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi}}\left[r_\ell(a,x_\ell)\right] - \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi_\ell^1}\left[r_\ell(a,x_\ell)\right] \leq 4\widehat{\Delta}(t-t_e) + V_e(t)(t-t_e). \tag{28}$$

In the above, we used the notation $V_e(t)$ to denote the $V(t)$ of epoch e. Hence the regret collected during rounds $\{t_e+1,\ldots,t\}$ can be upper bounded by the sum of three terms

$$\underbrace{\max\left(\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi}}\left[r_\ell(a,x_\ell)\right] - \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi_\ell}\left[r_\ell(a,x_\ell)\right], 0\right)}_{\mathbf{B}}$$
$$+ \underbrace{\max\left(\max_{\pi\in\Pi_{\mathbb{A}}\backslash\{\widehat{\pi}\}} \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi}\left[r_\ell(a,x_\ell)\right] - \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi}}\left[r_\ell(a,x_\ell)\right], 0\right)}_{\mathbf{C}}$$
$$+ \underbrace{\sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\widehat{\pi}}\left[r_\ell(a,x_\ell)\right] - r_\ell(a_\ell,x_\ell)}_{\mathbf{D}}.$$

From t_e+1 to t the variable b_t equals 1 an average of $\rho^e(t-t_e)$ times. Note that $V(t)(t-t_e) \approx \mathcal{O}\left(\widehat{\Delta}\sqrt{k_e(t-t_e)}\right)$. We can bound $\mathbf{B}, \mathbf{C}, \mathbf{D}$ individually as follows:

- $\mathbf{B}$ is the Pseudo-Regret of $\mathbb{A}$ w.r.t $\widehat{\pi}$, and can be upper bounded by $\mathcal{O}\left(\widehat{\Delta}\cdot\rho^e(t-t_e) + \widehat{\Delta}\cdot\rho^e\sqrt{(t-t_e)k_e}\right) = \mathcal{O}\left(\widehat{\Delta}\cdot\rho^e\sqrt{(t-t_e)k_e}\right)$ as a consequence of multiplying the right hand side of Equation 28 by ρ^e.
- $\mathbf{C}$ is the Pseudo-Regret of $\widehat{\pi}$ w.r.t. $\Pi_{\mathbb{A}}\backslash\{\widehat{\pi}\}$, and can be upper bounded by $\mathbf{II}(t,t_e) - \widehat{\Delta}(t-t_e) + V_e(t)(t-t_e)$ as a consequence of combining Equations 27 and the left hand side of 28.
- $\mathbf{D}$ is the difference between sample rewards vs. Pseudo-Rewards and can be bound by Hoeffding's inequality.

Let $t'_e = \min(t,t_e)$. Summing the upper bound $\mathbf{C}$ over all epochs $e' \leq e$, we choose the multiplier c in $\rho^e = \frac{cR^2(\Pi_{\mathbb{A}})}{k_e\widehat{\Delta}^2}\ln\frac{k_e}{\delta_e}$ such that[9]

$$\sum_{e'\leq e} \mathbf{II}(t'_{e'},t_{e'}) - \widehat{\Delta}(t'_{e'}-t_{e'}) + V_{e'}(t'_{e'})(t'_{e'}-t_{e'}) = \sum_{e'\leq e} \mathcal{O}\left(\widehat{\Delta}\sqrt{\frac{k_e(t'_{e'}-t_{e'})}{c}}\right) - \widehat{\Delta}(t'_{e'}-t_{e'})$$
$$\leq \mathcal{O}\left(\widehat{\Delta}k_0\right). \tag{29}$$

Combining the bounds for $\mathbf{B}$ and $\mathbf{D}$,

$$\mathcal{O}\left(\underbrace{\widehat{\Delta}\cdot\rho^e(t-t_e)}_{\mathbf{B}} + \underbrace{\sqrt{(t-t_e)\ln\left(\frac{t-t_e}{\delta_e}\right)}}_{\mathbf{D}}\right) = \mathcal{O}\left(\frac{R^2(\Pi_{\mathbb{A}})}{\widehat{\Delta}}\ln\left(\frac{k_e}{\delta_e}\right) + \sqrt{(t-t_e)\ln\left(\frac{t-t_e}{\delta_e}\right)}\right).$$

[9] Increasing the value of c implies we have to set k_0 to be larger. This has the only effect of increasing the constant on the RHS of Eq. 29.

Summing over all epochs and using the upper bounds

$$\sum_{e'=0}^{e} \ln\left(\frac{k_{e'}}{\delta_{e'}}\right) = \mathcal{O}\left(\ln(t)\ln\left(\frac{t}{\delta}\right) + \ln\left(\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right)\right)$$

$$\sum_{e'=0}^{e} \sqrt{(t'_{e'} - t_{e'})\ln\left(\frac{t'_{e'} - t_{e'}}{\delta_{e'}}\right)} = \mathcal{O}\left(\sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right)$$

and the bound from Eq. 29 along with the observation $\widehat{\Delta}k_0 = \mathcal{O}\left(\frac{R^2(\Pi_{\mathbb{A}})}{\widehat{\Delta}}\ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right)$ allows us to conclude,

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}}\cup\{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right),$$

and with it we finish the proof sketch of Lemma 41. Combining these results finalizes the proof of Lemma 25.

D.2 Analysis of the First Phase

We will use the notation t_{gap} to denote the (random) time when the Arbe-Gap estimation phase ends (see Algorithm 2).

Lemma 27 (Guarantee for First Phase)**.** *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \ldots, \mathbb{A}_M$. Further, let $t_{\mathrm{gap}} \in \mathbb{N} \cup \{\infty\}$ be the round where the Arbe-GapExploit subroutine is called. Then with probability at least $1 - poly(M)\delta$, the following conditions hold for all rounds $t \in [2i_\star, t_{\mathrm{gap}}]$. The regret is bounded as*

$$\mathrm{Reg}_{\mathbb{M}}(t,\Pi_M) = \mathcal{O}\left(\left(M + \frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)R(\widetilde{\Pi}_{i_\star})\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right).$$

If $\mathbb{B}$ is stochastic and there is a unique policy $\pi_\star$ with gap $\Delta > 0$, then the gap estimator $\widehat{\Delta}$ and policy $\widehat{\pi}$ passed onto Arbe-GapExploit satisfy $\frac{\Delta}{2} \le \widehat{\Delta} \le \Delta$ and $\widehat{\pi} = \pi_\star$. Further, the pseudo-regret is bounded as

$$\mathrm{PseudoReg}_{\mathbb{M}}(t,\Pi_M) = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right).$$

D.2.1 Adversarial Guarantees

We first start by bounding the number of restarts of the algorithm:

Lemma 28 (Number of Restarts of Arbe-Gap)**.** *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \ldots, \mathbb{A}_M$. Then for any total number of rounds t, there are at most $\ln(t)$ restarts due to a change in candidate policy (Line 17 in Algorithm 2) up to that round t. Further, with probability at least $1 - poly(M)\delta$, there are at most $i_\star - 1$ restarts due an elimination of a base learner (Line 11 in Algorithm 2).*

Proof. We first show the bound on the number of restarts due to changes in the candidate policy $\widehat{\pi}$. Let $t_1, t_2, \ldots$ be the rounds at which a restart is triggered in Line 17 of Algorithm 2 and $\widehat{\pi}_1, \widehat{\pi}_2, \ldots$ be the candidate policies selected at those restarts. For each restart i, we know that $\widehat{\pi}_i$ was selected in at least $\frac{3t_i}{4}$ of the first t_i rounds and therefore also t_{i+1} rounds. However, since the policy changed from $\widehat{\pi}_i$ to $\widehat{\pi}_{i+1}$ at round t_{i+1}, we also know that $\widehat{\pi}_i$ can only be selected at most $\frac{t_{i+1}}{4}$ of the first t_{i+1} rounds. Combining both bounds yields

$$\frac{3t_i}{4} \le \frac{t_{i+1}}{4}$$

and thus $t_{i+1} \ge 3t_i$ holds for all i. Since also $t_1 \ge 9$ by the condition in the algorithm, up to round t, there can only be $\log_3(t) - 1 \le \ln(t) - 1 \le \ln(t)$ restarts.

Finally, the number of restarts due to base learner elimination is bounded by $i_\star - 1$ since this condition can never trigger for $i \ge i_\star$ by Lemma 22 (which also holds for Arbe-Gap). □

The following lemma now bounds the regret within each restart:

Lemma 29. *Consider a run of Algorithm 2 with* Arbe-Gap$(\delta, s, t_0, \widehat{\pi})$ *where* $s \leq i_\star$ *and let* $T \in \mathbb{N} \cup \{\infty\}$ *be the round when the algorithm restarts or calls* Arbe-GapExploit *(*$T = \infty$ *if there is no restart or transition to the second phase). Then the regret against* Π_M *is bounded with probability at least* $1 - poly(M)\delta$ *for all* $t \in [t_0 + 1, T] = \{t_0 + 1, t_0 + 2, \ldots, T\}$ *simultaneously as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_M) = \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right).$$

Proof. The regret of Arbe-Gap in rounds $[t_0 + 1, t]$ against any policy $\pi' \in \Pi_M$ can be written as

$$\sum_{\ell=t_0+1}^{t} [\mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - r_\ell(a_\ell, x_\ell)] = \sum_{i=s}^{M+1}\sum_{\ell=t_0+1}^{t} [\rho_i \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = i\} r_\ell(a_\ell, x_\ell)] \ .$$

For $i \in [s, M]$, we can follow the analysis of Arbe and apply the arguments in the proof of Lemma 21 verbatim. This yields with probability at least $1 - \mathrm{poly}(M)\delta$

$$\begin{aligned}
&\sum_{i=s}^{M}\sum_{\ell=t_0+1}^{t} [\rho_i \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = i\} r_\ell(a_\ell, x_\ell)] \\
&\qquad = \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{(t-t_0)\ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right). \qquad (30)
\end{aligned}$$

It only remains to bound the regret contribution of the special base learner $\mathbb{A}_{M+1}$, which does not exist in Arbe. To do so, we will use the fact that the gap test in Line 14 can only trigger in the last round before a restart happens. This allows us to relate the regret of $\mathbb{A}_{M+1}$ to that of $\mathbb{A}_M$. For the regret of $\mathbb{A}_M$, we again use the arguments in the proof of Lemma 21 (Equation 23 specifically) verbatim to show

$$\sum_{\ell=t_0+1}^{t} [\rho_M \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = M\} r_\ell(a_\ell, x_\ell)] = \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right). \qquad (31)$$

The regret contribution of $\mathbb{A}_{M+1}$ can now be bounded as

$$\begin{aligned}
&\sum_{\ell=t_0+1}^{t} [\rho_{M+1} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathbf{1}\{b_\ell = M+1\} r_\ell(a_\ell, x_\ell)] \\
&= \rho_{M+1}\left(\left[\sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)]\right] - \widetilde{\mathrm{CRew}}_{M+1}(t_0+1, t)\right) && \text{(definition of } \widetilde{\mathrm{CRew}}_{M+1}) \\
&= \rho_{M+1}\left(\left[\sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)]\right] - \widetilde{\mathrm{CRew}}_{M}(t_0+1, t)\right) \\
&\quad + \rho_{M+1}\left(\widetilde{\mathrm{CRew}}_{M}(t_0+1, t) - \widetilde{\mathrm{CRew}}_{M+1}(t_0+1, t)\right) \\
&\leq \rho_{M+1}\left(\left[\sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)]\right] - \widetilde{\mathrm{CRew}}_{M}(t_0+1, t)\right) \\
&\quad + 7\rho_{M+1} W(t_0, t)(t - t_0) + 1 && \text{(gap test not triggered at } t-1) \\
&\leq \frac{\rho_{M+1}}{\rho_M}\mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) + 7\rho_{M+1} W(t_0, t)(t - t_0) + 1 && \text{(Equation 31)} \\
&\leq \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) + 7\rho_{M} W(t_0, t)(t - t_0) && (\rho_M \geq \rho_{M+1})
\end{aligned}$$

$$= \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) + \mathcal{O}\left(R(\widetilde{\Pi}_M)\sqrt{\rho_M(t-t_0)\ln\frac{t}{\delta}}\right) \quad \text{(definition of } W\text{)}$$

$$= \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta} + \sqrt{\frac{t-t_0}{\sum_{i=s}^{M+1} R(\widetilde{\Pi}_i)^{-2}}\ln\frac{t}{\delta}}\right) \quad \text{(definition of } \rho_M\text{)}$$

$$= \mathcal{O}\left(R(\widetilde{\Pi}_{i_\star})\sqrt{(t-t_0)\ln\frac{t}{\delta}} + \ln\frac{\ln t}{\delta}\right) \quad (s \le i_\star)$$

Note that $\rho_M \ge \rho_{M+1}$ holds without loss of generality since $\widetilde{\Pi}_{M+1}$ contains 2 fewer policies than $\widetilde{\Pi}_M$. Finally, the desired statement follows by combining the previous display with Equation 30. □

Equipped with the previous two lemmas, we can now prove the regret bound of the first phase for adversarial environments:

Lemma 30. *Consider a run of Algorithm 2 with inputs $t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$. Further, let $t_{\mathrm{gap}} \in \mathbb{N} \cup \{\infty\}$ be the round where the* Arbe-GapExploit *subroutine is called. Then with probability at least $1 - \mathrm{poly}(M)\delta$, the following conditions hold for all rounds $t \in [2i_\star, t_{\mathrm{gap}}]$. The regret is bounded as*

$$\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{t(\ln(t) + i_\star)\ln\frac{t}{\delta}}\right).$$

Proof. Let $\tau_0, \tau_1, \dots$ be the rounds where Arbe-Gap restarts or eventually calls Arbe-GapExploit. By convention, we set $\tau_0 = 0$ and $\tau_i = \infty$ when there are less than i total calls to Arbe-Gap. By Lemma 28, there are at most $\ln(t) + i_\star$ calls of Arbe-Gap up to round t with probability at least $1 - \mathrm{poly}(M)\delta$ for all $t \in \mathbb{N}$ jointly. Further, by Lemma 29, the regret in each of these calls is bounded with probability $1 - \mathrm{poly}(M)\delta$ as well. If we were to apply a naive union bound over all $\ln(t) + i_\star$, then our failure probability would increase at a rate of $\ln(t)$. However, we can easily avoid this by choosing the absolute constants in the definition of D_i appropriately. A factor of 3 larger is sufficient. This ensures that each of these terms is effectively at least as large as if we had invoked them with $\frac{\delta}{n^2}$ instead of δ in the n-th restart of Arbe-Gap. We now illustrate why this is true. Let c' be the absolute constant such that

$$\mathrm{D}_i(t_0, t) = c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln t}{\delta}} + \frac{c'}{\rho_i}\ln\frac{\ln t}{\delta}$$

satisfies the necessary concentration bounds (see Section C.2) for a single restart of Arbe-Gap. Now, we have

$$\begin{aligned}
3c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln t}{\delta}} + \frac{3c'}{\rho_i}\ln\frac{\ln t}{\delta} &\ge c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{(\ln t)^9}{\delta^9}} + \frac{c'}{\rho_i}\ln\frac{(\ln t)^3}{\delta^3} \\
&\ge c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{(\ln t)^3}{\delta}} + \frac{c'}{\rho_i}\ln\frac{(\ln t)^3}{\delta} \\
&\ge c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln t}{\frac{\delta}{\ln^2 t}}} + \frac{c'}{\rho_i}\ln\frac{\ln t}{\frac{\delta}{\ln^2 t}} \\
&\ge c'\sqrt{\frac{t-t_0}{\rho_i}\ln\frac{\ln t}{\frac{\delta}{n^2}}} + \frac{c'}{\rho_i}\ln\frac{\ln t}{\frac{\delta}{n^2}}
\end{aligned}$$

where the last step holds because the number of calls n to Arbe-Gap due to a change in candidate policy at round t is bounded as $\ln(t)$. Hence, with this choice of constant, we can ensure that the statement of Lemma 29 holds with probability at least $1 - \mathrm{poly}(M)\delta$ jointly for all calls of Arbe-Gap (for the remaining $i_\star \le M$ restarts possible due to elimination of a base learner, we apply a standard union bound).

Now, just as in the proof of Theorem 4, we write the regret of Arbe-Gap using $\bar{t}_s = \max\{\min\{\tau_s, t\} - \tau_{s-1}, 0\}$ as

$$
\begin{aligned}
&\mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) \\
&= \sum_{i=1}^{\ln(t)+i_\star} \mathrm{Reg}([\tau_{i-1}+1, \min\{\tau_i, t\}], \Pi_M) \\
&= \sum_{i=1}^{\ln(t)+i_\star} \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{\bar{t}_i \ln\frac{t}{\delta}} + M\ln\frac{\ln t}{\delta}\right) && \text{(Lemma 29)} \\
&= \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{(\ln(t)+i_\star)\sum_{i=1}^{\ln(t)+i_\star}\bar{t}_i \ln\frac{t}{\delta}}\right) \\
&\quad + \mathcal{O}\left(M(\ln(t)+i_\star)\ln\frac{\ln t}{\delta}\right) && \text{(Jensen's inequality)} \\
&= \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right) \\
&\quad + \mathcal{O}\left(M(\ln(t)+i_\star)\ln\frac{\ln t}{\delta}\right) && (\textstyle\sum_s \bar{t}_s \le t) \\
&= \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right) && (t \ge 2i_\star \text{ by assumption})
\end{aligned}
$$

This concludes the proof. □

D.2.2 Stochastic Guarantees

As a first step, we show that Arbe-Gap always maintains valid confidence bounds on the gap of the candidate policy in a stochastic environment:

Lemma 31 (Confidence bounds on the gap)**.** *Consider a run of Algorithm 2 with inputs $n = 1, t_0 = 0$, arbitrary policy policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \dots, \mathbb{A}_M$ with $1 \le R(\widetilde{\Pi}_1) \le R(\widetilde{\Pi}_2) \le \dots \le R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M+1-i)$ additional actions. Assume the environment $\mathbb{B}$ is stochastic and there is a policy $\pi_\star$ with gap $\Delta > 0$. Then with probability at least $1 - poly(M)\delta$ in all rounds $t \in \mathbb{N}$*

$$\widehat{\Delta}_t \le \Delta_{\widehat{\pi}} \le \widehat{\Delta}_t + 2W(t_0, t)$$

where $W(t_0, t)$ is the term used in the definition of $\widehat{\Delta}_t$ in the algorithm with

$$W(t_0, t) = \Theta\left(\frac{R(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{n(t-t_0)}{\delta}}{t-t_0}} + \frac{1}{\rho_M}\frac{\ln\frac{n\ln(t-t_0)}{\delta}}{t-t_0}\right),$$

and t_0 and n are the time and number of the last restart and before t and $\Delta_{\widehat{\pi}} = \mathbf{1}\{\widehat{\pi} = \pi_\star\}\Delta$ is the gap of the candidate policy $\widehat{\pi}$ in round t.

Proof. First, consider a single restart of Arbe-Gap. We note that both base learner $\mathbb{A}_M$ and $\mathbb{A}_{M+1}$ are h-stable on their respective policy classes $\widetilde{\Pi}_M$ and $\widetilde{\Pi}_{M+1}$ by assumption (removing a single policy usually does not impede h-stability, see e.g. Appendix B). Further, denote by $\pi_\star^i = \mathrm{argmin}_{\pi \in \widetilde{\Pi}_i} \mathbb{E}_{a\sim\pi, x\sim\mathcal{D}}[r(a,x)]$ a best policy in policy class $\widetilde{\Pi}_i$. We can use these insights and definition to derive the following lower-bound on $\widetilde{\mathrm{CRew}}_i(t_0, t)$ for $i \in \{M, M+1\}$ that holds

uniformly with probability at least $1 - \text{poly}(M)\delta$

$$
\begin{aligned}
&\widetilde{\text{CRew}}_i(t_0, t) \\
&\geq \sum_{\ell=t_0+1}^{t} r_\ell(a_\ell^i, x_\ell) - \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && \text{(Lemma 23)} \\
&\geq \max_{\pi' \in \widetilde{\Pi}_i} \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t-t_0}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) \\
& && \text{(h-stability of } \mathbb{A}_i) \\
&\geq \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\star^i}[r_\ell(a, x_\ell)] - \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t-t_0}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && (\pi_\star^i \in \widetilde{\Pi}_i) \\
&\geq \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\star^i, x\sim\mathcal{D}}[r(a, x)] - \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t-t_0}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) \\
&\quad - \mathcal{O}\left(\sqrt{(t-t_0) \ln \frac{\ln t}{\delta}}\right) && \text{(Lemma 43)} \\
&= \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\star^i, x\sim\mathcal{D}}[r(a, x)] - \mathcal{O}\left(R(\widetilde{\Pi}_i)\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{t-t_0}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) .
\end{aligned}
$$

Conversely, using similar concentration arguments, we can derive the following upper-bound for $\widetilde{\text{CRew}}_i(t_0, t)$ for $i \in \{M, M+1\}$ that holds uniformly with probability at least $1 - \text{poly}(M)\delta$:

$$
\begin{aligned}
&\widetilde{\text{CRew}}_i(t_0, t) \\
&\leq \sum_{\ell=t_0+1}^{t} r_\ell(a_\ell^i, x_\ell) + \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && \text{(Lemma 23)} \\
&\leq \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\ell^i}[r_\ell(a, x_\ell)] + \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && \text{(Lemma 23)} \\
&\leq \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\ell^i, x\sim\mathcal{D}}[r(a, x)] + \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && \text{(Lemma 43)} \\
&\leq \sum_{\ell=t_0+1}^{t} \mathbb{E}_{a\sim\pi_\star^i, x\sim\mathcal{D}}[r(a, x)] + \mathcal{O}\left(\sqrt{\frac{t-t_0}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}} + \frac{1}{\rho_i} \ln \frac{\ln(t-t_0)}{\delta}\right) && \text{(def. of } \pi_\star^i)
\end{aligned}
$$

Note that $\Delta_{\widehat{\pi}} = \mathbb{E}_{a\sim\pi_\star^M, x\sim\mathcal{D}}[r(a, x)] - \mathbb{E}_{a\sim\pi_\star^{M+1}, x\sim\mathcal{D}}[r(a, x)]$ which is either Δ if $\widehat{\pi} = \pi_\star$ or 0 if $\widehat{\pi} \neq \pi^\star$. Combining the bounds on $\widetilde{\text{CRew}}_i$ with a union bound, we can derive the following deviation bound

$$
\begin{aligned}
&- \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_M)}{\sqrt{\rho_M}} + \frac{1}{\sqrt{\rho_{M+1}}}\right)\sqrt{\frac{\ln \frac{t-t_0}{\delta}}{t-t_0}} + \left(\frac{1}{\rho_M} + \frac{1}{\rho_{M+1}}\right)\frac{\ln \frac{\ln(t-t_0)}{\delta}}{t-t_0}\right) \\
&\qquad \leq \frac{\widetilde{\text{CRew}}_M(t_0, t) - \widetilde{\text{CRew}}_{M+1}(t_0, t)}{t-t_0} - \Delta_{\widehat{\pi}} \leq \\
&+ \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_{M+1})}{\sqrt{\rho_{M+1}}} + \frac{1}{\sqrt{\rho_M}}\right)\sqrt{\frac{\ln \frac{t-t_0}{\delta}}{t-t_0}} + \left(\frac{1}{\rho_M} + \frac{1}{\rho_{M+1}}\right)\frac{\ln \frac{\ln(t-t_0)}{\delta}}{t-t_0}\right) .
\end{aligned}
$$

We can further simplify those bounds by noting that $R(\widetilde{\Pi}_{M+1}) \leq R(\widetilde{\Pi}_M)$ and thus also $\rho_{M+1} \geq \rho_M$ since $\widetilde{\Pi}_M$ is identical to $\widetilde{\Pi}_{M+1}$ but contains two more policies. Thus, we can bound the magnitude

of the upper and lower bound further by

$$\mathcal{O}\left(\frac{R(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{t-t_0}{\delta}}{t-t_0}}+\frac{1}{\rho_M}\frac{\ln\frac{\ln(t-t_0)}{\delta}}{t-t_0}\right).$$

We now rebind δ by $\frac{\delta}{n^2}$ and apply a union bound over all restarts of Arbe-Gap. Thus, we can choose a constant in the definition of

$$\mathrm{W}(t_0,t)=\Theta\left(\frac{R(\widetilde{\Pi}_M)}{\sqrt{\rho_M}}\sqrt{\frac{\ln\frac{n(t-t_0)}{\delta}}{t-t_0}}+\frac{1}{\rho_M}\frac{\ln\frac{n\ln(t-t_0)}{\delta}}{t-t_0}\right)$$

large enough so that

$$-\mathrm{W}(t_0,t)\le\frac{\widetilde{\mathrm{CRew}}_M(t_0,t)-\widetilde{\mathrm{CRew}}_{M+1}(t_0,t)}{t-t_0}-\Delta_{\widehat{\pi}}\le\mathrm{W}(t_0,t)$$

holds for all t in all possible restarts of Arbe-Gap with the desired $1-\mathrm{poly}(M)\delta$ probability. □

The lemma above immediately implies the correctness of the first phase, in the sense that if the algorithm moves on to the second phase in a stochastic environment, the candidate policy has to be optimal and the gap estimate is accurate up to a multiplicative factor:

Corollary 32. *Consider a run of Algorithm 2 with inputs $n=1, t_0=0$, arbitrary policy policy $\widehat{\pi}\in\Pi_M$ and M base learners $\mathbb{A}_1,\dots,\mathbb{A}_M$ with $1\le R(\widetilde{\Pi}_1)\le R(\widetilde{\Pi}_2)\le\cdots\le R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M+1-i)$ additional actions. Assume the environment $\mathbb{B}$ is stochastic and there is a policy $\pi_\star$ with gap $\Delta>0$. Then with probability at least $1-\mathrm{poly}(M)\delta$ the policy $\widehat{\pi}$ and gap estimate $\widehat{\Delta}$ passed to Arbe-GapExploit satisfy*

$$\widehat{\pi}=\pi_\star\qquad\text{and}\qquad\widehat{\Delta}\le\Delta\le2\widehat{\Delta}.$$

Proof. The statement follows from Lemma 31 and the condition in Line 14 of Algorithm 2. First, note that the test cannot trigger when $\widehat{\pi}\ne\pi_\star$ since $\widehat{\Delta}_t\le0$ in this case. Second, since $\widehat{\Delta}$ satisfies $2W(t_0,t)\le\widehat{\Delta}$ and $\widehat{\Delta}\le\Delta\le\widehat{\Delta}+2W(t_0,t)$ when the test triggers, we have

$$\widehat{\Delta}\le\Delta\le\widehat{\Delta}+2W(t_0,t)\le2\widehat{\Delta}\ ,$$

as claimed. □

We now move on to show that if there is a policy with a gap, the alorithm has to identify it within a certain number of rounds:

Lemma 33 (Arbe-Gap selects the right candidate policy)**.** *Consider a run of Algorithm 2 with inputs $t_0=0$, arbitrary policy policy $\widehat{\pi}\in\Pi_M$ and M base learners $\mathbb{A}_1,\dots,\mathbb{A}_M$ with $1\le R(\widetilde{\Pi}_1)\le R(\widetilde{\Pi}_2)\le\cdots\le R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M+1-i)$ additional actions. Assume the environment $\mathbb{B}$ is stochastic and there is a policy $\pi_\star$ with gap $\Delta>0$. Then with probability at least $1-\mathrm{poly}(M)\delta$ the number of rounds until $\pi_\star$ is always chosen as the candidate policy $\widehat{\pi}$ is bounded as*

$$\mathcal{O}\left(\left(M^2+\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)^2}i_\star\right)\frac{R(\widetilde{\Pi}_{i_\star})^2 i_\star}{\Delta^2}\ln^2\frac{MR(\widetilde{\Pi}_{i_\star})}{\Delta\delta}\right).$$

Proof. By Lemma 30, with probability at least $1-\mathrm{poly}(M)\delta$ the regret of Arbe-Gap is bounded for all rounds $t\in[2i_\star,t_{\mathrm{gap}}]$ as

$$\begin{aligned}\mathrm{Reg}_{\mathbb{M}}(t,\Pi_M)&=\mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star})+\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{t(\ln(t)+i_\star)\ln\frac{t}{\delta}}\right)\\&=\mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star})\sqrt{i_\star}+\frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}i_\star\right)\sqrt{t}\ln\frac{t}{\delta}\right).\end{aligned}$$

By the concentration argument in Lemma 42, the same bound can be established for the pseudo-regret

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) \leq c \left(MR(\widetilde{\Pi}_{i_\star})\sqrt{i_\star} + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)} i_\star \right) \sqrt{t} \ln \frac{t}{\delta} \tag{32}$$

for some sufficiently large absolute constant c. Now denote $\gamma = c \left(MR(\widetilde{\Pi}_{i_\star})\sqrt{i_\star} + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)} i_\star \right)$ and consider the value

$$t' = \frac{16^2\gamma^2}{\Delta^2} \ln^2 \frac{8\gamma}{\Delta\delta} \, .$$

Then by the properties of $\frac{\ln(t)}{\sqrt{t}}$ investigated in Lemma 52, we can bound for $t \geq t'$

$$\begin{aligned}
\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) &\leq \gamma\sqrt{t} \ln \frac{t}{\delta} && \text{(Equation 32)} \\
&= \gamma t \frac{\ln(t/\delta)}{\sqrt{t}} \leq \gamma t \frac{\Delta}{4\gamma} && \text{(Lemma 52)} \\
&= \frac{t}{4}\Delta \, .
\end{aligned}$$

We have shown that the adversarial regret rate implies that the pseudo-regret for rounds $t \geq t'$ has to be bounded by $\frac{t}{4}\Delta$. Since each policy but $\pi^\star$ incurs a pseudo-regret at least Δ per round, Arbe-Gap has to select $\pi^\star$ in at least $\frac{3}{4}t$ among all t rounds to satisfy this pseudo-regret bound. As a result, a switch of the candidate policy to $\pi^\star$ would be triggered if it is not already the candidate policy. Further, no other policy can be selected more than a quarter of the times, thus the candidate policy has to be $\pi^\star$ in all rounds $t \geq t'$. □

Lemma 34 (Number of Rounds in the First Phase)**.** *Consider a run of Algorithm 2 with inputs $t_0 = 0, n = 1$, arbitrary policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \ldots, \mathbb{A}_M$ with $1 \leq R(\widetilde{\Pi}_1) \leq R(\widetilde{\Pi}_2) \leq \cdots \leq R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M + 1 - i)$ additional actions. Assume the environment $\mathbb{B}$ is stochastic and there is a policy $\pi_\star$ with gap $\Delta > 0$. Then with probability at least $1 - poly(M)\delta$ the number of rounds until the algorithm enters the second phase by calling ArbeGap-Exploit is bounded as*

$$\mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_M)^4}{R(\widetilde{\Pi}_1)^2} + MR(\widetilde{\Pi}_{i_\star})^2 \right) \frac{Mi_\star}{\Delta^2} \ln^2 \frac{MR(\widetilde{\Pi}_{i_\star})}{\Delta\delta} \right) \, .$$

Proof. By Lemma 33, after a certain number of rounds t_{pol}, the candidate policy has to be $\pi_\star$ at all rounds. Hence, there can be no restarts due to candidate policy switches anymore. According to Lemma 28, there can only be up to $i_\star$ restarts after round t_{pol} due to elimination of a base learner. We will in the following show that if ArbeGap has been (re)started with candidate policy $\pi_\star$ and there are no other restarts in the meantime, it has to switch to the second phase within a certain number of rounds k. The total number of rounds in the first phase, is then bounded by

$$t_{pol} + i_\star \cdot k \, .$$

We will now show a bound on k. By Lemma 31, we have at all times that $\widehat{\Delta}_t \leq \Delta$ (since $\widehat{\pi} = \pi_\star$ by assumption) and the algorithm moves on to the second phase as soon as $2W(t_0, t) \leq \widehat{\Delta}_t$. Note that the condition of $\widehat{\Delta}_t \leq \Delta \leq 1 \leq R(\widetilde{\Pi}_M)^2$ is always satisfies in stochastic environments. Hence,the algorithm cannot stay in the first phase if $2W(t_0, t) \leq \Delta$ or, plugging in the definition of $W(t_0, t)$ with an appropriate absolute constant c

$$\frac{cR(\widetilde{\Pi}_M)}{\sqrt{\rho_M}} \sqrt{\frac{\ln \frac{n(t-t_0)}{\delta}}{t - t_0}} + \frac{c}{\rho_M} \frac{\ln \frac{n \ln(t-t_0)}{\delta}}{t - t_0} \leq \Delta.$$

Hence, we can obtain an value for the bound k by identifying a value that satisfies

$$\frac{\ln \frac{k}{\delta/n}}{k} \leq \frac{\Delta\rho_M}{2c} \quad \text{and} \frac{\ln \frac{k}{\delta/n}}{k} \leq \frac{\Delta^2\rho_M}{4c^2 R(\widetilde{\Pi}_M)^2} \, .$$

Since $\Delta \in (0,1]$ and $c, R(\widetilde{\Pi}_M) \geq 1$ without loss of generality, it is sufficient to only consider the condition on the right. By Lemma 53, we can set k as

$$k = \frac{16c^2 R(\widetilde{\Pi}_M)^2}{\Delta^2 \rho_M} \ln\left(\frac{2n}{\delta}\frac{4c^2 R(\widetilde{\Pi}_M)^2}{\Delta^2\rho_M}\right) = \mathcal{O}\left(\frac{MR(\widetilde{\Pi}_M)^4}{R(\widetilde{\Pi}_1)^2\Delta^2}\ln\frac{R(\widetilde{\Pi}_M)}{\delta\Delta}\right),$$

where we used the fact that $n \leq \ln t_{pol} + i_\star = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_M)}{\delta\Delta}\right)$. Hence, the total length of the first phase can be at most

$$\begin{aligned}
t_{pol} + i_\star \cdot k &= \mathcal{O}\left(\frac{i_\star MR(\widetilde{\Pi}_M)^4}{R(\widetilde{\Pi}_1)^2\Delta^2}\ln\frac{R(\widetilde{\Pi}_M)}{\Delta\delta} + \left(M^2 + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)^2}i_\star\right)\frac{R(\widetilde{\Pi}_{i_\star})^2 i_\star}{\Delta^2}\ln^2\frac{MR(\widetilde{\Pi}_{i_\star})}{\Delta\delta}\right)\\
&= \mathcal{O}\left(\left(\frac{R(\widetilde{\Pi}_M)^4}{R(\widetilde{\Pi}_1)^2} + MR(\widetilde{\Pi}_{i_\star})^2\right)\frac{Mi_\star}{\Delta^2}\ln^2\frac{MR(\widetilde{\Pi}_{i_\star})}{\Delta\delta}\right)\\
&= \mathcal{O}\left(\frac{M^2 i_\star R(\widetilde{\Pi}_M)^4}{R(\widetilde{\Pi}_1)^2\Delta^2}\ln^2\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right),
\end{aligned}$$

as claimed. □

Lemma 35 (Pseudo-Regret of the First Phase)**.** *Consider a run of Algorithm 2 with inputs $t_0 = 0, n = 1$, arbitrary policy $\widehat{\pi} \in \Pi_M$ and M base learners $\mathbb{A}_1, \ldots, \mathbb{A}_M$ with $1 \leq R(\widetilde{\Pi}_1) \leq R(\widetilde{\Pi}_2) \leq \cdots \leq R(\widetilde{\Pi}_M)$ where $\widetilde{\Pi}_i$ is the extended version of policy class Π_i with $(M+1-i)$ additional actions. Assume the environment $\mathbb{B}$ is stochastic and there is a policy $\pi_\star$ with gap $\Delta > 0$. Let t_{gap} be the round where* ArbeGap-Exploit *was called. Then with probability at least $1 - poly(M)\delta$, the pseudo-regret in all rounds $t \in [2i_\star, t_{gap}]$ is bounded as*

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\frac{R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2}\frac{M^2 i_\star}{\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right).$$

Proof. First, we bound the pseudo-regret by regret through a simple concentration argument in Lemma 42

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) \leq \mathrm{Reg}_{\mathbb{M}}(t, \Pi_M) + \mathcal{O}\left(\sqrt{t\ln\frac{\ln t}{\delta}}\right).$$

Next, we bound the regret by Lemma 30 which holds for any environment (adversarial but also stochastic). This yields

$$\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M) = \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\sqrt{t(\ln(t) + i_\star)\ln\frac{t}{\delta}}\right)$$

and finally, we use the bound on $t \leq t_{gap}$, the length of the first phase from Lemma 34 to replace $t(\ln(t) + i_\star)$ above which gives

$$\begin{aligned}
&\mathrm{PseudoReg}_{\mathbb{M}}(t, \Pi_M)\\
&= \mathcal{O}\left(\left(MR(\widetilde{\Pi}_{i_\star}) + \frac{R(\widetilde{\Pi}_{i_\star})^2}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\frac{R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)}\frac{M\sqrt{i_\star}}{\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\sqrt{(\ln(t) + i_\star)\ln\frac{t}{\delta}}\right)\\
&= \mathcal{O}\left(\left(M + \frac{R(\widetilde{\Pi}_{i_\star})}{R(\widetilde{\Pi}_1)}\sqrt{i_\star}\right)\frac{Mi_\star R(\widetilde{\Pi}_{i_\star})R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right)\\
&= \mathcal{O}\left(\frac{M^2 i_\star R(\widetilde{\Pi}_{i_\star})^2 R(\widetilde{\Pi}_M)^2}{R(\widetilde{\Pi}_1)^2\Delta}\ln\left(\frac{MR(\widetilde{\Pi}_M)}{\Delta\delta}\right)\ln\frac{t}{\delta}\right),
\end{aligned}$$

as claimed. □

D.3 Analysis of **Arbe-GapExploit**

In this section we prove the following result:

Lemma 36 (Guarantee for Second Phase)**.** *Let $\mathbb{A}$ be an h-stable learner with policy class $\Pi_{\mathbb{A}}$. Then the regret of Algorithm 4 against $\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}$ is bounded with probability at least $1 - \mathcal{O}(\delta)$ for all rounds $t > t_0$ that the algorithm has not terminated yet as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

Further, if the environment is stochastic with an optimal policy $\pi_\star$ that has a gap Δ compared to the best policy in $\Pi_{\mathbb{A}}$ and the inputs satisfy $\widehat{\pi} = \pi_\star$ and $\widehat{\Delta} \le \Delta \le 2\widehat{\Delta}$, then with probability at least $1 - \mathcal{O}(\delta)$ the pseudo-regret of Algorithm 4 is bounded in all rounds $t > t_0$ as

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\} \cup \Pi_{\mathbb{A}}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right)\right)$$

and the algorithm never terminates.

D.3.1 Guarantees for stochastic environments

Lemma 37 (Algorithm 4 does not terminate in stochastic environemnts)**.** *Assume the environment is stochastic and there is an optimal policy $\pi_\star$ with a gap Δ compared to the best policy in $\Pi_{\mathbb{A}}$. If Algorithm 4 is called with inputs $\widehat{\pi} = \pi_\star$ and $\widehat{\Delta} \le \Delta \le 2\widehat{\Delta}$, then with probability at least $1 - \mathcal{O}(\delta)$ it never terminates.*

Proof. Let e be an arbitrary epoch. By following the steps of Lemma 23, we can show that with probability at least $1 - \mathcal{O}(\delta_e)$ for all time steps in epoch e

$$\left|\widetilde{\mathrm{CRew}}_1(t_e+1,t) - \sum_{\ell=t_e+1}^{t} r_\ell(a_\ell^1, x_\ell)\right| \le \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right)$$

$$\left|\widetilde{\mathrm{CRew}}_0(t_e+1,t) - \sum_{\ell=t_e+1}^{t} r_\ell(a_\ell^0, x_\ell)\right| \le \mathcal{O}\left(\sqrt{\frac{t-t_e}{1-\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right)$$

and for $i \in \{0,1\}$

$$\left|\sum_{\ell=t_e+1}^{t} r_\ell(a_\ell^i, x_\ell) - \sum_{\ell=t_e+1}^{t} \mathbb{E}_{a\sim\pi_\ell^i}[r_\ell(a,x_\ell)]\right| \le \mathcal{O}\left(\sqrt{(t-t_e)\ln\frac{\ln(t-t_e)}{\delta_e}}\right).$$

Combining these bound together with Lemma 42, we have with $\pi_{\star,\mathbb{A}}$ being the best policy in $\Pi_{\mathbb{A}}$

$$\begin{aligned}
&\widetilde{\mathrm{CRew}}_0(t_e+1,t) - \widetilde{\mathrm{CRew}}_1(t_e+1,t) \\
&\le \sum_{\ell=t_e+1}^{t}\left[\mathbb{E}_{a\sim\pi_\ell^0,x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\ell^1,x\sim\mathcal{D}}[r(a,x)]\right] \\
&\quad + \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \\
&= \sum_{\ell=t_e+1}^{t}\left[\mathbb{E}_{a\sim\pi_\star,x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_{\star,\mathbb{A}},x\sim\mathcal{D}}[r(a,x)]\right] + \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}}\right) \\
&\quad + \sum_{\ell=t_e+1}^{t}\left[\mathbb{E}_{a\sim\pi_{\star,\mathbb{A}},x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\ell^1,x\sim\mathcal{D}}[r(a,x)]\right] + \mathcal{O}\left(\frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \\
&= \Delta(t-t_e) + \mathrm{PseudoReg}_{\mathbb{A}}([t_e+1,t],\Pi_{\mathbb{A}}) + \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right)
\end{aligned}$$

$$\leq \Delta(t-t_e) + \mathrm{Reg}_{\mathbb{A}}([t_e+1,t],\Pi_{\mathbb{A}}) + \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \quad \text{(Lemma 42)}$$

$$\leq \Delta(t-t_e) + \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{t-t_e}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \quad (\mathbb{A} \text{ is h-stable})$$

$$\leq 2\widehat{\Delta}(t-t_e) + \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{t-t_e}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \quad (\Delta \leq 2\widehat{\Delta})$$

where we used the fact that $\rho^e \leq 1/2 \leq 1-\rho^e$ by the choice of constants (see Lemma 39). This chain of inequalities holds with probability $1-\mathcal{O}(\delta_e)$. Combining this with a union bound, this implies that with this probability at least $1-\mathcal{O}(\delta)$, the test Line 16 never triggers. Similarly, we can lower-bound the same term as

$$\begin{aligned}
&\widetilde{\mathrm{CRew}}_0(t_e+1,t) - \widetilde{\mathrm{CRew}}_1(t_e+1,t) \\
&\geq \sum_{\ell=t_e+1}^{t}\left[\mathbb{E}_{a\sim\pi_\ell^0,x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\ell^1,x\sim\mathcal{D}}[r(a,x)]\right] \\
&\quad - \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \\
&\geq \sum_{\ell=t_e+1}^{t}\left[\mathbb{E}_{a\sim\pi_\star,x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_{\star,\mathbb{A}},x\sim\mathcal{D}}[r(a,x)]\right] \quad (\pi_\ell^0 = \pi_\star) \\
&\quad - \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \\
&\geq \Delta(t-t_e) - \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \\
&\geq \widehat{\Delta}(t-t_e) - \mathcal{O}\left(\sqrt{\frac{t-t_e}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(t-t_e)}{\delta_e}\right) \quad (\widehat{\Delta} \leq \Delta)
\end{aligned}$$

Hence, after combining these statements with a union bound, this implies that with this probability at least $1-\mathcal{O}(\delta)$, the test Line 14 never triggers. A final union bound for both tests completes the proof. □

Lemma 38 (Pseudo-regret of Algorithm 4)**.** *Assume the environment is stochastic and there is an optimal policy $\pi_\star$ with a gap Δ. If Algorithm 4 is called with inputs $\widehat{\pi} = \pi_\star$ and $\widehat{\Delta} = \Theta(\Delta)$, then the pseudo-regret of the algorithm is bounded with probability at least $1-\mathcal{O}(\delta)$ as*

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta} + \ln(t)\ln\frac{t}{\delta}\right)\right)$$

for all rounds $t \geq t_0$ where the algorithm has not terminated.

Proof. The pseudo-regret of Algorithm 4 can be decomposed into the regret in each epoch as

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) = \sum_{e=0}^{j(t)}\mathrm{PseudoReg}_{\mathbb{M}}([t_e+1,t'_e],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}})$$

where $t'_e = \min\{t_{e+1}, t\}$ and $j(t) = \min\{e \in \mathbb{N} \colon t_{e+1} \geq t\}$ is the epoch at time t. We consider the regret in each epoch separately as

$$
\begin{aligned}
&\mathrm{PseudoReg}_{\mathbb{M}}([t_e+1, t'_e], \{\widehat{\pi}\} \cup \Pi_{\mathbb{A}}) \\
&= \sum_{\ell=t_e+1}^{t'_e} \left(\mathbb{E}_{a\sim\pi^\star, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_t, x\sim\mathcal{D}}[r(a,x)]\right) \\
&= \sum_{\ell=t_e+1}^{t'_e} b_\ell \left(\mathbb{E}_{a\sim\pi^\star, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi^1_\ell, x\sim\mathcal{D}}[r(a,x)]\right) && (\widehat{\pi} = \pi^\star) \\
&\leq \sum_{\ell=t_e+1}^{t'_e} b_\ell \left(\Delta + \mathbb{E}_{a\sim\pi^\star_{\mathbb{A}}, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi^1_\ell, x\sim\mathcal{D}}[r(a,x)]\right)
\end{aligned}
$$

where $\pi^\star_{\mathbb{A}}$ is the best policy in $\Pi_{\mathbb{A}}$ which incurs a pseudo-regret of at most Δ per round. We now apply a concentration argument. Denote $Y_\ell = \mathbf{1}\{\ell > t_e\}\left(\Delta + \mathbb{E}_{a\sim\pi^\star_{\mathbb{A}}, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)]\right)$ and let $\mathcal{F}_\ell$ be the sigma-field that includes $\{\pi^1_t\}_{t>t_0}^{\ell}$, $\{b_t\}_{t>t_0}^{\ell-1}$ and t_e. Note that Y_ℓ is $\mathcal{F}_\ell$-measurable and $\sum_{\ell=t_e+1}^{t'_e}(b_\ell - \rho^e)Y_\ell$ is a martingale difference sequence w.r.t. $\mathcal{F}_\ell$. Since $|(b_\ell - \rho^e)Y_\ell| \leq 2$ and the sequence of conditional variance is bounded as $V_\ell = \sum_{\ell=t_e+1}^{t'_e} \rho^e(1-\rho^e)Y_\ell \leq \sum_{\ell=t_e+1}^{t'_e} \rho^e Y_\ell$, we can apply Lemma 47 and get with probability at least $1-\delta_e$ for all rounds $t_e \leq \tilde{t} \leq t_{e+1}$

$$
\begin{aligned}
\sum_{\ell=t_e+1}^{\tilde{t}} (b_\ell - \rho^e)Y_\ell &\leq \mathcal{O}\left(\sqrt{\sum_{\ell=t_e+1}^{\tilde{t}} \rho^e Y_\ell \ln \frac{\ln(\tilde{t}-t_e)}{\delta_e}} + \ln\frac{\ln(\tilde{t}-t_e)}{\delta_e}\right) \\
&\leq \sum_{\ell=t_e+1}^{\tilde{t}} \rho^e Y_\ell + \leq \mathcal{O}\left(\ln\frac{\ln(\tilde{t}-t_e)}{\delta_e}\right) && \text{(AM-GM inequality)}
\end{aligned}
$$

This holds in particular for $\tilde{t} = t'_e$ and with shorthand $k'_e = t'_e - t_e$, this gives

$$
\begin{aligned}
&\mathrm{PseudoReg}_{\mathbb{M}}([t_e+1, t'_e], \{\widehat{\pi}\} \cup \Pi_{\mathbb{A}}) \\
&\leq 2\rho^e \sum_{\ell=t_e+1}^{t'_e} \left(\Delta + \mathbb{E}_{a\sim\pi^\star_{\mathbb{A}}, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)]\right) + \mathcal{O}\left(\ln\frac{\ln(k'_e)}{\delta_e}\right) \\
&= 2\rho^e k'_e\Delta + \rho^e\,\mathrm{PseudoReg}_{\mathbb{A}}([t_e+1, t'_e], \Pi_{\mathbb{A}}) + \mathcal{O}\left(\ln\frac{\ln(k'_e)}{\delta_e}\right) \\
&\leq 2\rho^e k'_e\Delta + \rho^e\,\mathrm{Reg}_{\mathbb{A}}([t_e+1, t'_e], \Pi_{\mathbb{A}}) + \mathcal{O}\left(\rho^e\sqrt{k'_e \ln\frac{\ln(k'_e)}{\delta_e}} + \ln\frac{\ln(k'_e)}{\delta_e}\right) && \text{(Lemma 42)} \\
&\leq 2\rho^e k'_e\Delta + \mathcal{O}\left(\rho^e R(\Pi_{\mathbb{A}})\sqrt{\frac{k'_e}{\rho^e}\ln\frac{k'_e}{\delta_e}} + \ln\frac{\ln(k'_e)}{\delta_e}\right) && (\mathbb{A} \text{ is h-stable}) \\
&\leq 2\rho^e k'_e\Delta + \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\rho^e k'_e \ln\frac{k'_e}{\delta_e}} + \ln\frac{\ln(k'_e)}{\delta_e}\right) \\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2\Delta}{\widehat{\Delta}^2}\ln\frac{k_e}{\delta_e} + R(\Pi_{\mathbb{A}})\sqrt{\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}^2}\ln\left(\frac{k_e}{\delta_e}\right)\ln\frac{k'_e}{\delta_e}} + \ln\frac{\ln(k'_e)}{\delta_e}\right) \\
& && \text{(definition of } \rho^e \text{ and } k'_e \leq k_e) \\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln\frac{k_e}{\delta_e}\right) && (\widehat{\Delta} = \Theta(\Delta))
\end{aligned}
$$

Now, we can combine this pseudo-regret bound across all epochs and get

$$\begin{aligned}\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) &\le \sum_{e=0}^{j(t)}\mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln\frac{k_e}{\delta_e}\right)\\ &\le \mathcal{O}\left(j(t)\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln\frac{tj(t)}{\delta}\right)\\ &= \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln(t)\ln\frac{t}{\delta}\right) \qquad (j(t)=\mathcal{O}(\ln(t)))\end{aligned}$$

when $j(t)\ge 1$ since $k_e\le k_{j(t)}\le 2t$ in this case. In the other case, where $j(t)=0$, we have

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) \le \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln\frac{k_0}{\delta_0}\right) \le \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right).$$

Hence, combining both cases gives the final bound

$$\mathrm{PseudoReg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}\cup\Pi_{\mathbb{A}}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\Delta}\left(\ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}+\ln(t)\ln\frac{t}{\delta}\right)\right).$$

□

D.3.2 Exploitation Subroutine Guarantees for Adversarial Environments

Lemma 39. *Assume the absolute constant in the length of the initial epoch k_0 of Algorithm 4 is chosen large enough. Then the regret of Algorithm 4 against $\Pi_{\mathbb{A}}$ is bounded with probability at least $1-\mathcal{O}(\delta)$ for all rounds t that the algorithm has not terminate yet as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}}) = \mathcal{O}(\widehat{\Delta}k_0).$$

Proof. Consider any round $t>t_0$ before the test in Line 14 or 16 triggers. The total reward in relevant rounds can be decomposed into epochs $e=0,\dots,j(t)$ as follows where $j(t)=\max\{e\in\mathbb{N}\colon t_e<t\}$ is the epoch of round t and $t'_e=\min\{t_{e+1},t\}$.

$$\sum_{\ell=t_0+1}^{t} r_\ell(a_\ell,x_\ell) = \sum_{e=0}^{j(t)}\sum_{\ell=t_e+1}^{t'_e} r_\ell(a_\ell,x_\ell)$$

Further, let $k'_e=t'_e-t_e$ be the number of rounds in the e-th epoch and consider now a single epoch e. We can write

$$\begin{aligned}&\sum_{\ell=t_e+1}^{t'_e} r_\ell(a_\ell,x_\ell)\\ &= \rho^e\widetilde{\mathrm{CRew}}_1(t_e+1,t'_e)+(1-\rho^e)\widetilde{\mathrm{CRew}}_0(t_e+1,t'_e)\\ &\ge \rho^e\widetilde{\mathrm{CRew}}_1(t_e+1,t'_e)+(1-\rho^e)\left(\widetilde{\mathrm{CRew}}_1(t_e+1,t'_e)+\widehat{\Delta}k'_e-V(t'_e)k'_e\right)\\ &\qquad\qquad\text{(first test did not trigger)}\\ &\ge \widetilde{\mathrm{CRew}}_1(t_e+1,t'_e)+(1-\rho^e)\left(\widehat{\Delta}k'_e-V(t'_e)k'_e\right)\\ &\ge \sum_{\ell=t_e+1}^{t'_e} r_\ell(a^1_\ell,x_\ell)+(1-\rho^e)\left(\widehat{\Delta}k'_e-V(t'_e)k'_e\right)-\mathcal{O}\left(\sqrt{\frac{k'_e}{\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}}+\frac{1}{\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}\right)\\ &\qquad\qquad\text{(see proof of Lemma 37)}\\ &\ge \max_{\pi\in\Pi_{\mathbb{A}}}\sum_{\ell=t_e+1}^{t'_e}\mathbb{E}_{a\sim\pi}[r_\ell(a,x_\ell)]-\mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{k'_e}{\rho^e}\ln\frac{k'_e}{\delta_e}}+\frac{1}{\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}\right)\\ &\quad+(1-\rho^e)\left(\widehat{\Delta}k'_e-V(t'_e)k'_e\right) \qquad\qquad (\mathbb{A}\text{ is h-stable})\end{aligned}$$

$$\geq \max_{\pi \in \Pi_{\mathbb{A}}} \sum_{\ell=t_e+1}^{t'_e} \mathbb{E}_{a\sim\pi}[r_\ell(a, x_\ell)] + (1-\rho^e)\widehat{\Delta}k'_e - \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{k'_e}{\rho^e}\ln\frac{k'_e}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}\right)$$

(definition of $V(t'_e)$)

Hence, by rearranging terms, we get that the regret in a single epoch is bounded as

$$\mathrm{Reg}_{\mathbb{M}}([t_e+1, t'_e], \Pi_{\mathbb{A}}) \leq (\rho^e - 1)\widehat{\Delta}k'_e + \mathcal{O}\left(R(\Pi_{\mathbb{A}})\sqrt{\frac{k'_e}{\rho^e}\ln\frac{k'_e}{\delta_e}} + \frac{1}{\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}\right).$$

To further upper-bound these terms, we first derive useful expression for the inverse probability of playing $\mathbb{A}$. Here, we make the constants in the definition of the initial epoch length k_0 explicit. Specifically, we assume that $k_0 = \frac{c_0 R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}^2}\ln\frac{2c_0 R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}$ where c_0 is the absolute constant. We have

$$\frac{1}{\rho^e} = \frac{k_e\widehat{\Delta}^2}{c_\rho R(\Pi_{\mathbb{A}})^2 \ln\frac{k_e}{\delta_e}}$$

Plugging this bound on the inverse probability back into the expression for the regret per epoch above, we get with c as the constant in the $\mathcal{O}$ notation

$$\begin{aligned}
\mathrm{Reg}_{\mathbb{M}}([t_e+1, t'_e], \Pi_{\mathbb{A}}) &\leq (\rho^e - 1)\widehat{\Delta}k'_e + \widehat{\Delta}\sqrt{\frac{6ck'_e k_e}{c_\rho}} + \frac{6c\widehat{\Delta}^2 k_e}{c_0 R(\Pi_{\mathbb{A}})^2} \\
&= \widehat{\Delta}k'_e\left(-1 + \rho^e + \sqrt{\frac{6ck_e}{c_\rho k'_e}} + \frac{6c\widehat{\Delta}k_e}{k'_e c_\rho R(\Pi_{\mathbb{A}})^2}\right) \\
&\leq \frac{\widehat{\Delta}}{2}\left(-k'_e + \sqrt{\frac{12c}{c_\rho}k_e k'_e} + \frac{12c}{c_\rho}\frac{\widehat{\Delta}}{R(\Pi_{\mathbb{A}})}k_e\right). \qquad (\rho^e \leq 1/2)
\end{aligned}$$

In the last step, we used $\rho^e \leq \rho^0 \leq 1/2$ which we can establish by choosing the constants c_ρ and c_0 in the definition of ρ and k_0 appropriately. Specifically,

$$\rho^e \leq \rho^0 = c_\rho\frac{R(\Pi_{\mathbb{A}})^2\ln\frac{k_0}{\delta_0}}{\widehat{\Delta}^2 k_0} = \frac{c_\rho}{c_0}\frac{\ln\frac{k_0}{\delta_0}}{\ln\frac{R(\Pi_{\mathbb{A}})}{\delta\widehat{\Delta}}} \leq \frac{c_\rho}{c_0}\frac{\ln\frac{c_0 R(\Pi_{\mathbb{A}})^3}{\delta^2\widehat{\Delta}^3}}{\ln\frac{c_0 R(\Pi_{\mathbb{A}})}{\delta\widehat{\Delta}}} \leq 3\frac{c_\rho}{c_0}.$$

Thus, choosing $c_0 \geq 6c_\rho$ is sufficient to ensure $\rho^e \leq 1/2$. Summing now the bound above over epochs gives

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_{\mathbb{A}}) \leq \frac{\widehat{\Delta}}{2}\left(-t + t_0 + \sqrt{\frac{12c}{c_\rho}}\sum_{e=0}^{j(t)}\sqrt{k_e k'_e} + \frac{12c}{c_\rho}\frac{\widehat{\Delta}}{R(\Pi_{\mathbb{A}})}\sum_{e=0}^{j(t)}k_e\right).$$

We now distinguish between two cases. First consider the case where $j(t) \geq 1$ and assume that $c_\rho \geq 9 \cdot 12c$. Since in this case, the number of rounds $t - t_0$ is at least half the sum of epoch lengths, $\sum_{e=0}^{j(t)} k_e$, we have

$$\begin{aligned}
\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_{\mathbb{A}}) &\leq \frac{\widehat{\Delta}}{2}\left(-t + t_0 + \frac{1}{3}\sum_{e=0}^{j(t)}\sqrt{k_e k'_e} + \frac{1}{9}\sum_{e=0}^{j(t)}k_e\right) \\
&\leq \frac{\widehat{\Delta}}{2}\left(-t + t_0 + \frac{4}{9}\sum_{e=0}^{j(t)}k_e\right) \leq \frac{\widehat{\Delta}}{2}\left(-t + t_0 + \frac{8}{9}(t - t_0)\right) \leq 0\,.
\end{aligned}$$

In the other case, when $j(t) = 0$, we have

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_{\mathbb{A}}) \leq \frac{\widehat{\Delta}}{2}\left(-k'_0 + \sqrt{\frac{12c}{c_\rho}}\sqrt{k_0 k'_0} + \frac{12c}{c_\rho}\frac{\widehat{\Delta}}{R(\Pi_{\mathbb{A}})}k_0\right) = \mathcal{O}(\widehat{\Delta}k_0).$$

Hence, the regret against $\Pi_{\mathbb{A}}$ is always bounded as

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \Pi_{\mathbb{A}}) = \mathcal{O}(\widehat{\Delta}k_0).$$

□

Lemma 40. *The regret of Algorithm 4 against $\widehat{\pi}$ is bounded with probability at least $1 - \mathcal{O}(\delta)$ for all rounds t that the algorithm has not terminated yet as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1, t], \{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\Delta\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

Proof. Consider any round $t > t_0$ before the test in Line 14 or 16 triggers. The total reward in relevant rounds can be decomposed into epochs $e = 0, \dots, j(t)$ as follows where $j(t) = \max\{e \in \mathbb{N} \colon t_e < t\}$ is the epoch of round t and $t'_e = \min\{t_{e+1}, t\}$.

$$\sum_{\ell=t_0+1}^{t} r_\ell(a_\ell, x_\ell) = \sum_{e=0}^{j(t)} \sum_{\ell=t_e+1}^{t'_e} r_\ell(a_\ell, x_\ell)$$

Further, let $k'_e = t'_e - t_e$ be the number of rounds in the e-th epoch and consider now a single epoch e. We can write

$$\begin{aligned}
&\sum_{\ell=t_e+1}^{t'_e} r_\ell(a_\ell, x_\ell)\\
&= (1-\rho^e)\widetilde{\mathrm{CRew}}_0(t_e+1, t'_e) + \rho^e\widetilde{\mathrm{CRew}}_1(t_e+1, t'_e)\\
&\geq (1-\rho^e)\widetilde{\mathrm{CRew}}_0(t_e+1, t'_e) + \rho^e\left(\widetilde{\mathrm{CRew}}_0(t_e+1, t'_e) - 4\widehat{\Delta}k'_e - V(t'_e)k'_e\right) && \text{(second test did not trigger)}\\
&\geq \widetilde{\mathrm{CRew}}_0(t_e+1, t'_e) - \rho^e\left(4\widehat{\Delta}k'_e + V(t'_e)k'_e\right)\\
&\geq \sum_{\ell=t_e+1}^{t'_e} r_\ell(a^1_\ell, x_\ell) - \rho^e\left(4\widehat{\Delta}k'_e + V(t'_e)k'_e\right) - \mathcal{O}\left(\sqrt{\frac{k'_e}{1-\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln k'_e}{\delta_e}\right) && \text{(see proof of Lemma 37)}\\
&\geq \sum_{\ell=t_e+1}^{t'_e} \mathbb{E}_{a\sim\widehat{\pi}}[r_\ell(a, x_\ell)] - \rho^e\left(4\widehat{\Delta}k'_e + V(t'_e)k'_e\right) - \mathcal{O}\left(\sqrt{\frac{k'_e}{1-\rho^e}\ln\frac{\ln(k'_e)}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln k'_e}{\delta_e}\right).
\end{aligned}$$

The last step here follows since $\rho^e k'_e V(t'_e) \leq \mathcal{O}(R(\Pi_{\mathbb{A}})\sqrt{\rho^e k'_e \ln\frac{k'_e}{\delta_e}} + \ln\frac{\ln k'_e}{\delta_e})$ By rearranging terms, we can bound the regret against $\widehat{\pi}$ in epoch e as

$$\begin{aligned}
&\mathrm{Reg}_{\mathbb{M}}([t_e+1, t'_e], \{\widehat{\pi}\})\\
&\leq \mathcal{O}\left(\rho^e k'_e(\widehat{\Delta} + V(t'_e)) + \sqrt{\frac{k'_e}{1-\rho^e}\ln\frac{k'_e}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln k'_e}{\delta_e}\right)\\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\ln\frac{k_e}{\delta_e} + \sqrt{\frac{k'_e}{1-\rho^e}\ln\frac{k'_e}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln k'_e}{\delta_e}\right) && \text{(definition of } \rho^e, V(t'_e) \text{ and } k_e \geq k'_e)\\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\ln\frac{k_e}{\delta_e} + \sqrt{\frac{k'_e}{1-\rho^e}\ln\frac{k'_e}{\delta_e}} + \frac{1}{1-\rho^e}\ln\frac{\ln k'_e}{\delta_e}\right)\\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\ln\frac{k_e}{\delta_e} + \sqrt{k'_e\ln\frac{k'_e}{\delta_e}} + \ln\frac{\ln k'_e}{\delta_e}\right) && (\rho^e \leq 1/2)\\
&\leq \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\ln\frac{k_e}{\delta_e} + \sqrt{k'_e\ln\frac{k'_e}{\delta_e}}\right).
\end{aligned}$$

Note that $\rho^e \leq 1/2$ holds for appropriate constants in the definition of k_0 and ρ^e (see Lemma 39). We can now sum the regret over all epochs e and get

$$\begin{aligned}
&\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}) \leq \sum_{e=0}^{j(t)} \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln \frac{k_e}{\delta_e} + \sqrt{k'_e \ln \frac{k'_e}{\delta_e}}\right) \\
&\leq \sum_{e=0}^{j(t)} \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln \frac{(j(t)+1)k_{j(t)}}{\delta} + \sqrt{k'_e \ln \frac{(j(t)+1)k_{j(t)}}{\delta}}\right) \\
&\leq \mathcal{O}\left((j(t)+1)\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln \frac{(j(t)+1)k_{j(t)}}{\delta} + \sqrt{(j(t)+1)(t-t_0) \ln \frac{(j(t)+1)k_{j(t)}}{\delta}}\right).
\end{aligned}$$

We now distinguish between two cases. First, $j(t) = 0$, in which case

$$\begin{aligned}
\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}) &= \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln \frac{R(\Pi_{\mathbb{A}})}{\Delta\delta} + \sqrt{k_0 \ln \frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}}\right) \\
&= \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln \frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right)
\end{aligned}$$

and the case where $j(t) > 0$, where $j(t) = \mathcal{O}(\ln(t-t_0)) = \mathcal{O}(\ln(t))$ and $k_{j(t)} \leq 2(t-t_0)$

$$\begin{aligned}
\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}) &= \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln(t-t_0) \ln \frac{t-t_0}{\delta} + \sqrt{(t-t_0)\ln(t-t_0) \ln \frac{t-t_0}{\delta}}\right) \\
&= \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}} \ln(t) \ln \frac{t}{\delta} + \sqrt{(t-t_0)\ln(t) \ln \frac{t}{\delta}}\right).
\end{aligned}$$

Combining both cases gives

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

□

Lemma 41 (Regret in adversarial environments)**.** *Assume the absolute constant in the length of the initial epoch k_0 and sampling probabilities ρ^e of Algorithm 4 is chosen large enough. Then the regret of Algorithm 4 against $\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}$ is bounded with probability at least $1 - \mathcal{O}(\delta)$ for all rounds t that the algorithm has not terminated yet as*

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

Proof. By combining Lemma 39 and Lemma 40, we have

$$\begin{aligned}
&\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}) \\
&\quad = \mathcal{O}\left(\widehat{\Delta}k_0 + \frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right)
\end{aligned}$$

and by plugging in the definition of k_0, we get

$$\mathrm{Reg}_{\mathbb{M}}([t_0+1,t],\Pi_{\mathbb{A}} \cup \{\widehat{\pi}\}) = \mathcal{O}\left(\frac{R(\Pi_{\mathbb{A}})^2}{\widehat{\Delta}}\left(\ln(t)\ln\frac{t}{\delta} + \ln\frac{R(\Pi_{\mathbb{A}})}{\widehat{\Delta}\delta}\right) + \sqrt{(t-t_0)\ln(t)\ln\frac{t}{\delta}}\right).$$

□

D.4 Concentration Bounds in Stochastic Environments

Lemma 42. *In stochastic environments, the regret and pseudo-regret of any algorithm $\mathbb{A}$ against a policy class Π' satisfy with probability at least $1-\delta$ for all rounds $t \in \mathbb{N}$*

$$\mathrm{PseudoReg}_{\mathbb{A}}(t, \Pi') - \mathrm{Reg}_{\mathbb{A}}(t, \Pi') = \mathcal{O}\left(\sqrt{t \ln \frac{\ln t}{\delta}}\right) . \tag{33}$$

Proof. Let $\pi_\star \in \mathrm{argmax}_{\pi \in \Pi'} \mathbb{E}_{a\sim\pi, x\sim\mathcal{D}}[r(a,x)]$ be the best policy in Π'. Then

$$\begin{aligned}
&\mathrm{PseudoReg}_{\mathbb{A}}(t, \Pi') - \mathrm{Reg}_{\mathbb{A}}(t, \Pi') \\
&= \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\star, x\sim\mathcal{D}}[r(a,x)] - \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)] \\
&\quad - \left(\max_{\pi'\in\Pi'} \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] - \sum_{\ell=1}^{t} r_\ell(a_\ell, x_\ell)\right) \\
&= \sum_{\ell=1}^{t} r_\ell(a_\ell, x_\ell) - \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)] \\
&\quad + \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\star, x\sim\mathcal{D}}[r(a,x)] - \max_{\pi'\in\Pi'} \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi'}[r_\ell(a, x_\ell)] \\
&\le \sum_{\ell=1}^{t} r_\ell(a_\ell, x_\ell) - \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)] \\
&\quad + \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\star, x\sim\mathcal{D}}[r(a,x)] - \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\star}[r_\ell(a, x_\ell)] && (\pi_\star \in \Pi') \\
&\le 2 \times 1.44 \sqrt{\max(2t, 2)\left(1.4 \ln\ln\left(2\left(\max\left(\frac{2t}{2}, 1\right)\right)\right) + \ln\frac{5.2}{\delta}\right)} && \text{(Lemma 46)} \\
&= \mathcal{O}\left(\sqrt{t \ln \frac{\ln t}{\delta}}\right) .
\end{aligned}$$

Here, the last main step is to apply the time-uniform Hoeffding bound from Lemma 46 to the the two differences individually. Both are martingale sequences that are bounded, i.e., $r_\ell(a_\ell, x_\ell) - \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)] \in [-1, 1]$ and $\mathbb{E}_{a\sim\pi_\star, x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\star}[r_\ell(a, x_\ell)] \in [-1, 1]$. □

Lemma 43. *Let $t_0 \in \mathbb{N}$ be a possibly random time and let $\{\pi_\ell\}_{\ell \ge t_0}$ be a possibly random sequence of policies $\pi_\ell \colon \mathcal{X} \to \Delta_{\mathcal{A}}$ so that for all $\ell \ge t_0$, π_ℓ is independent of $\{(r_j, x_j)\}_{j \ge \ell}$ and t_0 is independent of $\{(r_j, x_j)\}_{j \ge t_0}$. Then with probability at least $1 - 2\delta$*

$$\left|\sum_{\ell=t_0+1}^{t} \left[\mathbb{E}_{a\sim\pi_\ell}[r_\ell(a, x_\ell)] - \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)]\right]\right| = \mathcal{O}\left(\sqrt{(t - t_0) \ln \frac{\ln(t - t_0)}{\delta}}\right) .$$

Proof. Define the sigma-algebra $\mathcal{F}_\ell = \sigma(\{r_j, x_j, \pi_j\}_{j \le \ell} \cup \{\pi_\ell, t_0\})$ and let

$$Z_\ell = \begin{cases} 0 & \text{if } \ell \le t_0 \\ \mathbb{E}_{a\sim\pi_\ell}[r_\ell(a, x_\ell)] - \mathbb{E}_{a\sim\pi_\ell, x\sim\mathcal{D}}[r(a,x)] & \text{otherwise.} \end{cases}$$

The sequence $\{Z_\ell\}_{\ell\in\mathbb{N}}$ is a martingale difference sequence w.r.t. $\mathcal{F}_\ell$ and $Z_\ell \in [-\mathbf{1}\{\ell > t_0\}, +\mathbf{1}\{\ell > t_0\}]$. As a result, we can apply Lemma 46 and get that

$$\begin{aligned}\sum_{\ell=t_0+1}^{t} [\mathbb{E}_{a\sim\pi_\ell}[r_\ell(a,x_\ell)] - \mathbb{E}_{a\sim\pi_\ell,x\sim\mathcal{D}}[r(a,x)]] &= \sum_{\ell=1}^{t} Z_t \\ &\leq 1.44\sqrt{\max(2(t-t_0),2)\left(1.4\ln\ln\left(2\left(\max\left(\frac{2(t-t_0)}{2},1\right)\right)\right)+\ln\frac{5.2}{\delta}\right)} \\ &= \mathcal{O}\left(\sqrt{(t-t_0)\ln\frac{\ln(t-t_0)}{\delta}}\right)\end{aligned}$$

holds for all t with probability at least $1-\delta$. Applying the same argument to $-Z_\ell$ and a union bound completes the proof. □

Notice that, because contexts are i.i.d., a simple anytime Hoeffding bound implies the random variable $\mathrm{MaxRew}(t)$ (recall this quantity is defined as the maximum sum of pseudo-expectations over realized contexts) is larger than $t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)]$ (up to a factor of $\widetilde{\mathcal{O}}(\sqrt{t})$).

Lemma 44. *If the environment $\mathbb{B}$ is stochastic then with probability at least $1-\delta$*

$$t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)] \leq \max_{\pi\in\Pi}\sum_{\ell=1}^{t}\mathbb{E}_{a\sim\pi}[r_\ell(a,x_\ell)] + \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right)$$

for all $t\in\mathbb{N}$.

Proof. Consider the martingale sequence $\{Z_\ell\}_{\ell=1}^{\infty}$ defined as $Z_\ell = \mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)] - \mathbb{E}_{a\sim\pi_\star(\cdot|x_t)}[r(a,x)|x_t]$. By definition $|Z_\ell|\leq 2$ for all ℓ. Therefore by an anytime Hoeffding bound (Lemma 46 in Appendix E), with probability at least $1-\delta$,

$$\sum_{\ell=1}^{t} Z_\ell = \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right)$$

for all $t\in\mathbb{N}$. Thus,

$$\begin{aligned}t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)] &\leq \sum_{\ell=1}^{t}\mathbb{E}_{a\sim\pi_\star}[r_\ell(a,x_\ell)] + \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right) \\ &\leq \max_{\pi\in\Pi}\sum_{\ell=1}^{t}\mathbb{E}_{a\sim\pi}[r_\ell(a,x_\ell)] + \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right)\end{aligned}$$

□

Lemma 44 implies that an algorithm that an algorithm that competes with $\mathrm{MaxRew}(t)$ in turn can compete against $t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)]$. This fact will help us argue that an adversarial algorithm has good performance in a stochastic environment.

Let's show that when the contexts are i.i.d., the counterfactual reward of any algorithm that decides what policy to play before the context is revealed is (up to a $\sqrt{t\ln\frac{t}{\delta}}$ factor) upper bounded by $t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)]$,

Lemma 45. *If the environment is stochastic with i.i.d. contexts and the algorithm decides on its policy π_t before observing context x_t then with probability at least $1-\delta$,*

$$\sum_{\ell=1}^{t}\mathbb{E}_{a\sim\pi_\ell(\cdot|x_\ell)}[r(a,x_\ell)|x_\ell] \leq t\mathbb{E}_{a\sim\pi_\star(\cdot|x),x\sim\mathcal{D}}[r(a,x)] + \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right)$$

for all $t\in\mathbb{N}$.

Proof. Similar to the proof of Lemma 44, consider the martingale sequence $\{Z_\ell\}_{\ell=1}^{\infty}$ defined as $Z_\ell = \mathbb{E}_{a\sim\pi_\ell(\cdot|x_\ell)}\left[r(a,x_\ell)|x_\ell\right] - \mathbb{E}_{a\sim\pi_\ell(\cdot|x),x\sim\mathcal{D}}\left[r(a,x)\right]$. By definition $|Z_\ell| \leq 2$. Therefore by the anytime Hoeffding bound of Lemma 46, with probability at least $1-\delta$

$$\sum_{\ell=1}^{t} Z_\ell = \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\ell(\cdot|x_\ell)}\left[r(a,x_\ell)|x_\ell\right] - \sum_{\ell=1}^{t} \mathbb{E}_{a\sim\pi_\ell(\cdot|x),x\sim\mathcal{D}}\left[r(a,x)\right] = \mathcal{O}\left(\sqrt{t\ln\frac{t}{\delta}}\right)$$

for all $t \in \mathbb{N}$. By definition $\pi_\star$ is the policy satisfying $\pi_\star = \operatorname{argmax}_\pi \mathbb{E}_{a\sim\pi,x\sim\mathcal{D}}\left[r(a,x)\right]$ and therefore for any π_ℓ, the inequality $\mathbb{E}_{a\sim\pi_\ell,x\sim\mathcal{D}}\left[r(a,x)\right] \leq \mathbb{E}_{a\sim\pi_\star,x\sim\mathcal{D}}\left[r(a,x)\right]$ holds. □

Lemma 45 implies that in the case of i.i.d. contexts a learner that selects a policy based on historical data (that is the learner selects a policy to play before the context is revealed) cannot do substantially better than playing $\pi_\star = \operatorname{argmax}_{\pi\in\Pi} \mathbb{E}_{a\sim\pi(\cdot|x),x\sim\mathcal{D}}\left[r(a,x)\right]$ during all timesteps. This will prove helpful when deriving bounds for the gap estimation phase.

E Additional Technical Lemmas

Lemma 46 (Time-uniform Hoeffding bound)**.** *Let $S_t = \sum_{i=1}^{t} Y_t$ be a martingale sequence w.r.t. some sigma algebra $\mathcal{F}_t$ and let $Y_t \in [a_t, b_t]$ almost surely for a_t, b_t measurable in $\mathcal{F}_t$. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

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

where $W_t = \sum_{i=1}^{t} \frac{(b_i-a_i)^2}{4}$ and $m > 0$ arbitrary but fixed.

Proof. By entry "Hoeffding I" in Table 3 of Howard et al. [18], S_t is a sub-ψ_N process with variance process W_t. Further, by Proposition 2 in Howard et al. [19], this implies that it is also a sub-ψ_P process with $c = 0$. We now apply Lemma 48 to achieve the desired result. □

Lemma 47 (Time-uniform Bernstein bound)**.** *Let $S_t = \sum_{i=1}^{t} Y_i$ be a martingale sequence w.r.t. a sigma algebra $\mathcal{F}_t$ and let $Y_t \leq c$ a.s. for some parameter $c > 0$. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\begin{aligned} S_t \leq\ & 1.44\sqrt{\max(W_t, m)\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \ln\frac{5.2}{\delta}\right)} \\ & + 0.41c\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \ln\frac{5.2}{\delta}\right) \end{aligned}$$

where $W_t = \sum_{i=1}^{t} \mathbb{E}[Y_i^2|\mathcal{F}_i]$ and $m > 0$ is arbitrary but fixed.

Proof. By entry "Bennett" in Table 3 of Howard et al. [18], S_t is a sub-ψ_P process with variance process W_t and parameter $c = 0$. We now apply Lemma 48 to achieve the desired result. □

Lemma 48 (General concentration result)**.** *In the terminology of Howard et al. [18], let $S_t = \sum_{i=1}^{t} Y_i$ be a sub-ψ_P process with parameter $c = 0$ and variance process W_t. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\begin{aligned} S_t \leq\ & 1.44\sqrt{\max(W_t, m)\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \ln\frac{5.2}{\delta}\right)} \\ & + 0.41c\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \ln\frac{5.2}{\delta}\right) \end{aligned}$$

where $m > 0$ is arbitrary but fixed.

Proof. This bound follows from Howard et al. [18] Theorem 1 with Equation 10 in that paper by setting $s = 1.4$ and $\eta = 2$. □

Lemma 49. *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \le b$. Let*

$$\mathrm{Var}_\ell(X_\ell) = \mathbf{Var}(X_\ell | X_1, \cdots, X_{\ell-1})$$

Let $V_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$ be the sum of conditional variances of X_t. Then we have that for any $\delta' \in (0,1)$ and $t \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 2\sqrt{V_t}A_t + 3bA_t^2\right) \le \delta' ,$$

where $A_t = \sqrt{2\ln\ln\left(2\left(\max\left(\frac{V_t}{b^2}, 1\right)\right)\right) + \ln\frac{6}{\delta'}}$.

Proof. We are in a position to use 47 (with $c = b$). Let $S_t = \sum_{\ell=1}^{t} X_t$ and $W_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$. Let's set $m = b^2$. It follows that with probability $1 - \delta'$ for all $t \in \mathbb{N}$

$$\begin{aligned}
S_t &\le 1.44\sqrt{\max(W_t, b^2)\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{b^2}, 1\right)\right)\right) + \ln\frac{5.2}{\delta'}\right)} \\
&\quad + 0.41b\left(1.4\ln\ln\left(2\left(\max\left(\frac{W_t}{b}, 1\right)\right)\right) + \ln\frac{5.2}{\delta'}\right) \\
&\le 2\sqrt{\max(W_t, b^2)\left(2\ln\ln\left(2\left(\max\left(\frac{W_t}{b^2}, 1\right)\right)\right) + \ln\frac{6}{\delta'}\right)} \\
&\quad + b\left(2\ln\ln\left(2\left(\max\left(\frac{W_t}{b^2}, 1\right)\right)\right) + \ln\frac{6}{\delta'}\right) \\
&= 2\max(\sqrt{W_t}, b)A_t + bA_t^2 \\
&\le 2\sqrt{W_t}A_t + 2bA_t + bA_t^2 \\
&\overset{(i)}{\le} 2\sqrt{W_t}A_t + 3bA_t^2 ,
\end{aligned}$$

where $A_t = \sqrt{2\ln\ln\left(2\left(\max\left(\frac{W_t}{b^2}, 1\right)\right)\right) + \ln\frac{6}{\delta'}}$. Inequality (i) follows because $A_t \ge 1$.

Setting $V_t = W_t$ we conclude the proof. □

In turn, a corollary of the previous lemma is the following.

Lemma 50. *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \le b_t$ with b_t a non-decreasing deterministic sequence. Let*

$$\mathrm{Var}_\ell(X_\ell) = \mathbf{Var}(X_\ell | X_1, \cdots, X_{\ell-1})$$

Let $V_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$ be the sum of conditional variances of X_t. Then for any $\delta \in (0,1)$ and $t \in \mathbb{N}$ we have

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 2\sqrt{V_t}A_t + 3B_tA_t^2\right) \le \delta ,$$

where $A_t(\delta) = 2\sqrt{\ln\frac{12t^2}{\delta}}$.

Proof. For any t define a martingale difference sequence $\{X_\ell^{(t)}\}_{\ell=1}^{\infty}$ as follows:

$$X_\ell^{(t)} = \begin{cases} X_\ell & \text{if } \ell \le t \\ 0 & \text{otherwise} \end{cases}$$

We apply Lemma 49 with parameter $\delta' = \frac{\delta}{2t^2}$ and $b = B_t$, and then overapproximate. A union bound over $t \in \mathbb{N}$ gets the desired result. □

Lemma 51. *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \le b_t$ with b_t a non-decreasing deterministic sequence. Let*

$$\mathrm{Var}_\ell(X_\ell) = \mathbf{Var}(X_\ell | X_1, \cdots, X_{\ell-1})$$

Let $V_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$ be the sum of conditional variances of X_t. Then we have that for any $\delta \in (0,1)$ and $t \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 4\sqrt{V_t \ln \frac{12t^2}{\delta}} + 12 b_t \ln \frac{12t^2}{\delta}\right) \le \delta \ ,$$

Proof. Notice that by definition $V_t(\mathbf{a}) \le t b_t^2$. Therefore $\max(\frac{V_t(\mathbf{a})}{b_t^2}, 1) \le t$ and

$$A_t(\delta') \le \sqrt{2 \ln \ln 2t + \ln \frac{12t^2}{\delta'}} \le 2\sqrt{\ln \frac{12t^2}{\delta'}} := \tilde{A}_t(\delta') \ .$$

Substituting this upper bound in the statement of Lemma 50 yields the result. □

Lemma 52. *Let $c, \delta \in (0,1]$ and $t \ge \frac{16}{c^2} \ln^2 \frac{2}{c\delta}$. Then*

$$\frac{\ln \frac{t}{\delta}}{\sqrt{t}} \le c$$

Proof. The function $f(t) = \frac{\ln \frac{t}{\delta}}{\sqrt{t}}$ is non-increasing on $(e^2\delta, \infty)$. Thus, it is sufficient to show that the inequality holds for $t = 16\frac{\ln^2 \frac{2}{c\delta}}{c^2}$. Since $\ln(x) \le \frac{x}{2}$ for all $x \in \mathbb{R}_+$, we can upper-bound this value of t as

$$t \le \frac{16}{c^4\delta^2}.$$

This implies

$$\ln \frac{t}{\delta} \le \ln \frac{16}{c^4\delta^4} = 4 \ln \frac{2}{c\delta} = c\sqrt{16\frac{\ln^2(2/c/\delta)}{c^2}} = c\sqrt{t}$$

which proves the claim. □

Lemma 53. *Let $c, \delta \in (0,1]$ and $t \ge \frac{4}{c} \ln \frac{2}{c\delta}$. Then*

$$\frac{\ln \frac{t}{\delta}}{t} \le c$$

Proof. The function $f(t) = \frac{\ln \frac{t}{\delta}}{t}$ is non-increasing on $(e\delta, \infty)$. Thus, it is sufficient to show that the inequality holds for $t = 4\frac{\ln \frac{2}{c\delta}}{c}$. Since $\ln(x) \le \frac{x}{2}$ for all $x \in \mathbb{R}_+$, we can upper-bound this value of t as

$$t \le \frac{4}{c^2\delta}.$$

This implies

$$\ln \frac{t}{\delta} \le \ln \frac{4}{c^2\delta^2} = 2 \ln \frac{2}{c\delta} = \frac{c}{2} t \le ct$$

which proves the claim. □