Proportional Response: Contextual Bandits for Simple and Cumulative Regret Minimization

Sanath Kumar Krishnamurthy
Management Science and Engineering
Stanford University
sanathsk@stanford.edu

Ruohan Zhan
Industrial Engineering and Decision Analytics
Hong Kong University of Science and Technology
rhzhan@ust.hk

Susan Athey
Graduate School of Business
Stanford University
athey@stanford.edu

Emma Brunskill
Computer Science Department
Stanford University
ebrun@cs.stanford.edu

Abstract

In many applications, e.g. in healthcare and e-commerce, the goal of a contextual bandit may be to learn an optimal treatment assignment policy at the end of the experiment. That is, to minimize simple regret. However, this objective remains understudied. We propose a new family of computationally efficient bandit algorithms for the stochastic contextual bandit setting, where a tuning parameter determines the weight placed on cumulative regret minimization (where we establish near-optimal minimax guarantees) versus simple regret minimization (where we establish state-of-the-art guarantees). Our algorithms work with any function class, are robust to model misspecification, and can be used in continuous arm settings. This flexibility comes from constructing and relying on "conformal arm sets" (CASs). CASs provide a set of arms for every context, encompassing the context-specific optimal arm with a certain probability across the context distribution. Our positive results on simple and cumulative regret guarantees are contrasted with a negative result, which shows that no algorithm can achieve instance-dependent simple regret guarantees while simultaneously achieving minimax optimal cumulative regret guarantees.

1 Introduction

Learning and deploying personalized treatment assignment policies is crucial across domains such as healthcare and e-commerce Murphy [2003], Li et al. [2010]. Traditional randomized control trials (RCTs), while foundational for policy learning [Banerjee et al., 2016, Das et al., 2016], can be inefficient and costly Offer-Westort et al. [2021]. This motivates the study of adaptive sequential experimentation algorithms for the stochastic contextual bandit (CB) settings. The algorithm interacts with a finite sequence of users drawn *stochastically* from a fixed but unknown distribution. At each round, the algorithm receives a *context* (a user's feature vector), selects an action, and gets a corresponding reward. At the end of this adaptive experiment, the algorithm outputs a learned policy (mapping between contexts and actions).

Our algorithms are designed with the dual objectives of minimizing *simple regret* and *cumulative regret*. Simple regret quantifies the difference between the expected rewards achieved by the optimal policy and the policy learned at the conclusion of the experimental process. In contrast, cumulative regret encapsulates the summation of differences between the expected rewards generated by the

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

optimal policy and the exploration policies employed at each sequential round of decision-making.[1] Although there are many settings where simple regret is an important consideration, the majority of research in the contextual bandit field has focused on the minimization of cumulative regret. To the best of our knowledge, there is no general-purpose computationally efficient algorithm for pure exploration objectives like simple regret minimization in the contextual bandit setting. Further, there has been relatively little work so far into algorithms that explore the trade-off between multiple objectives like cumulative regret and simple regret (though see Athey et al. [2022], Erraqabi et al. [2017], Yao et al. [2021] for studies that address this empirically or juxtapose minimizing cumulative regret with estimating treatment effects or arm parameters). Our work seeks to address these gaps. We show that there is a trade-off between simple and cumulative regret minimization (formalized later in a lower-bound result). To navigate this trade-off, we proposes a new algorithm called Risk Adjusted Proportional Response (RAPR) with a tuning parameter $\omega \in [1, K]$, which governs the weight placed on the two objectives.[2] The algorithm is general-purpose (in that it can address any user-specified reward and policy classes), ensures near-optimal guarantees, and is also computationally efficient.

Types of guarantees. In our analysis, we consider two different types of bounds on simple and cumulative regret, worst-case and instance-dependent guarantees. Here instance-dependent guarantees refer to bounds that surpass worst-case rates by exploiting instances with large gaps between the conditional expected rewards of the optimal and sub-optimal arms. Recent work by [Foster and Rakhlin, 2020] has shown that it is not possible for contextual algorithms to have instance-dependent guarantees on cumulative regret (without suffering an exponential dependence on model class complexity); the authors instead develop algorithms that achieve minimax optimal (worst case optimal) cumulative regret guarantees (with square-root dependence on model class complexity). Li et al. [2022] developed the first general-purpose contextual bandit algorithm for pure exploration, and their algorithm achieved instance-dependent guarantees. They also show that instance-dependent best policy identification guarantees must come at the cost of worse than minimax optimal cumulative regret (discussed in detail later). We show a similar lower bound on cumulative regret for algorithms that achieve better instance-dependent simple regret guarantees, and propose the first family of algorithms that flexibly navigate such trade-offs.

Overview of our guarantees. The simple regret guarantees of RAPR are never worse than the minimax optimal rates (Theorem 2). Depending on the instance, RAPR achieves simple regret guarantees that are up to $O(1/\sqrt{\omega})$ times smaller compared to minimax optimal rates (Theorem 2). This improvement factor of $O(1/\sqrt{\omega})$ over minimax optimal rates is asymptotically achieved for instances where *realizability* holds (the reward model class is well specified) and the gap between the best and second best arm in terms of conditional expected reward is at least $\Delta > 0$ at every context (best-case instance in Theorem 2). RAPR provides these instance-dependent guarantees without the knowledge of any instance information. Unfortunately, the corresponding cumulative regret for the above instances is a factor of $O(\sqrt{\omega})$ times larger compared to minimax optimal rates (Theorem 1). The cumulative regret guarantees of our algorithm only degrade relative to the minimax optimal rate if the instance allows for better simple regret guarantees. Our lower bound (Theorem 3) considers the instances described above with $\Delta = 0.24$ (the gap between best and second best arm in terms of conditional expected reward). Theorem 3 shows that, for any algorithm that bounds the simple regret on these instances to $O(1/\sqrt{\omega})$ of the minimax optimal rates, its cumulative regret will be at least $\Omega(\sqrt{\omega})$ times the minimax optimal rates. RAPR thus achieves a near-optimal trade-off between guarantees on simple vs cumulative regrets when T is large enough. The trade-off contrasts with non-contextual bandits, where successive elimination ensures improved (compared to minimax) instance-dependent guarantees for both simple and cumulative regret [Even-Dar et al., 2006, Slivkins et al., 2019].

Types of CB algorithms. Contextual bandit algorithms broadly fall into two categories: regression-free and regression-based. Regression-free algorithms create an explicit policy distribution, randomly choosing a policy for decision-making at any time-step [Agarwal et al., 2014, Beygelzimer et al., 2011, Dudik et al., 2011, Li et al., 2022]. While these algorithms provide worst-case cumulative regret guarantees [Agarwal et al., 2014, Beygelzimer et al., 2011, Dudik et al., 2011] or instance-dependent

[1] Our formal definition of simple regret compares against the best policy in our policy class, while our cumulative regret definition compares against the global optimal policy (induced by the true conditional expected reward model). The reason for this discrepancy is because we use a regression based approach (due to computational considerations) for constructing our exploration policies.

[2] K is the number of arms for the finite arm setting.

PAC guarantees for policy learning [Li et al., 2022] without additional assumptions, they can be computationally intensive Foster and Rakhlin [2020]: they require solving and storing the output of $\Omega(\text{poly}(T))$ cost-sensitive classification (CSC) problems [Krishnamurthy et al., 2017] at every epoch (or update step). In contrast, regression-based algorithms [e.g., Abbasi-Yadkori et al., 2011, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2020] construct a conditional arm distribution using regression estimates of the expected reward, allowing for methods that need only solve $\mathcal{O}(1)$ regression or CSC problems at every epoch (or update step). Traditionally, these algorithms relied on realizability assumptions for optimal regret guarantees, but recent advances allow for misspecified reward model classes [Carranza et al., 2023, Foster et al., 2020a, Krishnamurthy et al., 2021]. We develop regression-based algorithms and do not assume realizability. RAPR is the first general-purpose regression-based algorithm with attractive pure exploration (simple regret) guarantees.

Overview of our algorithm. We now describe the RAPR algorithm in more detail. We first define a surrogate objective for simple regret, the optimal cover, which is inversely proportional to the probability that the bandit exploration policy chooses the arm recommended by the unknown optimal policy. The optimal cover bounds the variance of evaluating the unknown optimal policy under our exploration policy. This surrogate objective can be minimized by appropriately designing our exploration policy/action selection kernels. To maintain the attractive computational properties of regression-based algorithms, RAPR does not construct an explicit distribution over policies as that distribution would have large support and would be computationally and memory intensive to maintain. Instead, the goal of minimizing the optimal cover is attained by directly constructing a distribution over arms for each arriving context. This in turn builds on a novel general-purpose uncertainty quantification at each context. Much of the existing literature constructs confidence intervals with point-wise guarantees, but existing approaches to constructing them rely on assumptions like linear realizability. For general function classes, these intervals may be too wide and are often computationally expensive to construct. To overcome this issue, we develop Conformal Arm Sets (CASs), which are a set of potentially optimal arms at each context. This uncertainty quantification is regression-based and computationally efficient to construct; it's general-purpose and shrinks at "fast rates" (with square-root dependency on expected squared error bounds for regression). Unfortunately, these sets come with some risk of not containing the arm recommended by the optimal policy at every context. Nevertheless, we can use this uncertainty quantification to construct a distribution over arms at each context that helps us minimize the optimal cover by balancing the benefits and risks of relying on these CASs. The unavoidable trade-off between our simple and cumulative regret guarantees is an artifact of these risky sets. Beyond allowing us to trade off simple and cumulative regret guarantees, the flexibility of the approach also helps us extend to continuous arm settings and allows us to handle model misspecification.

Other Related Work. Our work connects to the literature on pure exploration, extensively studied in MAB settings (see overview in [Lattimore and Szepesvári, 2020]). [Even-Dar et al., 2006, Hassidim et al., 2020] study elimination-based algorithms for fixed confidence best-arm identification (BAI). [Kasy and Sautmann, 2021, Russo, 2016] study variants of Thompson Sampling with optimal asymptotic designs for BAI. [Karnin et al., 2013] propose sequential halving for fixed budget BAI. Our algorithm provides fixed confidence simple regret guarantees and can be seen as a generalization of successive elimination [Even-Dar et al., 2006] to the contextual bandit setting. The key technical difference is that it is often impossible to construct sub-gaussian confidence intervals on conditional expected rewards. The uncertainty quantification we use is similar to the notion of conformal prediction (see [Vovk et al., 2005] for a detailed exposition). Until recently, pure exploration had been nearly unstudied in contextual bandits. [Zanette et al., 2021] provide a static exploration algorithm that achieves the minimax lower bound on sample complexity for linear contextual bandits. [Li et al., 2022] then provided the first algorithm with instance-dependent (ϵ, δ)-PAC guarantees for contextual bandits. This algorithm is regression-free (adapts techniques from [Agarwal et al., 2014]) and requires a sufficiently large dataset of offline contexts as input. Hence, unfortunately, it inherits high memory and runtime requirements [See Foster and Rakhlin, 2020, for a more detailed discussion]. However, these costs come with the benefit that their notion of instance dependence leverages structure not only in the true conditional expected reward (as in Theorem 2) but also in the policy class (similar to policy disagreement coefficient Foster et al. [2020b]). They also prove a negative result, showing that it is not possible for an algorithm to have instance-dependent $(0, \delta)$-PAC guarantees and achieve minimax optimal cumulative regret guarantees. Our hardness result is similar but complementary to their result, for we show a similar result for simple regret (rather than their $(0, \delta)$-PAC sample

complexity).[3] Our work also recovers some cumulative regret guarantees for the continuous arm case [Majzoubi et al., 2020, Zhu and Mineiro, 2022], with new guarantees on simple regret and robustness to misspecification. Note that our restriction to "slightly randomized" policies for the continuous arm case results in regret bounds with respect to a "slightly randomized" (smooth) benchmark [see Zhu and Mineiro, 2022, for smooth regret].

1.1 Stochastic Contextual Bandits

We consider the stochastic contextual bandit setting, with context space $\mathcal{X}$, (compact) arm space $\mathcal{A}$, and a fixed but unknown distribution D over contexts and arm rewards. $D_{\mathcal{X}}$ refers to the marginal distribution over contexts, and T signifies the number of rounds or sample size. At each time $t \in [T]$[4], the environment draws a context x_t and a reward vector $r_t \in [0,1]^{\mathcal{A}}$ from D; the learner chooses an arm a_t and observes a reward $r_t(a_t)$. To streamline notation for discrete and continuous arm spaces, we consider a finite measure space $(\mathcal{A}, \Sigma, \mu)$ over the set of arms, with K shorthand for $\mu(\mathcal{A})$.[5] For ease of exposition, we focus on the finite/discrete arm setting. Here $\mathcal{A} = [K]$ and μ is the count measure, and $\mu(S) = |S|$ for any $S \subseteq \mathcal{A}$. A (deterministic) policy π maps contexts to *singleton arm sets* $\Sigma_1 := \{a | a \in \mathcal{A}\}$[6]. With some abuse of notation, we also let π refer to the kernel given by $\pi(a|x) = I(a \in \pi(x))$. An action selection kernel (randomized policy) $p : \mathcal{A} \times \mathcal{X} \to [0,1]$ is a probability kernel that describes a distribution $p(\cdot|x)$ over arms at every context x. We let $D(p)$ be the induced distribution over $\mathcal{X} \times \mathcal{A} \times [0,1]$, where sampling $(x, a, r(a)) \sim D(p)$ is equivalent to sampling $(x, r) \sim D$ and then sampling $a \sim p(\cdot|x)$.

A reward model f maps $\mathcal{X} \times \mathcal{A}$ to $[0,1]$, with $f^*(x,a) := \mathbb{E}_D[r_t(a)|x_t = x]$ denoting the true conditional expected reward model. Our algorithm works with a reward model class $\mathcal{F}$ and a policy class Π. For a given model f and an action selection kernel p, we denote the expected instantaneous reward of p with f as $R_f(p)$. We write $R_{f^*}(p)$ as $R(p)$ to simplify notation when no confusion arises. The optimal policy associated with reward function f is defined as π_f[7].

$$R_f(p) := \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p(\cdot|x)} [f(x,a)], \text{ and } \pi_f \in \arg\max_{\pi} R_f(\pi).$$

The policy π_f induced by $f \in \mathcal{F}$ is assumed to be within policy class Π without loss of generality.[8] For any $S \subseteq \mathcal{A}$, with some abuse of notation, we let $f(x, S) = \int_{a \in S} f(x,a) d\mu(a) / \mu(S)$. Note that $\pi_f(x) \in \arg\max_{S \in \Sigma_1} f(x, S)$ for all x. The *regret* of a policy π with respect to f is the difference between the optimal value and the actual value of π, denoted as $\text{Reg}_f(\pi) := R_f(\pi_f) - R_f(\pi)$. Finally, we let π^* denote the optimal policy in the class Π and let $\text{Reg}_\Pi(\cdot)$ denote the regret with respect to π^*. That is, $\pi^* \in \arg\max_{\pi \in \Pi} R(\pi)$ and $\text{Reg}_\Pi(\pi) := R(\pi^*) - R(\pi)$.

Objectives. Contextual bandit algorithms adaptively construct action sampling kernels (exploration policies) $\{p_t\}_{t \in [T]}$ used to collect data over the T rounds. At the end of the adaptive experiment, the adaptively collected data is used to learn a policy $\hat{\pi} \in \Pi$. We study two main objectives to measure quality of these outputs: [Objective 1] *Cumulative regret minimization.* Cumulative regret (CReg_T) is given by $\text{CReg}_T := \sum_{t=1}^{T} \text{Reg}_{f^*}(p_t)$. It compares the cumulative expected reward obtained during the experiment with the expected reward of the policy (π_{f^*}) induced by the true conditional expected reward (f^*). We seek to minimize cumulative regret which is equivalent to maximizing cumulative expected reward during the experiment. [Objective 2] *Simple regret minimization.* Simple regret is given by $\text{Reg}_\Pi(\hat{\pi})$. It compares the expected reward of the learnt policy $\hat{\pi} \in \Pi$ against the value of the optimal policy in the class Π. We seek to minimize simple regret which is equivalent to maximizing expected reward of the policy learnt at the end of the experiment. To understand the kind of exploration kernels ($\{p_t\}_{t \in [T]}$) that help with policy learning, we now identify a surrogate objective for simple regret (called optimal cover) that is in terms of the kernels used for exploration.

[3] In (ϵ, δ) PAC sample complexity results, given an input (ϵ, δ), the objective is to minimize the number of samples needed in order to output an ϵ-optimal policy with probability at least $1 - \delta$ (a "fix accuracy, compute budget" setting). In contrast, in our simple regret case, we consider how to minimize the error ϵ as the number of samples increases.

[4] For any $n \in \mathbb{N}^+$, we use notation $[n]$ to denote the set $\{1, ..., n\}$

[5] Here Σ is a σ-algebra over $\mathcal{A}$ and μ is a bounded set function from Σ to the real line.

[6] The introduction of Σ_1 is to allow for easy generalization to the continuous arm setting.

[7] subject to any tie-breaking rule.

[8] Note that Π may contain policies that are not induced by models in the class $\mathcal{F}$.

Definition 1 (Cover). *Given a kernel p and a policy π, we define the cover of policy π under the kernel p to be,*

$$V(p,\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim\pi(\cdot|x)}\left[\frac{\pi(a|x)}{p(a|x)}\right]. \tag{1}$$

Additionally, for any pair of kernels (p,q), we let $V(p,q) := \mathbb{E}_{x\sim D_{\mathcal{X}}, a\sim q(\cdot|x)}[q(a|x)/p(a|x)]$. Finally, we use the term optimal cover for kernel p to refer to $V(p,\pi^*)$.

The cover measures the quality of data collected under the action selection kernel p for evaluating a given policy π and bounds the variance of commonly used unbiased estimators for policy value [e.g., Agarwal et al., 2014, Hadad et al., 2021, Zhan et al., 2021]. In particular, the cover under optimal policy $\frac{1}{T}\sum_{t=1}^{T} V(p_t,\pi^*)$ can be treated as a surrogate objective for simple regret minimization (proven in Appendix E.2), which is particularly instructional in designing our algorithm to minimize simple regret.

Extending notation to continuous arms. In the continuous arm setting, evaluating arbitrary deterministic policies can be infeasible without extra assumptions [Mou et al., 2023]. Thus, we focus on "slightly randomized" policies by generalizing Σ_1 to be the arm sets with measure one ($\Sigma_1 := \{S \in \Sigma | \mu(S) = 1\}$).[9] The granularity of these sets can be adjusted by scaling the finite measure μ, which also affects the value of $K = \mu(\mathcal{A})$. We then continue defining policies be maps from $\mathcal{X}$ to Σ_1 and Π is a class of such policies. We overload notation and define the induced kernel as $\pi(a|x) = I(a \in \pi(x))$, which is a valid definition since $\int_a I(a \in \pi(x)) d\mu(a) = \mu(\pi(x)) = 1$. All the remaining definitions, including $R_f(\pi), \pi_f, \pi^*$ and $V(p,\pi)$, relied on these induced kernels and continue to hold. While there are some measure theoretic issues that remain to be discussed, we defer these details to Appendix A.

Uniform sampling. Our algorithm frequently selects an arm uniformly from a constructed set of arms. In the context of a set $S \subseteq \mathcal{A}$, uniform sampling refers to selecting an arm from the distribution $q(a) := I(a \in S)/\mu(S)$. This constitutes a probability measure since its integral over $\mathcal{A}$ equals 1. In the discrete arm setting, uniform sampling from a set $S \subseteq \mathcal{A}$ implies selecting an arm according to the distribution $I(a \in S)/|S|$.

1.2 Oracle Assumptions

Our algorithm relies on two sub-routines. For generality, we abstract away these sub-routines by stating them as oracle assumptions, for which we describe two oracles, EstOracle and EvalOracle, in Assumptions 1 and 2 respectively. The EstOracle sub-routine is for estimating conditional expected reward models (Assumption 1), and the EvalOracle sub-routine is for estimating policy values (Assumption 2) according to the true and estimated reward models.

These sub-routine tasks are supervised learning problems. Hence, the average errors for the corresponding tasks can be bounded in terms of the number of samples (n) and a confidence parameter (δ'). The oracle assumptions specify the estimation rates. We let $\xi : \mathbb{N} \times [0,1] \to [0,1]$ denote the estimation rate for these oracles. For simplicity, we assume that they share the same rate and that $\xi(n,\delta')$ scales polynomially in $1/n$ and $\log(1/\delta')$. In order to simplify the analysis, we also require $\xi(n/3, \delta'/n^3)$ be non-increasing in n.[10] We now formally describe these oracle assumptions, starting with EstOracle.

Assumption 1 (Estimation Oracle). *We assume access to a reward model estimation oracle (EstOracle) that takes as input an action selection kernel p, and n independently and identically drawn samples from the distribution $D(p)$. The oracle then outputs an estimated model $\hat{f} \in \mathcal{F}$ such that for any $\delta' \in (0,1)$, the following holds with probability at least $1-\delta'$:*

$$\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p(\cdot|x)} [(\hat{f}(x,a) - f^*(x,a))^2] \leq B + \xi(n,\delta')$$

[9] Note that our restriction to "slightly randomized" policies for the continuous arm case results in regret bounds with respect to a "slightly randomized" (smooth) benchmark. Hence for the continuous arm case, our cumulative regret bounds translate to smooth regret bounds from Zhu and Mineiro [2022] with $K = 1/h$. Where h is the measure of smoothness in smooth regret (a leading objective for this setting).

[10] This ensures that ξ_m defined in Lemma 1 is non-increasing in m for any epoch schedule with increasing epoch lengths.

Where $B \geq 0$ is a fixed but unknown constant that may depend on the model class $\mathcal{F}$ and distribution D, but is independent of the action selection kernel p.

In Assumption 1, the parameter B measures the bias of model class $\mathcal{F}$; under realizability, B equals 0. The function ξ characterizes the estimation variance, which decreases with increasing sample size. *As long as the variance term (which shrinks as we gather more data) is larger than the fixed unknown bias (B)*, we have from Assumption 1 that the expected squared error for the estimated reward model is bounded by 2ξ. We use this bound on expected squared error to further bound how accurately the estimated reward model evaluates policies in the class Π (Lemma 6). However, since B is unknown, we need a test to detect when this policy evaluation bounds starts failing (which can only happen after the variance term gets dominated by the unknown bias term). To construct this test, our algorithm relies on EvalOracle, which provides consistent independent policy value estimates and helps compare them with policy value estimates with respect to the estimated reward model.

Assumption 2 (Evaluation Oracle)**.** *We assume access to an oracle (EvalOracle) that takes as input an action selection kernel p, n independently and identically drawn samples from the distribution $D(p)$, a set of m models $\{g_i | i \in [m]\} \subseteq \mathcal{F}$, and another action selection kernel q. The oracle then outputs a policy evaluation estimator $\hat{R}$ of true policy value, and a set of m policy evaluation estimators $\{\hat{R}_{g_i} | i \in [m]\}$ that estimate policy value with respect to the models $g_1, g_2, \ldots, g_m$ respectively. Such that for any $\delta' \in (0,1)$, the following conditions simultaneously hold with probability at least $1-(m+1)\delta'$:*

- $|\hat{R}(\pi) - R(\pi)| \leq \sqrt{2V(p,\pi)\xi(n,\delta')} + 2\xi(n,\delta')/(\min_{(x,a)\in\mathcal{X}\times\mathcal{A}} p(a|x))$ *for all* $\pi \in \Pi \cup \{q\}$.
- $|\hat{R}_f(\pi) - R_f(\pi)| \leq \sqrt{2\xi(n,\delta')}$ *for all* $\pi \in \Pi \cup \{q\}$ *and for all* $f \in \{g_i | i \in [m]\}$.

When $\mathcal{F}$ and Π are finite, one can construct oracles such that Assumptions 1 and 2 hold with $\xi(n,\delta') = \mathcal{O}(\log(\max(|\mathcal{F}|, |\Pi|)/\delta')/n)$. One example of such a construction is given by using empirical squared loss minimization for EstOracle, using inverse propensity scores (IPS) for estimating $R(\pi)$ in EvalOracle, and using the empirical average for estimating $R_f(\pi)$ in EvalOracle. The guarantees of these assumptions can be derived using Bernstein's inequality and union bounding. When $\mathcal{F}$ has pseudo-dimension [Koltchinskii, 2011] bounded by d and Π has the Natarajan-dimension bounded by d [Jin et al., 2022, Jin, 2022], one can construct oracles such that Assumptions 1 and 2 hold with $\xi(n,\delta') = \mathcal{O}(d\log(nK/\delta')/n)$.

2 Algorithm

At a high level, our algorithm operates in two modes, indicated by a Boolean variable "**safe**". During mode one (**safe** = **True**), where estimated reward models are sufficiently accurate at evaluating policies in the class Π,[11] we use our estimated models to update our action selection kernel used during exploration. During mode two (**safe** = **False**), where the condition for mode one no longer holds, we stop updating the action selection kernel used for exploration. Operationally our algorithm runs in epochs/batches indexed by m. Epoch m begins at round $t = \tau_{m-1} + 1$ and ends at $t = \tau_m$, and we use $m(t)$ to denote the epoch index containing round t. We let $\hat{m}$ denote the critical epoch, at the end of which our algorithm changes mode (with "**safe**" being updated from "**True**" to "**False**"); we refer to $\hat{m}$ as the *algorithmic safe epoch*. For all rounds in epoch $m \leq \hat{m}$, our algorithm samples action using the action selection kernel p_m defined later in (4). For $m > \hat{m}$, our algorithm samples action using $p_{\hat{m}}$ –the action selection kernel used in the algorithmic safe epoch $\hat{m}$.

We now describe the critical components of our algorithm. These include (i) data splitting and using oracle sub-routines; (ii) *misspecification tests*, which we use to identify the **safe**-mode switching epoch $\hat{m}$; and (iii) *conformal arm sets*, which presents a new form of uncertainty quantification that is critical in constructing p_{m+1} at the end of each epoch $m \in [\hat{m}]$. Finally, we use these components to describe our final algorithm.

Data splitting and oracle sub-routines. Consider an epoch $m \in [\hat{m}]$. Let S_m denote the set of samples collected in this epoch: $S_m = \{(x_t, a_t, r_t(a_t)) | t \in [\tau_{m-1}, \tau_m]\}$. Our algorithm splits S_m into three equally-sized subsets: $S_{m,1}, S_{m,2}$ and $S_{m,3}$. Algorithm 1 outlines using these subsets and the oracles (described in Section 1.2) to estimate reward models and evaluate policies. Based

[11] Where the estimated reward models pass the misspecification test.

Algorithm 1 ω Risk Adjusted Proportional Response (ω-RAPR)

input: Trade-off parameter $\omega \in [1, K]$, proportional response threshold $\beta_{\max} = 1/2$, and confidence parameter δ (used in definition of ξ_m).

Let $p_1(a|x) \equiv 1/\mu(\mathcal{A}) = 1/K$, $\hat{f}_1 \equiv 0$, $\alpha_1 = 3K$, $\tau_1 = 3$, and **safe** = **True**.
for epoch $m = 1, 2, \ldots$ **do**
  $\tau_m = 2\tau_{m-1}$. ▷ Doubling epochs.
  if safe then
    for round $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
      Observe context x_t, sample $a_t \sim p_m(\cdot|x_t)$, and observe $r_t(a_t)$.
    end for
    Let S_m denote the data collected in epoch m.
    We split S_m into three equally sized sets $S_{m,1}, S_{m,2}$ and $S_{m,3}$.
    Let $\hat{f}_{m+1} \leftarrow \text{EstOracle}(p_m, S_{m,1})$, and let C_{m+1} be given by Definition 2.
    Let η_{m+1} be the solution to (5) and let $\alpha_{m+1} := 3K/\eta_{m+1}$. ▷ $S_{m,2}$ is used here.
    Now let p_{m+1} be given by (4).
    Let $\hat{R}_{m+1}, \{\hat{R}_{m+1,\hat{f}_i} | i \in [m+1]\} \leftarrow \text{EvalOracle}(p_m, S_{m,3}, \{\hat{f}_i | i \in [m+1]\}, p_{m+1})$.
    if (2) does not hold. **then**
      $\hat{m}$, **safe** $\leftarrow m$, **False**.
    end if
  else
    for round $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
      Observe context x_t, sample $a_t \sim p_{\hat{m}}(\cdot|x_t)$, and observe $r_t(a_t)$.
    end for
    Let S_m denote the data collected in epoch m.
    Let $\hat{R}_{m+1}, \{\hat{R}_{m+1,\hat{f}_i} | i \in [\hat{m}]\} \leftarrow \text{EvalOracle}(p_{\hat{m}}, S_m, \{\hat{f}_i | i \in [\hat{m}]\}, p_{\hat{m}})$.
  end if
end for

on Assumptions 1 and 2, we bound the errors for these estimates in terms of $\xi_{m+1} = 2\xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3))$, where δ is a specified confidence parameter. As we will see later, our algorithm relies on these bounds to test for misspecification and construct action selection kernels.

Misspecification test. We first discuss the need for our misspecification test. Note that Assumption 1 is flexible and allows our reward model class $\mathcal{F}$ to be misspecified. In particular, the squared error of our reward model estimate may depend on an unknown bias term B. To account for this unknown B, it is useful to center our analysis around the safe epoch $m^* := \arg\max\{m \geq 1 | \xi_{m+1} \geq 2B\}$, which denotes the last epoch where variance dominates bias. We show that for any epoch $m \in [m^*]$, the estimated reward model $\hat{f}_{m+1}$ is "sufficiently accurate" at evaluating the expected reward of any policy in $\Pi \cup \{p_{m+1}\}$. This property is critical in ensuring that the constructed action selection kernel p_{m+1} has low exploration regret $\text{Reg}_{f^*}(p_{m+1})$ and a small optimal cover ($V(p_{m+1}, \pi^*)$). Since B and m^* are unknown, we need to test whether the estimated reward model is sufficiently accurate at evaluating these policies. When the test fails, the algorithm sets the variable "safe" to **False** and stops updating the action selection kernel used for exploration. The core idea for this test comes from Krishnamurthy et al. [2023] although its application to simple regret minimization is new, and the form of our test differs a bit. We now state our misspecification test (2). At the end of each epoch m, the test is passed if (2) holds:

$$\max_{\pi \in \Pi \cup \{p_{m+1}\}} |\hat{R}_{m+1,\hat{f}_{m+1}}(\pi) - \hat{R}_{m+1}(\pi)| - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi)}{40\bar{m}^2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}$$
$$\leq 2.05\sqrt{\alpha_m \xi_{m+1}} + 1.1\sqrt{\xi_{m+1}}, \tag{2}$$

where $\alpha_{\bar{m}}$ empirically bounds $V(p_{\bar{m}}, \pi^*)$, the optimal cover for the action selection kernel used in epoch $\bar{m}$ (see (49)). The first term in (2) measures how well the estimated reward model $\hat{f}_{m+1}$ evaluates the policy π, and the second term accounts for under-explored policies (policies that have high regret under the reward model $\hat{f}_m$ would be less explored in epoch m).

Conformal arm sets. We proceed to introduce the notion of *conformal arm sets* (CASs), based on which we construct the action selection kernels employed by our algorithms. At the beginning of each epoch m, we construct CASs, denoted as $\{C_m(x,\zeta)|x \in \mathcal{X}, \zeta \in [0,1]\}$; here ζ controls the probability with which the set C_m contains the optimal arm. The construction of these sets rely on the models $(\hat{f}_1, \ldots, \hat{f}_m)$ estimated from data up to epoch $m-1$, as defined below.

Definition 2 (Conformal Arm Sets)**.** *Consider $\zeta \in (0,1)$. At epoch m, for context x, the arm set $C_m(x,\zeta)$ is given by (3).*

$$
\begin{aligned}
&C_m(x,\zeta) := \pi_{\hat{f}_m}(x) \bigcup \bar{C}_m(x,\zeta), \quad \bar{C}_m(x,\zeta) := \bigcap_{\bar{m}\in[m]} \tilde{C}_{\bar{m}}\Big(x, \frac{\zeta}{2\bar{m}^2}\Big), \\
&\tilde{C}_{\bar{m}}(x,\zeta') := \left\{ a : \hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a) \leq \frac{20\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}{\zeta'} \right\} \forall \bar{m} \in [m], \zeta' \in (0,1).
\end{aligned} \tag{3}
$$

Similar to conformal prediction (CP) [Vovk et al., 2005, Shafer and Vovk, 2008], CASs have marginal coverage guarantees. We show that with high probability, we have $\pi^*(x)$ lies in $C_m(x,\zeta)$ with probability at least $1-\zeta$ over the context distribution. That is, $\Pr_{x\sim D_{\mathcal{X}}}(\pi^*(x) \in C_m(x,\zeta)) \geq 1-\zeta$ with high-probability (see Appendix E.1). However, there is also a key technical difference. While CP provides coverage guarantees for the conditional random outcome, CASs provide coverage guarantees for $\pi^*(x)$ – which is not a random variable given the context x. Hence, intervals estimated by CP need to be wide enough to account for conditional outcome noise, whereas CASs do not. CASs also have several advantages compared to pointwise confidence intervals used in UCB algorithms. First, CASs are computationally easier to construct. Second, CAS widths have a polynomial dependency on model class complexity, whereas pointwise intervals may have an exponential dependence for some function classes [see lower bound examples in Foster et al., 2020b]. Third, pointwise intervals require realizability, whereas the guarantees of CASs hold even without realizability (as long as the misspecification test in (2) holds). However, it's important to remember that these benefits of CASs come with the risk of only covering $\pi^*(x)$ marginally over the context distribution – that is, these sets may not contain $\pi^*(x)$ at all x.

Risk Adjusted Proportional Response Algorithm. We now describe the design of our algorithm, which is summarized in Algorithm 1. The algorithm depends on the following input parameters: $\omega \in [1,K]$ which controls the trade-off between simple and cumulative regret, the proportional response threshold $\beta_{\max} = 1/2$, and confidence parameter δ. The algorithm also computes η_{m+1} (risk adjustment parameter for p_{m+1}), α_{m+1} (empirical bound on optimal cover for p_{m+1}), and $\lambda_{m+1}(\cdot)$ (empirical bound on average CAS size). At the end of every epoch $m \in [\hat{m}]$, we construct the action selection kernel p_{m+1} given by (4).

$$
p_{m+1}(a|x) = \frac{(1-\beta_{\max})I[a \in C_{m+1}(x, \beta_{\max}/\eta_{m+1})]}{\mu\big(C_{m+1}(x, \beta_{\max}/\eta_{m+1})\big)} + \int_0^{\beta_{\max}} \frac{I[a \in C_{m+1}(x, \beta/\eta_{m+1})]}{\mu\big(C_{m+1}(x, \beta/\eta_{m+1})\big)} \mathrm{d}\beta. \tag{4}
$$

At any context x, sampling arm a from $p_{m+1}(\cdot|x)$ is equivalent to the following. Sample β uniformly from $[0,1]$, then sample arm a uniformly from the set $C_{m+1}(x, \min(\beta_{\max}, \beta)/\eta_{m+1})$. A small β results in a larger CAS and a higher probability of containing the optimal arm for the sampled context. However, uniformly sampling an arm from a larger CAS also implies a lower probability on every arm in the set. Sampling β uniformly allows us to respond proportionately to the risk of not sampling the optimal arm while enjoying the benefits of smaller CASs. We refer to this as the *Proportional Response Principle*.

Similarly note that, a larger risk-adjustment parameter η_{m+1} encourages reliance on less risky albeit larger CASs. We want to choose η_{m+1} to tightly bound the the optimal cover (surrogate for simple regret), subject to cumulative regret constrains imposed by the trade-off parameter ω. To do this, we first let $\lambda_{m+1}(\eta)$ be a high-probability empirical upper bound on $E_{x\sim D_{\mathcal{X}}}[\mu(C_{m+1}(x, \beta_{\max}/\eta))]$. Hence, using (49), we can upper bound the optimal cover ($V(p_{m+1}, \pi^*)$) by $\frac{\lambda_{m+1}(\eta_{m+1})}{1-\beta_{\max}} + \frac{K}{\eta_{m+1}}$. Our choice of η_{m+1} approximately minimizes this upper bound on the optimal cover, by choosing the largest feasible $\eta \in [\eta_m, \sqrt{\omega K/\alpha_m}]$ such that $\lambda_{m+1}(\eta) \leq \frac{K}{\eta}$ (see (5)). Note that this choice of

η_{m+1} balances the risk of a small η (large $\frac{K}{\eta}$) with the benefits of a small $\lambda_{m+1}(\eta)$ (small $\frac{\lambda_{m+1}(\eta)}{1-\beta_{\max}}$).

$$\lambda_{m+1}(\eta) := \min\left(1 + \frac{1}{|S_{m,2}|}\sum_{t\in S_{m,2}} \mu(\bar{C}_{m+1}\Big(x, \frac{\beta_{\max}}{\eta}\Big)) + \sqrt{\frac{K^2\ln(8|S_{m,2}|(m+1)^2/\delta)}{2|S_{m,2}|}}, K\right),$$
$$\eta_{m+1} \leftarrow \max\left\{\eta_m,\ \max\left\{\eta = \frac{|S_{m,2}|}{n}\middle| n \in [|S_{m,2}|], \eta \le \sqrt{\frac{\omega K}{\alpha_m}}, \lambda_{m+1}(\eta) \le \frac{K}{\eta}\right\}\right\}. \tag{5}$$

With η_{m+1} chosen, the action selection kernel p_{m+1} is decided. Now let $\alpha_{m+1} = 3K/\eta_{m+1}$, which is a high-probability empirical upper bound on $V(p_{m+1}, \pi^*)$. We then use α_{m+1} at the end of epoch $m+1$ to construct CASs, compute the risk-adjustment parameter, and test for misspecification.

Computation. We have $\mathcal{O}(\log(T))$ epochs. At the end of any epoch $m \in [\hat{m}]$, we solve three optimization problems. The first is for estimating $\hat{f}_{m+1}$, which often reduces to empirical squared loss minimization and is computationally tractable for several function classes $\mathcal{F}$. The second is for computing the risk-adjustment parameter in (5) which can be solved via binary search. The third is for the misspecification test in (2), which can be solved via two calls to a cost-sensitive classification (CSC) solver (don't need this when assuming realizability, further if we only care about cumulative regret, sufficient to use the simpler test in Krishnamurthy et al. [2021]). Finally, to learn a policy $\hat{\pi}$ at the end of T rounds, we need to solve (8) using a CSC solver (under realizability we can set $\hat{\pi} = \pi_{\hat{f}_{m(T)-1}}$). Hence, overall, ω-RAPR makes exponentially fewer calls to solvers compared to regression-free algorithms like Li et al. [2022].

3 Main Results

Our algorithm/analysis/results hold for both the discrete and continuous arm cases. As discussed before, minimizing optimal cover helps us ensure improved simple regret guarantees. Hence $\alpha_m \in [1, 3K]$ (the high-probability empirical upper bound on the optimal cover $V(p_m, \pi^*)$) will play a crititcal role thoughout this results section. We start with stating our cumulative regret bounds.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Then with probability $1-\delta$, ω-RAPR attains the following cumulative regret guarantee. Here $\xi_{m+1} = \xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3)$ for all m.*

$$\mathit{CReg}_T \le \tilde{\mathcal{O}}\left(\sum_{t=\tau_1+1}^{T}\sqrt{\frac{K}{\alpha_{m(t)}}\frac{\alpha_{m(t)-1}}{\alpha_{m(t)}}}\left(\sqrt{KB} + \sqrt{K\xi_{m(t)}}\right)\right) \tag{6a}$$
$$\le \tilde{\mathcal{O}}\left(\sqrt{\omega KB}T + \sum_{t=\tau_1+1}^{T}\sqrt{\omega K\xi_{m(t)}}\right). \tag{6b}$$

Where we use $\tilde{\mathcal{O}}$ to hide terms logarithmic in $T, K, \xi(T, \delta)$.

We start with discussing (6b). The first part $\sqrt{\omega KB}T$ comes from the bias of the regression oracle with model class $\mathcal{F}$ and will vanish under the realizability assumption. The second part $\sum_{t=\tau_1+1}^{T}\sqrt{\omega K\xi_{m(t)}}$, when setting $\omega = 1$, recovers near-optimal (upto logarithmic factors) minimax cumulative regret guarantees for common model classes, as demonstrated by the following examples.

Corollary 1. *We consider ω-RAPR with appropriate oracles in the following cases and let B denote the corresponding bias terms. When $\mathcal{F}$ and Π are finite, $\mathit{CReg}_T \le \tilde{\mathcal{O}}(\sqrt{\omega KB}T + \sqrt{\omega KT\log(\max(|\mathcal{F}|, |\Pi|)/\delta)})$ with probability at least $1-\delta$. When $\mathcal{F}$ has a finite pseudo dimension d, Π has a finite Natarajan dimension d, and $\mathcal{A}$ is finite, $\mathit{CReg}_T \le \tilde{\mathcal{O}}(\sqrt{\omega KB}T + \sqrt{\omega KTd\log(TK/\delta)})$ with probability at least $1-\delta$. Note that under realizability ($B = 0$), 1-RAPR achieves near-optimal minimax cumulative guarantees.*

In (6a), we observe that the multiplicative $\sqrt{\omega}$ cost to cumulative regret is only incurred if the empirical bound on optimal cover ($\alpha_m \in [1, 3K]$) can get small. That is, our cumulative regret

bounds degrade only if our algorithm better bounds the optimal cover and thus ensures better simple regret guarantees. [12] We now provide instance dependent simple regret guarantees for our algorithm.

Theorem 2. *Suppose Assumptions 1 and 2 hold. For some $(\lambda, \Delta, A) \in [0,1] \times (0,1] \times [1, K]$, consider instances where for $1-\lambda$ fraction of contexts at most A arms are Δ optimal (i.e. (7) holds).*

$$\mathbb{P}_{x \sim D_{\mathcal{X}}}\Big(\mu\big(\{a \in \mathcal{A} : f^*(x, \pi_{f^*}(x)) - f^*(x,a) \le \Delta\}\big) \le A\Big) \ge 1 - \lambda. \tag{7}$$

Let $m' = \min(\hat{m}, m(T)) - 1$. Let the learned policy $\hat{\pi}$ be given by (8) (equivalent to variance penalized policy optimization).

$$\hat{\pi} \in \arg\max_{\pi \in \Pi} \hat{R}_{m(T)}(\pi) - \frac{1}{2}\sqrt{\alpha_{m'}\xi_{m(T)}} \sum_{\bar{m} \in [m']} \frac{\hat{R}_{m(T),\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m(T),\hat{f}_{\bar{m}}}(\pi)}{40\bar{m}^2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}. \tag{8}$$

Then with probability $1-\delta$, ω-RAPR has the following simple regret bound when T samples.

$$\begin{aligned} &\mathit{Reg}_{\Pi}(\hat{\pi}) \le \mathcal{O}\left(\sqrt{\alpha_{m'}\xi_{m(T)}}\right) \\ &\le \mathcal{O}\left(\sqrt{\xi_{m(T)} \min\left(K, A + K\lambda + \frac{K}{\omega} + \frac{K^{3/2}\omega^{1/2}}{\Delta}\sqrt{\xi_{\min(m^*, m(T)-1) - \lceil \log_2 \log_2(K) \rceil}}\right)}\right). \end{aligned}$$

Under (7), we can only argue that the expected (over context distribution) measure of Δ optimal arms is at most $(1-\lambda)A + K\lambda = O(A + K\lambda)$. Hence for large T, the best we can hope for is instance-dependant simple regret guarantees that shrink/improve over minimax guarantees by a factor of $\mathcal{O}(\sqrt{(A+K\lambda)/K})$. We show that this is guaranteed by Theorem 2. Suppose $\omega = K$, $\mathcal{F}$ has a finite pseudo dimension bounded by d, and Π has a finite Natarajan dimension bounded by d. The simple regret guarantee of Theorem 2 reduces to $\tilde{\mathcal{O}}(\min((\sqrt{Kd/T}, \sqrt{(A+K\lambda)d/T} + (K/\sqrt{\Delta})\sqrt{d/T}\sqrt[4]{B + d/T}))$. When the reward model estimation bias B is small enough, the term $(K/\sqrt{\Delta})\sqrt{d/T}\sqrt[4]{B+d/T})$ is dominated by the remaining terms for large T. Hence, in this case, we get a simple regret bound of $\tilde{\mathcal{O}}(\sqrt{(A+K\lambda)d/T})$ for large T. As promised, this improves upon the minimax guarantees by a factor of $\mathcal{O}(\sqrt{(A+K\lambda)/K})$.

Note that Theorem 1 guarantees are better for ω closer to 1 whereas Theorem 2 guarantees are better for ω closer to K. Hence these theorems show a tradeoff between the cumulative and simple regret guarantees for ω-RAPR. Theorem 3 shows that improving upon minimax simple regret guarantees for instances satisfying (7) may come at the unavoidable cost of worse than minimax optimal cumulative regret guarantees. This contrasts with non-contextual bandits, where successive elimination ensures improved (compared to minimax) gap-dependent guarantees for both simple and cumulative regret.

Theorem 3. *Given parameters $K, F, T \in \mathbb{N}$ and $\phi \in [1, \infty)$. There exists a context space $\mathcal{X}$ and a function class $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ with K actions such that $|\mathcal{F}| \le F$ and the following lower bound on cumulative regret holds:*

$$\inf_{\mathbf{A} \in \Psi_\phi} \sup_{D \in \mathcal{D}} \mathop{\mathbb{E}}_{D}\left[\sum_{t=1}^{T}\big(r_t(\pi^*(x_t)) - r_t(a_t)\big)\right] \ge \tilde{\Omega}\left(\sqrt{\frac{K}{\phi}}\sqrt{KT\log F}\right)$$

Here $(a_1, \dots a_T)$ denotes the actions selected by an algorithm A. $\mathcal{D}$ denotes the set of environments such that $f^* \in \mathcal{F}$ and (7) hold with $(A, \lambda, \Delta) = (1, 0, 0.24)$. Π denotes policies induced by $\mathcal{F}$. Ψ_ϕ denotes the set of CB algorithms that run for T rounds and output a learned policy with a simple regret guarantee of $\sqrt{\phi \log F/T}$ for any instance in $\mathcal{D}$ with confidence at least 0.95, i.e., $\Psi_\phi := \{A : \mathbb{P}(\mathit{Reg}(\hat{\pi}_{\mathbf{A}}) \le \sqrt{\phi \log F/T}) \ge 0.95$ for any instance in $\mathcal{D}\}$. Finally, $\tilde{\Omega}(\cdot)$ hides factors logarithmic in K and T.

[12] For large t, once our bounds on optimal cover can't be significantly improved, we have $\alpha_{m(t)}/\alpha_{m(t)-1} = O(1)$. Hence for large T, our cumulative regret is a factor of $\sqrt{K/\alpha_{m(T)+1}}$ larger than the near optimal minimax guarantees. As we will see in Theorem 3, this multiplicative factor is unavoidable.

Near optimal trade-off of RAPR. Note that the environments constructed in Theorem 3 satisfy $f^* \in \mathcal{F}$ with $\max(|\mathcal{F}|, |\Pi|) \leq F$ and also satisfy (7) with $(A, \lambda, \Delta) = (1, 0, 0.24)$. With appropriate oracles, Assumptions 1 and 2 are satisfied with $B = 0$ (i.e. $m^* = \infty$) and $\xi(n, \delta') = \mathcal{O}(\log(F/\delta')/n)$. Hence for large enough T, ω-RAPR achieves a simple regret bound of $\tilde{\mathcal{O}}(\sqrt{(K/\omega)\log F/T})$ with probability at least 0.95 and thus is a member of Ψ_ϕ for some $\phi = \tilde{\mathcal{O}}(K/\omega)$. Theorem 3 lower bounds the cumulative regret of such algorithms by $\tilde{\Omega}(\sqrt{K/\phi}\sqrt{KT\log F}) = \tilde{\Omega}(\sqrt{\omega}\sqrt{KT\log F})$. Up to logarithmic factors, this matches the cumulative regret upper bound for ω-RAPR. Re-emphasizing that the trade-off observed in Theorems 1 and 2 is near optimal for large T.

Simulations. To demonstrate the computational tractability of our approach, we ran a simulation on setting within a $\mathbb{R}^2$context space, eight arms, linear models, and an exploration horizon of 5000. Our algorithms ran in less than 9 seconds on a Macbook M1 Pro. We also compare with other baselines on simple/cumulative regret. See Appendix E.4 for details.

Conclusion. We develop Risk Adjusted Proportional Response (RAPR), a computationally efficient regression-based contextual bandit algorithm. It is the first contextual bandit algorithm capable of trading-off worst-case cumulative regret guarantees with instance-dependent simple regret guarantees. The versatility of our algorithm allows for general reward models, handles misspecification, extends to finite and continuous arm settings, and allows us to choose the trade-off between simple and cumulative regret guarantees. The key ideas underlying RAPR are conformal arm sets (CASs) to quantify uncertainty, proportional response principle for cumulative regret minimization, optimal cover as a surrogate for simple regret, and risk adjustment for better bounds on the optimal cover. [13]

Limitations. A limitation of our approach is that we do not utilize the structure of the policy class being explored. Further refining CASs with other forms of uncertainty quantification that leverage such structure can lead to significant improvements, and potentially avoid trade-offs between simple/cumulative regret when policy class structure allows for it.

References

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

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

S. Athey, U. Byambadalai, V. Hadad, S. K. Krishnamurthy, W. Leung, and J. J. Williams. Contextual bandits in a survey experiment on charitable giving: Within-experiment outcomes versus policy learning. *arXiv preprint arXiv:2211.12004*, 2022.

A. V. Banerjee, E. Duflo, and M. Kremer. The influence of randomized controlled trials on development economics research and on development policy. *The state of Economics, the state of the world*, pages 482–488, 2016.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

A. G. Carranza, S. K. Krishnamurthy, and S. Athey. Flexible and efficient contextual bandits with heterogeneous treatment effect oracles. In *International Conference on Artificial Intelligence and Statistics*, pages 7190–7212. PMLR, 2023.

J. Das, A. Chowdhury, R. Hussam, and A. V. Banerjee. The impact of training informal health care providers in india: A randomized controlled trial. *Science*, 354(6308):aaf7384, 2016.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 169–178, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

[13] S.A. and S.K.K. are grateful for the support provided by Golub Capital Social Impact Lab and the ONR grant N00014-19-1-2468. E.B. is grateful for the support of NSF grant 2112926.

A. Erraqabi, A. Lazaric, M. Valko, E. Brunskill, and Y.-E. Liu. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR, 2017.

E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

D. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020a.

D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020b.

V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15): e2014602118, 2021.

A. Hassidim, R. Kupfer, and Y. Singer. An optimal elimination algorithm for learning a best arm. *Advances in Neural Information Processing Systems*, 33:10788–10798, 2020.

Y. Jin. Upper bounds on the natarajan dimensions of some function classes. *arXiv preprint arXiv:2209.07015*, 2022.

Y. Jin, Z. Ren, Z. Yang, and Z. Wang. Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

A. Krishnamurthy, A. Agarwal, T.-K. Huang, H. Daumé III, and J. Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

S. K. Krishnamurthy, V. Hadad, and S. Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*, pages 5805–5814. PMLR, 2021.

S. K. Krishnamurthy, A. Propp, and S. Athey. Towards costless model selection in contextual bandits: A bias-variance perspective. *arXiv preprint arXiv:2106.06483*, 2023.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Z. Li, L. Ratliff, K. G. Jamieson, L. Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

M. Majzoubi, C. Zhang, R. Chari, A. Krishnamurthy, J. Langford, and A. Slivkins. Efficient contextual bandits with continuous actions. *Advances in Neural Information Processing Systems*, 33:349–360, 2020.

W. Mou, P. Ding, M. J. Wainwright, and P. L. Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency. *arXiv preprint arXiv:2301.06240*, 2023.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

M. Offer-Westort, A. Coppock, and D. P. Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4):826–844, 2021.

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. *Advances in Neural Information Processing Systems*, 24, 2011.

D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9 (3), 2008.

D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN*, 2020.

A. Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

J. Yao, E. Brunskill, W. Pan, S. Murphy, and F. Doshi-Velez. Power constrained bandits. In *Machine Learning for Healthcare Conference*, pages 209–259. PMLR, 2021.

A. Zanette, K. Dong, J. N. Lee, and E. Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34:22720–22731, 2021.

R. Zhan, V. Hadad, D. A. Hirshberg, and S. Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135, 2021.

Y. Zhu and P. Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, pages 27574–27590. PMLR, 2022.

A Expanded Notations

We start with expanding our notation from Section 1.1 to include notation helpful for our proofs and expand to the continuous arm setting.

Measure over arms. To recap, our algorithm and analysis adapt to both discrete and continuous arm spaces, where we consider a finite measure space $(\mathcal{A}, \Sigma, \mu)$ over the set of arms (i.e. $\mu(\mathcal{A})$ is finite) to unify the notation.[14] As short hand, we use K in lieu of $\mu(\mathcal{A})$. We let Σ_1 be a set of arms in Σ with measure one, i.e. $\Sigma_1 := \{S \in \Sigma | \mu(S) = 1\}$.

Policies. Let $\tilde{\Pi}$ denote the universal set of policies. That is, $\tilde{\Pi}$ is the set of all functions from $\mathcal{X}$ to Σ_1. The policy class Π is a subset of $\tilde{\Pi}$. We use $\pi(x)$ to denote the set of arms given $x \in \mathcal{X}$ and use $p_\pi(a|x) = \mathbb{I}(a \in \pi(x))$ to denote the induced probability measure over arms at x.[15] With some abuse of notation, we use the notation $\pi(a|x)$ in lieu of $p_\pi(a|x)$. Below is the elaboration of our notation to both discrete and continuous arm spaces.

- *Discrete arm space.* We choose μ to be the count-measure, where $\mu(S) = |S|$ for any $S \subseteq \mathcal{A}$ and $\mu(\mathcal{A}) = K$. In this case, Σ_1 contains singleton arm sets, and $\tilde{\Pi}$ denotes deterministic policies from $\mathcal{X}$ to $\mathcal{A}$ where each policy maps a context to an action.
- *Continuous arm space.* We choose μ to any finite measure, where $\mu(S) = \int_S d\mu(a)$ for any $S \subseteq \mathcal{A}$, and in particular $\mu(\mathcal{A}) = K$. In this case, Σ_1 contains arm sets that may have an infinite number of arms but with total measure be 1 with respect to μ.

Space of action selection kernels. In this paper, we will always define our action selection kernels with respect to the reference measure μ, that is $p(S|x) = \int_{a \in S} p(a|x) d\mu$ for any $S \in \Sigma$ and $x \in \mathcal{X}$. Based on the notation in Section 1.1, for any kernel p, we let $\text{Reg}_f(p) = R_f(\pi_f) - R_f(p)$.

Now let $\mathcal{P}$ denote the set of action selection kernels such that $p(a|x) \leq 1$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, and in particular, we have the policy class $\tilde{\Pi} \subset \mathcal{P}$. We note that all action selection kernels (p) considered in this paper belong to the set $\mathcal{P}$, allowing our analysis to rely on the fact that $p(\cdot|\cdot) \leq 1$.

Note that, $\text{Reg}_f(p)$ is non-negative for any $p \in \mathcal{P}$. To see this, consider any context x. Recall that $\pi_f(x) \in \arg\max_{S \in \Sigma_1} f(x, S)$ for all x. Since $p(\cdot|\cdot) \leq 1$ for any $p \in \mathcal{P}$, we have $\int_{a \in \mathcal{A}} p(a|x) f(x, a) d\mu$ is maximized when $p(a|x) = 1$ for all $a \in \pi_f(x)$. That is, $f(x, \pi_f(x)) = \max_{p \in \mathcal{P}} \mathbb{E}_{a \sim p(\cdot|x)}[f^*(x, a)]$. Hence, $\max_{p \in \mathcal{P}} R_f(p) = R_f(\pi_f)$, so $\text{Reg}_f(p)$ is non-negative for any $p \in \mathcal{P}$.

Connection to smooth regret Zhu and Mineiro [2022]. Recall that we define cumulative regret as $\text{CReg}_T := \sum_{t=1}^T \text{Reg}_{f^*}(p_t)$, which measures regret w.r.t the benchmark $R_{f^*}(\pi_{f^*}) = \max_{p \in \mathcal{P}} R_{f^*}(p)$. As discussed earlier, Zhu and Mineiro [2022] shows that smooth regret bounds are stronger than several other definitions of cumulative regret in the continuous arm setting [e.g., Majzoubi et al., 2020]. Hence to show that our bounds are comparable/competitive for the continuous arm setting, we argue that our definition of cumulative regret (CReg_T) is equivalent to the definition of smooth regret in Zhu and Mineiro [2022].

Let the loss vectors l_t in Zhu and Mineiro [2022] be given by $-r_t$. Let the smoothness parameter h in Zhu and Mineiro [2022] be given by $1/K$. And, let the base probability measure in Zhu and Mineiro [2022] be given by μ/K. Then, our benchmark ($\max_{p \in \mathcal{P}} R_{f^*}(p)$) is equal to the smooth benchmark ($\mathbb{E}[\text{Smooth}_h(x)]$) considered in Zhu and Mineiro [2022]. Hence, our definition of cumulative regret (CReg_T) is equal to smooth regret ($\text{Reg}_{\text{CB},h}(T)$) when the loss, smoothness parameter, and base probability measure are given as above. This shows the equivalence in our definitions.

Hence our near-optimal cumulative regret bounds (with $\omega = 1$) recover several existing results for the stochastic contextual bandit setting up to logarithmic factors using only offline regression oracles. Our algorithm also handles reward model misspecification and does not assume realizability. We also provide instance-dependent simple regret bounds (for larger choices of ω). The parameter ω allows us to trade-off between simple and cumulative regret bounds.

Measure theoretic issues with continuous arms. To avoid measure-theoretic issues, we require that for all models $f \in \mathcal{F} \cup \{f^*\}$, all contexts $x \in \mathcal{X}$, and all real numbers $z \in \mathbb{R}$, we have the level set

[14] Here Σ is a σ-algebra over $\mathcal{A}$ and μ is a bounded set function from Σ to the real line.

[15] Note that for any $\pi \in \tilde{\Pi}$, we have $\int_{a \in \mathcal{A}} p_\pi(a|x) d\mu = \mu(\pi(x)) = 1$ at any $x \in \mathcal{X}$.

	Environment distribution
$\mathcal{X}, \mathcal{A}$	set of contexts and set of arms (respectively).
D	joint distribution over contexts and arm rewards.
$D_{\mathcal{X}}$	marginal distribution over contexts.
$D(p)$	distribution over $\mathcal{X} \times \mathcal{A} \times [0,1]$ induced by action selection kernel p.
f^*	true conditional expected reward, $f^*(x,a) := \mathbb{E}_D[r_t(a)\|x_t = x]$.
	Measure space over arm sets
$(\mathcal{A}, \Sigma, \mu)$	measure space over the set of arms with $K := \mu(\mathcal{A})$.
Σ_1	set of measurable arm sets with measure one, $\Sigma_1 := \{S \in \Sigma \| \mu(S) = 1\}$.
$\tilde{\Pi}$	set of all policies (functions from $\mathcal{X}$ to Σ_1).
$\mathcal{P}$	set of kernels such that $p(a\|x) \leq 1$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.
	Algorithm inputs
$\beta_{\max}$	proportional response threshold ($\beta_{\max} = 1/2$).
ω	trade-off parameter ($\omega \in [1, K]$).
δ	confidence parameter ($\delta \in (0,1)$).
$\mathcal{F}, \Pi$	give reward model class and policy class.
	Policy value and optimal policy
$R_f(p)$	$:= \mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim p(\cdot\|x)}[f(x,a)]$, denotes the value of kernel p under model f.
$R(p)$	$:= R_{f^*}(p)$, denotes the true value of kernel p.
π_f	$\in \arg\max_{\pi \in \tilde{\Pi}} R_f(\pi)$, denotes the best universal policy under model f.
π^*	$\in \arg\max_{\pi \in \Pi} R(\pi)$, denotes the best policy in the class Π.
	Regret and cover
$\mathrm{Reg}_f(p)$	$:= R_f(\pi_f) - R_f(p)$, denotes the regret of kernel p under model f.
$R(p)$	$:= R_{f^*}(p)$, denotes the true value of kernel p.
$\mathrm{Reg}_\Pi(p)$	$:= R_f(\pi_f) - R_f(p)$, denotes the regret of kernel p under model f.
$V(p,q)$	cover $V(p,q) := \mathbb{E}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot\|x)} \left[q(a\|x)/p(a\|x)\right]$.
	Epochs
m	epoch index.
ξ_{m+1}	$:= 2\xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3))$, where ξ is given in Section 1.2.
m^*	safe epoch, last epoch where $\xi_{m+1} \geq 2B$ (variance dominates bias).
$\hat{m}$	last epoch that starts with **safe** set as **True**.
	Algorithmic parameters
$\hat{f}_{m+1} \in \mathcal{F}$	fitted reward model via regressions on samples collected in epoch m.
C_{m+1}	conformal arm set defined in Definition 2.
η_{m+1}	risk adjustment parameter (5).
α_{m+1}	empirical bound on optimal cover ($\alpha_{m+1} = \frac{3K}{\eta_{m+1}}$).
p_{m+1}	action selection kernel corresponding to epoch $m+1$ defined in (4).
U_{m+1}	$:= 20\sqrt{\alpha_m \xi_{m+1}}$.

Table 1: Table of notations

of arms $\{a|f(x,a) \leq z\}$ must lie in Σ. That is the reward models $f \in \mathcal{F} \cup \{f^*\}$ are measurable at every context x with the Lebesgue measure on the range of $f(x,\cdot)$ and the measure $(\mathcal{A}, \Sigma, \mu)$ on the domain of $f(x,\cdot)$. We note that this isn't a strong condition and usually trivially holds.

Moreover, we require an additional condition as follows to simplify our arguments and allow for easy construction of our uncertainty sets (see Definition 2). We require that for all models $f \in \mathcal{F}$ and all contexts $x \in \mathcal{X}$, we have $f(x, \pi_f(x))$ is equal to $\max_{a \in \mathcal{A}} f(x,a)$. This condition trivially holds for the finite-arm setting with μ as a count measure. For the continuous arm setting, this condition follows from requiring $\arg\max_{a \in \mathcal{A}} f(x,a)$ lies in Σ and has measure of at least one.

Additional notation. For notational convenience, we let $U_m = 20\sqrt{\alpha_{m-1}\xi_m}$ for any epoch m. Note that by construction ((5) and $\alpha_m = 3K/\eta_m$) α_m is non-increasing in m. Further, from the conditions in Section 1.2, we have $\xi_{m+1} = 2\xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3))$ is non-increasing in m. Hence U_m is also non-increasing in m. We also let $\alpha_0 := \alpha_1 = 3K$, and let $\alpha_m := \alpha_{\hat{m}}$ for any epoch $m \geq \hat{m}$. Similarly, we let $\eta_0 := \eta_1 = 1$, and let $\eta_m := \eta_{\hat{m}}$ for any epoch $m \geq \hat{m}$. Sometimes, we use use $C_m(x, \beta, \eta)$ in lieu of $C_m(x, \beta/\eta)$.

B Bounding Cumulative Regret

This section derives the cumulative regret bounds for ω-RAPR. We start with analyzing the output of oracles described in Assumptions 1 and 2. Note that we do not make the "realizability" assumption in this work – i.e., we do not assume that f^* lies in $\mathcal{F}$. Hence, as in Assumption 1, the expected squared error of our estimated models need not go to zero (even with infinite data) and may contain an unknown non-zero irreducible error term (B) that captures the bias of the model class $\mathcal{F}$. It is useful to split our analysis into two regimes to handle this unknown term B, similar to the approach in Krishnamurthy et al. [2021]. In particular, we separately analyze oracle outputs for epochs before and after a so-called "safe epoch". Where we define the safe epoch m^* as the epoch where the variance of estimating from the model class $\mathcal{F}$ (ξ_m) is dominated by the bias of estimating from the class $\mathcal{F}$ (B). That is, $m^* := \arg\max\{m \geq 1 | \xi_{m+1} \geq 2B\}$.

B.1 High Probability Events

We start with defining high-probability events under which our key theoretical guarantees hold. The first high probability event characterizes the accuracy of estimated reward models and the policy evaluation estimators, the tail bound of which can be obtained by taking a union bound of each epoch-specific event that happens with probability $1 - \frac{\delta}{4m^2}$ under assumptions in Section 1.2.

Lemma 1. *Suppose Assumption 1 and Assumption 2 hold. The following event holds with probability* $1 - \delta/2$*,*

$$\begin{aligned}\mathcal{W}_1 := \Bigg\{ & \forall m, \forall \pi \in \Pi \cup \{p_{m+1}\}, \forall f \in \{\hat{f}_1, \ldots, \hat{f}_{m+1}\}, \\ & \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \leq B + \xi_{m+1}/2 \\ & |\hat{R}_{m+1,f}(\pi) - R_f(\pi)| \leq \sqrt{\xi_{m+1}}, \\ & |\hat{R}_{m+1}(\pi) - R(\pi)| \leq \sqrt{V(p_m, \pi)\xi_{m+1}} + \xi_{m+1}/\Big(\min_{(x,a) \in \mathcal{X} \times \mathcal{A}} p_m(a|x)\Big).\Bigg\}.\end{aligned} \tag{9}$$

Where $\xi_{m+1} = 2\xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3))$.[16]

The second high probability event characterizes the measure of conformal arm sets, which directly follows from Hoeffding's inequality and union bound.

Lemma 2. *The following event holds with probability* $1 - \delta/2$*,*

$$\begin{aligned}\mathcal{W}_2 := \Bigg\{ & \forall m, \forall \eta \in \Big\{ \frac{|S_{m,2}|}{n} | n \in [|S_{m,2}|] \Big\}, \\ & \Bigg| \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \left[\mu\left(\bar{C}_{m+1}\left(x, \frac{\beta_{\max}}{\eta} \right) \right) \right] - \frac{1}{|S_{m,2}|} \sum_{t \in S_{m,2}} \mu\left(\bar{C}_{m+1}\left(x_t, \frac{\beta_{\max}}{\eta} \right) \right) \Bigg| \\ & \leq \sqrt{\frac{K^2 \ln(8|S_{m,2}|m^2/\delta)}{2|S_{m,2}|}} \Bigg\}.\end{aligned} \tag{10}$$

Together both $\mathcal{W}_1$ and $\mathcal{W}_2$ hold with probability $1 - \delta$. The rest of our analysis works under these events.

B.2 Analyzing the Cover

In this sub-section, we upper bound the cover ($V(p_m, \cdot)$) for the action selection kernel used in epoch m.[17] To upper bound $V(p_m, q)$, we first lower bound $p_m(\cdot|\cdot)$. Recall that in Appendix A, we define

[16] Our epoch schedules will always be increasing in epoch length. Under such conditions, we have ξ_m is non-increasing in m.

[17] Recall that $V(p, q) := \mathbb{E}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)}[q(a|x)/p(a|x)]$.

$U_m = 20\sqrt{\alpha_{m-1}\xi_m}$ for any epoch m. Starting from here, our lemmas and proofs will use U_m and $20\sqrt{\alpha_{m-1}\xi_m}$ interchangeably.

Lemma 3. *For any epoch m, we have (11) holds.*

$$
\begin{aligned}
p_m(a|x) &\geq \begin{cases} \frac{1-\beta_{\max}}{\mu(C_m(x,\beta_{\max},\eta_m))} + \frac{\beta_{\max}}{\mu(\mathcal{A})}, & \text{if } a \in C_m(x,\beta_{\max},\eta_m) \\ \frac{\eta_m}{\mu(\mathcal{A})} \min_{\bar{m}\in[m]} \frac{2\bar{m}^2 U_{\bar{m}}}{\hat{f}_{\bar{m}}(x,\pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)}, & \text{if } a \notin C_m(x,\beta_{\max},\eta_m) \end{cases} \\
&\geq \begin{cases} \frac{1}{\mu(\mathcal{A})}, & \text{if } a \in C_m(x,\beta_{\max},\eta_m) \\ \frac{\eta_m \min_{\bar{m}\in[m]} 2\bar{m}^2 U_{\bar{m}}}{\mu(\mathcal{A})}, & \text{if } a \notin C_m(x,\beta_{\max},\eta_m) \end{cases}
\end{aligned} \tag{11}
$$

Proof. Recall that p_m given by (12).

$$
p_m(a|x) = \frac{(1-\beta_{\max}) I[a \in C_m(x,\beta_{\max},\eta_m)]}{\mu\big(C_m(x,\beta_{\max},\eta_m)\big)} + \int_0^{\beta_{\max}} \frac{I[a \in C_m(x,\beta,\eta_m)]}{\mu\big(C_m(x,\beta,\eta_m)\big)} \mathrm{d}\beta. \tag{12}
$$

We divide our analysis into two cases based on whether a lies in $C_m(x,\beta_{\max},\eta_m)$, and lower bound $p_m(a|x)$ in each case.

Case 1 ($a \in C_m(x,\beta_{\max},\eta_m)$). Note that $C_m(x,\beta_{\max},\eta_m) \subseteq C_m(x,\beta,\eta_m) \subseteq \mathcal{A}$ for all $\beta \in [0,\beta_{\max}]$. Hence, $a \in C_m(x,\beta,\eta_m)$ and $\mu(C_m(x,\beta,\eta_m)) \leq \mu(\mathcal{A})$ for all $\beta \in [0,\beta_{\max}]$. Therefore, in this case, $p_m(a|x) \geq \frac{(1-\beta_{\max})}{\mu\left(C_m(x,\beta_{\max},\eta_m)\right)} + \frac{\beta_{\max}}{\mu(\mathcal{A})} \geq \frac{1}{\mu(\mathcal{A})}$.

Case 2 ($a \notin C_m(x,\beta_{\max},\eta_m)$). For this case, the proof follows from (13).

$$
\begin{aligned}
p_m(a|x) &\geq \int_0^{\beta_{\max}} \frac{I[a \in C_m(x,\beta,\eta_m)]}{\mu\big(C_m(x,\beta,\eta_m)\big)} \mathrm{d}\beta \\
&\overset{(i)}{\geq} \frac{1}{\mu(\mathcal{A})} \int_0^{\beta_{\max}} I[a \in C_m(x,\beta,\eta_m)] \mathrm{d}\beta \\
&\overset{(ii)}{\geq} \frac{I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \int_0^{\beta_{\max}} I[a \in C_m(x,\beta,\eta_m)] \mathrm{d}\beta \\
&\overset{(iii)}{=} \frac{I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \int_0^{1} I[a \in C_m(x,\beta,\eta_m)] \mathrm{d}\beta \\
&\overset{(iv)}{=} \frac{I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \int_0^{1} \prod_{\bar{m}\in[m]} I\big[\hat{f}_{\bar{m}}(x,\pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a) \leq \frac{2\bar{m}^2 \eta_m U_{\bar{m}}}{\beta}\big] \mathrm{d}\beta \\
&= \frac{I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \int_0^{1} \prod_{\bar{m}\in[m]} I\big[\beta \leq \frac{2\bar{m}^2 \eta_m U_{\bar{m}}}{\hat{f}_{\bar{m}}(x,\pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)}\big] \mathrm{d}\beta \\
&= \frac{\eta_m I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \min_{\bar{m}\in[m]} \frac{2\bar{m}^2 U_{\bar{m}}}{\hat{f}_{\bar{m}}(x,\pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)} \\
&\overset{(v)}{\geq} \frac{\eta_m I(a \notin C_m(x,\beta_{\max},\eta_m))}{\mu(\mathcal{A})} \min_{\bar{m}\in[m]} 2\bar{m}^2 U_{\bar{m}}
\end{aligned} \tag{13}
$$

where (i) is because the measure of the conformal set C_m can be no larger than the measure of the action space $\mathcal{A}$; (ii) follows from $I(a \notin C_m(x,\beta_{\max},\eta_m)) \leq 1$; (iii) follows from the fact that if $a \notin C_m(x,\beta_{\max},\eta_m)$ then $a \notin C_m(x,\beta,\eta_m)$ for all $\beta \geq \beta_{\max}$; (iv) follows from Definition 2; and (v) follows from .[18] □

Using Lemma 3, we get an upper bound on $V(p_m,q)$ in terms of $\mathbb{E}[\mu(C_m(x,\beta_{\max},\eta_m))]$, K/η_m, and expected regret with respect to the models $\hat{f}_1,\ldots,\hat{f}_m$.

[18] Note that if $a \in \pi_{\hat{f}_m}(x)$ then $I(a \notin C_m(x,\beta_{\max},\eta_m)) = 0$.

Lemma 4. *For any epoch m and any action selection kernel $q \in \mathcal{P}$, we have (14) holds.*

$$V(p_m, q) \leq \frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1 - \beta_{\max}} + \frac{K}{\eta_m} \sum_{\bar{m} \in [m]} \frac{Reg_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2 U_{\bar{m}}}. \tag{14}$$

Proof. From Lemma 3, we have (15) holds. [19]

$$\begin{aligned} &\frac{I(a \in C_m(x, \beta_{\max}, \eta_m))}{p_m(a|x)} \leq \frac{\mu\big(C_m(x, \beta_{\max}, \eta_m)\big)}{1 - \beta_{\max}}, \\ \text{and} \quad &\frac{I(a \notin C_m(x, \beta_{\max}, \eta_m))}{p_m(a|x)} \leq \frac{\mu(\mathcal{A})}{\eta_m} \max_{\bar{m} \in [m]} \frac{\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x, a)}{2\bar{m}^2 U_{\bar{m}}}. \end{aligned} \tag{15}$$

We now bound the cover $V(p_m, q)$ as follows,

$$\begin{aligned} V(p_m, q) &= \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)} \left[\frac{q(a|x)}{p_m(a|x)} \right] \\ &\overset{(i)}{\leq} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)} \left[\frac{1}{p_m(a|x)} \right] \\ &= \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)} \left[\frac{I[a \in C_m(x, \beta_{\max}, \eta_m)] + I[a \notin C_m(x, \beta_{\max}, \eta_m)]}{p_m(a|x)} \right] \\ &\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)} \left[\frac{\mu\big(C_m(x, \beta_{\max}, \eta_m)\big)}{1 - \beta_{\max}} \right] + \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}, a \sim q(\cdot|x)} \left[\frac{\mu(\mathcal{A})}{\eta_m} \max_{\bar{m} \in [m]} \frac{\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x, a)}{2\bar{m}^2 U_{\bar{m}}} \right] \\ &\leq \frac{\mathbb{E}[\mu\big(C_m(x, \beta_{\max}, \eta_m)\big)]}{1 - \beta_{\max}} + \frac{\mu(\mathcal{A})}{\eta_m} \sum_{\bar{m} \in [m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2 U_{\bar{m}}}. \end{aligned} \tag{16}$$

Here (i) follows from the fact that $q \in \mathcal{P}$ and (ii) follows from (15). □

Having bounded the cover for the kernel p_m in terms of $\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]$ and K/η_m. We now bound these terms with α_m.

Lemma 5. *Suppose $\mathcal{W}_2$ holds. Then for any epoch m, we have (17) holds.*

$$\frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1 - \beta_{\max}} + \frac{K}{\eta_m} \leq \alpha_m \leq \frac{3K}{\eta_m}. \tag{17}$$

Proof. Since $\mu(C_m(x, \beta_{\max}, \eta_m)) \leq K$ and $\beta_{\max} = 1/2$, the bound trivially holds if $\eta_m = 1$. Suppose $\eta_m > 1$. Note that $\eta_{\bar{m}}$ is non-decreasing in $\bar{m}$ by construction. Let m' be the smallest epoch index such that $\eta_{m'} = \eta_m$. We now have the following holds.

$$\begin{aligned} &\frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1 - \beta_{\max}} + \frac{K}{\eta_m} \\ &\overset{(i)}{\leq} \frac{\min\{1 + \mathbb{E}[\mu(\bar{C}_m(x, \beta_{\max}, \eta_m))], K\}}{1 - \beta_{\max}} + \frac{K}{\eta_m} \\ &\overset{(ii)}{\leq} \frac{\min\{1 + \mathbb{E}[\mu(\bar{C}_{m'}(x, \beta_{\max}, \eta_m))], K\}}{1 - \beta_{\max}} + \frac{K}{\eta_m} \\ &\overset{(iii)}{\leq} \frac{\lambda_{m'}(\eta_m)}{1 - \beta_{\max}} + \frac{K}{\eta_m} \\ &\overset{(iv)}{\leq} 2\lambda_{m'}(\eta_m) + \frac{K}{\eta_m} \\ &\overset{(v)}{\leq} \frac{3K}{\eta_m} \end{aligned} \tag{18}$$

[19] Note that we require $f(x, \pi_f(x)) \geq f(x, a)$, for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $f \in \mathcal{F}$.

Here (i) follows from the definition of CASs. (ii) follows from $\bar{C}_m \subseteq \bar{C}_{m'}$ which follows from the fact that $\bar{C}_m = \cap_{\bar{m}\in[m]}\tilde{C}_{\bar{m}}$ and $m' \leq m$. (iii) follows from $\mathcal{W}_2$ and the definition of $\lambda_{m'}$ in (5). (iv) follows from the fact that $\beta_{\max} = 1/2$. (v) follows from (5) and $\eta_{m'-1} < \eta_{m'} = \eta_m$ – note that $\eta_{m'-1} \neq \eta_{m'}$ gives us that $\eta_{m'}$ was set using the constrained maximization procedure in (5), hence the constraint $\lambda_{m'}(\eta) \leq K/\eta$ is satisfied at $\eta = \eta_{m'} = \eta_m$. □

B.3 Evaluation Guarantees Under Safe Epoch

This sub-section provides guarantees on how accurate $R_{\hat{f}_{m+1}}$ is at evaluating policies when we are within the safe epoch.

Lemma 6. *Suppose $\mathcal{W}_1$ holds. For all epochs $m \in [m^*]$, for any $q \in \mathcal{P}$, we have,*

$$|R_{\hat{f}_{m+1}}(q) - R(q)| \leq \sqrt{V(p_m, q)\xi_{m+1}}. \tag{19}$$

Proof. Consider any epoch $m \in [m^*]$ and policy $q \in \mathcal{P}$. We then have,

$$\begin{aligned}
|R_{\hat{f}_{m+1}}(q) - R(q)| &= |\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim q}[\hat{f}_{m+1}(x,a) - f^*(x,a)]| \\
&\overset{(i)}{=} \left|\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[\frac{q(a|x)}{p_m(a|x)}\big(\hat{f}_{m+1}(x,a) - f^*(x,a)\big)\right]\right| \\
&\leq \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[\frac{q(a|x)}{p_m(a|x)}\big|\hat{f}_{m+1}(x,a) - f^*(x,a)\big|\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[\sqrt{\Big(\frac{q(a|x)}{p_m(a|x)}\Big)^2\big|\hat{f}_{m+1}(x,a) - f^*(x,a)\big|^2}\right] \\
&\overset{(ii)}{\leq} \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[\Big(\frac{q(a|x)}{p_m(a|x)}\Big)^2\right]}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[(\hat{f}_{m+1}(x,a) - f^*(x,a))^2\right]} \\
&\overset{(iii)}{=} \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim q}\left[\frac{q(a|x)}{p_m(a|x)}\right]}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim p_m}\left[(\hat{f}_{m+1}(x,a) - f^*(x,a))^2\right]} \\
&\overset{(iv)}{\leq} \sqrt{V(p_m, q)\xi_{m+1}},
\end{aligned} \tag{20}$$

where (i) and (iii) follow from change of measure arguments, (ii) follows from Cauchy-Schwartz inequality, and (iv) follows from $\mathcal{W}_1$. □

By combining the guarantees of Lemma 4 and Lemma 6, we get Lemma 7.

Lemma 7. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Then for any action selection kernel $q \in \mathcal{P}$, we have:*

$$|R_{\hat{f}_{m+1}}(q) - R(q)| \leq \sqrt{\alpha_m\xi_{m+1}} + \frac{1}{2}\sqrt{\alpha_m\xi_{m+1}}\sum_{\bar{m}\in[m]}\frac{Reg_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2U_{\bar{m}}}. \tag{21}$$

Proof. From Lemma 4, we have (22) holds for any $q \in \mathcal{P}$.

$$\begin{aligned}
&V(p_m, q) \\
&\leq \frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1-\beta_{\max}} + \frac{K}{\eta_m}\sum_{\bar{m}\in[m]}\frac{\text{Reg}_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2U_{\bar{m}}} \\
&\leq \Big(\frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1-\beta_{\max}} + \frac{K}{\eta_m}\Big) + \Big(\frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1-\beta_{\max}} + \frac{K}{\eta_m}\Big)\sum_{\bar{m}\in[m]}\frac{\text{Reg}_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2U_{\bar{m}}} \\
&= \alpha_m + \alpha_m\sum_{\bar{m}\in[m]}\frac{\text{Reg}_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2U_{\bar{m}}}
\end{aligned} \tag{22}$$

Combining (22) with Lemma 6 we have:

$$\begin{aligned}
&|R_{\hat{f}_{m+1}}(q) - R(q)| \\
\leq &\sqrt{V(p_m, q)\xi_{m+1}} \\
\overset{(i)}{\leq} &\frac{1}{2}\sqrt{\alpha_m \xi_{m+1}} + \frac{1}{2}\sqrt{\frac{\xi_{m+1}}{\alpha_m}} V(p_m, q) \\
\overset{(i)}{\leq} &\sqrt{\alpha_m \xi_{m+1}} + \frac{1}{2}\sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m}\in[m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(q)}{2\bar{m}^2 U_{\bar{m}}}.
\end{aligned} \tag{23}$$

Where (i) follows from AM-GM inequality and (ii) follows from (22). □

B.4 Testing Safety

The misspecification test (2) is designed to test if we are within the safe epoch. In principle, it works by comparing the accuracy of $R_{\hat{f}_{m+1}}$ (Lemma 7) and $\hat{R}_{m+1}$ (Lemma 8). Formally, Lemma 11 shows that the misspecification test in (2) fails only after m^*. Hence, $\hat{m} \geq m^* + 1$. Lemma 12 then describes the implication of (2) continuing to hold. In what follows, we let $\widehat{\mathrm{Reg}}_{m+1,\hat{f}_{\bar{m}}}(\pi) := \hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi)$. We start with Lemma 8 which provides accuracy guarantees for $\hat{R}_{m+1}$ in any epoch.

Lemma 8. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Then for any epoch m and all $\pi \in \Pi \cup \{p_{m+1}\}$, we have,*

$$|\hat{R}_{m+1}(\pi) - R(\pi)| \leq \sqrt{\alpha_m \xi_{m+1}} + \frac{1}{2}\sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m}\in[m]} \frac{\mathit{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + \frac{K\xi_{m+1}}{\eta_m \min_{\bar{m}\in[m]} U_{\bar{m}}}. \tag{24}$$

Proof. From Lemma 3, we have (25) holds, which provides a worst-case lower bound on p_m.

$$\min_{(x,a)\in\mathcal{X}\times\mathcal{A}} p_m(a|x) \geq \min\left(\frac{1}{K}, \frac{\eta_m \min_{\bar{m}\in[m]}(2\bar{m}^2)U_{\bar{m}}}{K}\right) \geq \frac{\eta_m \min_{\bar{m}\in[m]} U_{\bar{m}}}{K} \tag{25}$$

Where the last inequality follows from $U_m \leq 1$. Now from $\mathcal{W}_1$, we have,

$$\begin{aligned}
|\hat{R}_{m+1}(\pi) - R(\pi)| &\overset{(\mathcal{W}_1)}{\leq} \sqrt{V(p_m, \pi)\xi_{m+1}} + \frac{K\xi_{m+1}}{\eta_m \min_{\bar{m}\in[m]} U_{\bar{m}}} \\
&\overset{(i)}{\leq} \frac{1}{2}\sqrt{\alpha_m \xi_{m+1}} + \frac{1}{2}\sqrt{\frac{\xi_{m+1}}{\alpha_m}} V(p_m, \pi) + \frac{K\xi_{m+1}}{\eta_m \min_{\bar{m}\in[m]} U_{\bar{m}}} \\
&\overset{(ii)}{\leq} \sqrt{\alpha_m \xi_{m+1}} + \frac{1}{2}\sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m}\in[m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + \frac{K\xi_{m+1}}{\eta_m \min_{\bar{m}\in[m]} U_{\bar{m}}}.
\end{aligned} \tag{26}$$

Where (i) follows from AM-GM inequality, and (ii) follows from (22) in the proof of Lemma 7. □

Lemmas 9 and 10 provide useful inequalities that help construct the misspecification test (2).

Lemma 9. *For any epoch m, policy $\pi \in \mathcal{P}$, and model $f \in \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_{m+1}\}$. We have,*

$$\begin{aligned}
&||R_f(\pi) - R(\pi)| - |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)|| \\
&\leq |\hat{R}_{m+1}(\pi) - R(\pi)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)|.
\end{aligned}$$

Proof. The proof follows from noting that,

$$\begin{aligned}
&|R_f(\pi) - R(\pi)| \\
&= |\hat{R}_{m+1}(\pi) - R(\pi) + R_f(\pi) - \hat{R}_{m+1,f}(\pi) + \hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)| \\
&\leq |\hat{R}_{m+1}(\pi) - R(\pi)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)| + |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)|.
\end{aligned} \tag{27}$$

and from noting that,

$$\begin{aligned}
&|\hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi)| \\
&= |\hat{R}_{m+1}(\pi) - R(\pi) + R_f(\pi) - \hat{R}_{m+1,f}(\pi) + R(\pi) - R_f(\pi)| \\
&\leq |\hat{R}_{m+1}(\pi) - R(\pi)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)| + |R_f(\pi) - R(\pi)|.
\end{aligned} \tag{28}$$

□

Lemma 10. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Then for any epoch m, any model $f \in \{\hat{f}_i | i \in [m+1]\}$, and any policy $\pi \in \Pi \cup \{p_{m+1}\}$, we have,*

$$|\mathit{Reg}_f(\pi) - \widehat{\mathit{Reg}}_{m+1,f}(\pi)| \leq 2\sqrt{\xi_{m+1}}$$

Proof. Follows from triangle inequality and $\mathcal{W}_1$,

$$\begin{aligned}
&|\mathrm{Reg}_f(\pi) - \widehat{\mathrm{Reg}}_{m+1,f}(\pi)| \\
&\leq |R_f(\pi_f) - \hat{R}_{m+1,f}(\pi_f)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)| \leq 2\sqrt{\xi_{m+1}}
\end{aligned} \tag{29}$$

□

As discussed earlier, Lemma 11 shows that the misspecification test in (2) fails only after m^*. Hence, $\hat{m} \geq m^* + 1$.

Lemma 11. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Now for any epoch $m \in [m^*]$ we have that,*

$$\begin{aligned}
&\max_{\pi \in \Pi \cup \{p_{m+1}\}} |\hat{R}_{m+1,\hat{f}_{m+1}}(\pi) - \hat{R}_{m+1}(\pi)| - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\widehat{\mathit{Reg}}_{m+1,\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} \\
&\leq 2.05\sqrt{\alpha_m \xi_{m+1}} + 1.1\sqrt{\xi_{m+1}}.
\end{aligned} \tag{30}$$

Proof. For any epoch $m \in [m^*]$ and for any $\pi \in \Pi \cup \{p_{m+1}\}$, we have,

$$\begin{aligned}
&|\hat{R}_{\hat{f}_{m+1}}(\pi) - \hat{R}_{m+1}(\pi)| \\
&\overset{(i)}{\leq} |\hat{R}_{m+1}(\pi) - R(\pi)| + |R_{\hat{f}_{m+1}}(\pi) - R(\pi)| + |R_{\hat{f}_{m+1}}(\pi) - \hat{R}_{\hat{f}_{m+1}}(\pi)| \\
&\overset{(ii)}{\leq} 2\sqrt{\alpha_m \xi_{m+1}} + \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + \frac{K\xi_{m+1}}{\eta_m \min_{\bar{m} \in [m]} U_{\bar{m}}} + \sqrt{\xi_{m+1}} \\
&\overset{(iii)}{\leq} 2\sqrt{\alpha_m \xi_{m+1}} + \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + \frac{\alpha_m \xi_{m+1}}{U_m} + \sqrt{\xi_{m+1}} \\
&\overset{(iv)}{\leq} 2\sqrt{\alpha_m \xi_{m+1}} + \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\widehat{\mathrm{Reg}}_{m+1,\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + \frac{2\sqrt{\alpha_m}\xi_{m+1}}{U_m} + \frac{\alpha_m \xi_{m+1}}{U_m} + \sqrt{\xi_{m+1}} \\
&\overset{(v)}{\leq} 2.05\sqrt{\alpha_m \xi_{m+1}} + \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\widehat{\mathrm{Reg}}_{m+1,\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + 1.1\sqrt{\xi_{m+1}}
\end{aligned} \tag{31}$$

Where (i) follows from Lemma 9. (ii) follows from Lemma 7, Lemma 8, and $\mathcal{W}_1$. (iii) follows from Equation (17) and the fact that $U_{\bar{m}}$ is non-increasing in $\bar{m}$ (giving us $\min_{\bar{m} \in [m]} U_{\bar{m}} = U_m$). (iv) follows from Lemma 10, the fact that $U_{\bar{m}}$ is non-increasing in $\bar{m}$, and the fact that $\sum_{\bar{m}=1}^{\infty} 1/(2\bar{m}^2) \leq 1$. (v) follows from $U_m = 20\sqrt{\alpha_{m-1}\xi_m}$, $\alpha_m \leq \alpha_{m-1}$, and $\xi_{m+1} \leq \xi_m$. □

Lemma 12 now describes the implication of (2) continuing to hold.

Lemma 12. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Now for any epoch $m \in [\hat{m} - 1]$ and any policy $\pi \in \Pi \cup \{p_{m+1}\}$, we then have that,*

$$|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \leq 2.2\sqrt{\xi_{m+1}} + 3.1\sqrt{\alpha_m \xi_{m+1}} + \frac{3}{2}\sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\mathit{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}}. \tag{32}$$

Proof.

$$
\begin{aligned}
&|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \\
&\overset{(i)}{\leq} |\hat{R}_{m+1}(\pi) - R(\pi)| + |\hat{R}_{\hat{f}_{m+1}}(\pi) - \hat{R}_{m+1}(\pi)| + |R_{\hat{f}_{m+1}}(\pi) - \hat{R}_{\hat{f}_{m+1}}(\pi)| \\
&\overset{(ii)}{\leq} 3.1\sqrt{\alpha_m \xi_{m+1}} + \frac{3}{2}\sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\text{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2 U_{\bar{m}}} + 2.2\sqrt{\xi_{m+1}}
\end{aligned} \tag{33}
$$

Where (i) follows from Lemma 9. And (ii) follows from Equation (2), Lemma 8, Lemma 10, and $\mathcal{W}_1$. □

B.5 Inductive Argument

This sub-section leverages the guarantee of Lemma 12 and applies it inductively to derive Lemma 14. This lemma bounds $\text{Reg}_{\Pi}(\pi)$ in terms of $\text{Reg}_{\hat{f}_{m+1}}(\pi)$ and vice-versa for any policy $\pi \in \Pi$. The proof of Lemma 14, relies on the following helpful lemma.

Lemma 13. *Consider any class of policies $\Pi' \supseteq \Pi$ and consider any fixed constants $l_1, l_2, l_3, C' > 0$. At any epoch m, suppose the policy evaluation guarantee of Equation (34) holds.*

$$
\forall \pi \in \Pi', \;\; |R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \leq l_1\sqrt{\xi_{m+1}} + l_2\sqrt{\alpha_m \xi_{m+1}} + \frac{l_3}{C'} \sum_{\bar{m} \in [m]} \frac{z_{\bar{m},m+1} \textit{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2} \tag{34}
$$

Now consider fixed constants $C_1, C_2 \geq 0$. As an inductive hypothesis, suppose Equation (35) holds.

$$
\forall \bar{m} \in [m], \forall \pi \in \Pi', \;\; \textit{Reg}_{\hat{f}_{\bar{m}}}(\pi) \leq \frac{4}{3}\textit{Reg}_{\Pi}(\pi) + C_1\sqrt{\xi_{\bar{m}}} + C_2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}. \tag{35}
$$

We then have that Equation (36) holds.

$$
\forall \pi \in \Pi', \textit{Reg}_{\Pi}(\pi) \leq \frac{6}{5}\textit{Reg}_{\hat{f}_{m+1}}(\pi) + \frac{12}{5}\left(l_1 + \frac{l_3 C_1}{C'}\right)\sqrt{\xi_{m+1}} + \frac{12}{5}\left(l_2 + \frac{l_3 C_2}{C'}\right)\sqrt{\alpha_m \xi_{m+1}}. \tag{36}
$$

Now consider $C_3 \geq 0$ and further suppose Equation (37) holds.

$$
\forall \bar{m} \in [m], \textit{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{m+1}}) \leq C_3\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}. \tag{37}
$$

We then also have that Equation (38) holds,

$$
\forall \pi \in \Pi', \textit{Reg}_{\hat{f}_{m+1}}(\pi) \leq \frac{7}{6}\textit{Reg}_{\Pi}(\pi) + \left(2l_1 + \frac{l_3 C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{l_3(C_2 + C_3)}{C'}\right)\sqrt{\alpha_m \xi_{m+1}}. \tag{38}
$$

Where $C' \geq 8l_3$, $z_{\bar{m},m+1} := \sqrt{\frac{\alpha_m \xi_{m+1}}{\alpha_{\bar{m}-1}\xi_{\bar{m}}}} \leq 1$, and $\alpha_m \leq \alpha_{\bar{m}-1}$ for all $\bar{m} \in [m]$.

Proof. Consider any policy $\pi \in \Pi'$. Suppose (34) and (35) hold. We first show (39).

$$
\begin{aligned}
&\mathrm{Reg}_{\Pi}(\pi) - \mathrm{Reg}_{\hat{f}_{m+1}}(\pi) \\
&= R(\pi^*) - R(\pi) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) + R_{\hat{f}_{m+1}}(\pi) \\
&\le R(\pi^*) - R_{\hat{f}_{m+1}}(\pi^*) + (R_{\hat{f}_{m+1}}(\pi) - R(\pi)) \\
&\overset{(i)}{\le} 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}(\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi) + \mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi^*)) \\
&\overset{(ii)}{\le} 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} \\
&\qquad + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}\left(\frac{4}{3}\mathrm{Reg}_{\Pi}(\pi) + 2C_1\sqrt{\xi_{\bar{m}}} + 2C_2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}\right) \\
&= 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} \\
&\qquad + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{1}{2\bar{m}^2}\left(\frac{4z_{\bar{m},m+1}}{3}\mathrm{Reg}_{\Pi}(\pi) + \frac{2C_1\sqrt{\xi_{m+1}}}{\sqrt{\alpha_{\bar{m}-1}/\alpha_m}} + 2C_2\sqrt{\alpha_m\xi_{m+1}}\right) \\
&\overset{(iii)}{\le} \left(2l_1 + \frac{2l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{2l_3C_2}{C'}\right)\sqrt{\alpha_m\xi_{m+1}} + \frac{4}{3}\frac{l_3}{C'}\mathrm{Reg}_{\Pi}(\pi),
\end{aligned}
\tag{39}
$$

Where (i) follows from (34), (ii) follows from (35) and from $\mathrm{Reg}_{\Pi}(\pi^*) = 0$, and finally (iii) follows from $z_{\bar{m},m+1} \le 1$, $\alpha_m \le \alpha_{\bar{m}-1}$, and $\sum_{\bar{m}\in[m]} 1/(2\bar{m}^2) \le 1$. Now (39) immediately implies (40).

$$
\begin{aligned}
&\left(1 - \frac{4l_3}{3C'}\right)\mathrm{Reg}_{\Pi}(\pi) \le \mathrm{Reg}_{\hat{f}_{m+1}}(\pi) + \left(2l_1 + \frac{2l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{2l_3C_2}{C'}\right)\sqrt{\alpha_m\xi_{m+1}} \\
&\overset{(i)}{\Longrightarrow} \mathrm{Reg}_{\Pi}(\pi) \le \frac{6}{5}\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) + \frac{12}{5}\left(l_1 + \frac{l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \frac{12}{5}\left(l_2 + \frac{l_3C_2}{C'}\right)\sqrt{\alpha_m\xi_{m+1}}
\end{aligned}
\tag{40}
$$

Where (i) follows from the fact that $C' \ge 8l_3$. Similar to (39), we will now show (41).

$$
\begin{aligned}
&\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) - \mathrm{Reg}_{\Pi}(\pi) \\
&= R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi) - (R(\pi^*) - R(\pi)) \\
&\le \left(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R(\pi_{\hat{f}_{m+1}})\right) + \left(R(\pi) - R_{\hat{f}_{m+1}}(\pi)\right) \\
&\overset{(i)}{\le} 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}(\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{m+1}}) + \mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)) \\
&\overset{(ii)}{\le} 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} \\
&\qquad + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}\left(\frac{4}{3}\mathrm{Reg}_{\Pi}(\pi) + C_1\sqrt{\xi_{\bar{m}}} + (C_2 + C_3)\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}\right) \\
&= 2l_1\sqrt{\xi_{m+1}} + 2l_2\sqrt{\alpha_m\xi_{m+1}} \\
&\qquad + \frac{l_3}{C'}\sum_{\bar{m}\in[m]}\frac{1}{2\bar{m}^2}\left(\frac{4z_{\bar{m},m+1}}{3}\mathrm{Reg}_{\Pi}(\pi) + \frac{C_1\sqrt{\xi_{m+1}}}{\sqrt{\alpha_{\bar{m}-1}/\alpha_m}} + (C_2 + C_3)\sqrt{\alpha_m\xi_{m+1}}\right) \\
&\overset{(iii)}{\le} \left(2l_1 + \frac{l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{l_3(C_2 + C_3)}{C'}\right)\sqrt{\alpha_m\xi_{m+1}} + \frac{4l_3}{3C'}\mathrm{Reg}_{\Pi}(\pi)
\end{aligned}
\tag{41}
$$

Where (i) follows from (34), (ii) follows from (35), (37), and (iii) follows from $z_{\bar{m},m+1} \le 1$, $\alpha_m \le \alpha_{\bar{m}-1}$, and $\sum_{\bar{m}\in[m]} 1/(2\bar{m}^2) \le 1$. Now (41) immediately implies (42).

$$
\begin{aligned}
&\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) \le \left(1 + \frac{4l_3}{3C'}\right)\mathrm{Reg}_{\Pi}(\pi) + \left(2l_1 + \frac{l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{l_3(C_2 + C_3)}{C'}\right)\sqrt{\alpha_m\xi_{m+1}} \\
&\overset{(i)}{\Longrightarrow} \mathrm{Reg}_{\hat{f}_{m+1}}(\pi) \le \frac{7}{6}\mathrm{Reg}_{\Pi}(\pi) + \left(2l_1 + \frac{l_3C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{l_3(C_2 + C_4)}{C'}\right)\sqrt{\alpha_m\xi_{m+1}}.
\end{aligned}
\tag{42}
$$

Where (i) follows from the fact that $C' \ge 8l_3$. □

Lemma 14. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Now for any epoch $m \in [\hat{m}-1]$, we then have that (43) holds.*

$$\begin{aligned}\forall \pi \in \Pi,\ &Reg_{\Pi}(\pi) \le \frac{4}{3} Reg_{\hat{f}_{m+1}}(\pi) + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m \xi_{m+1}},\\ &Reg_{\hat{f}_{m+1}}(\pi) \le \frac{4}{3} Reg_{\Pi}(\pi) + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m \xi_{m+1}}.\end{aligned} \tag{43}$$

Moreover when $m \in [m^*]$, we have (43) holds for all policies $\pi \in \mathcal{P}$.

Proof. Note that (43) trivially holds for $m = 0$. We will now use an inductive argument. Consider any epoch $m \in [\hat{m}]$. As an inductive hypothesis, let us assume (44) holds. (i.e. (43) holds for epoch $m-1$.)

$$\begin{aligned}&\forall \pi \in \Pi, \bar{m} \in [m],\\ &\mathrm{Reg}_{\Pi}(\pi) \le \frac{4}{3}\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}},\\ &\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi) \le \frac{4}{3}\mathrm{Reg}_{\Pi}(\pi) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}.\end{aligned} \tag{44}$$

Hence from (44), we have (35) holds with $C_1 = 6.5$ and $C_2 = 12$. Since $m \in [\hat{m}]$, from Lemma 12, we have (45) holds.

$$\begin{aligned}&\forall \pi \in \Pi \cup \{p_{m+1}\},\\ &|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \le \frac{22}{10}\sqrt{\xi_{m+1}} + \frac{31}{10}\sqrt{\alpha_m \xi_{m+1}} + \frac{3}{40}\sum_{\bar{m}\in[m]} \frac{z_{\bar{m},m+1}\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2}\end{aligned} \tag{45}$$

Hence from (45), we have (34) holds with $l_1 = 2.2$, $l_2 = 3.1$, $l_3 = 1.5$, $C' = 20$, and $\tilde{\Pi} = \Pi$. Hence from Lemma 13, we have (46) holds.

$$\begin{aligned}&\forall \pi \in \Pi,\\ &\mathrm{Reg}_{\Pi}(\pi) \le \frac{6}{5}\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) + \frac{12}{5}\left(\frac{22}{10} + \frac{1.5*6.5}{20}\right)\sqrt{\xi_{m+1}} + \frac{12}{5}\left(\frac{31}{10} + \frac{1.5*12}{20}\right)\sqrt{\alpha_m\xi_{m+1}}\\ &= \frac{6}{5}\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) + 6.45\sqrt{\xi_{m+1}} + 9.6\sqrt{\alpha_m\xi_{m+1}}\\ &\le \frac{4}{3}\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m\xi_{m+1}}\end{aligned} \tag{46}$$

Now from (44) and (46), we have (47) holds.

$$\begin{aligned}&\forall \bar{m} \in [m],\\ &\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{m+1}}) \le \frac{4}{3}\mathrm{Reg}_{\Pi}(\pi_{\hat{f}_{m+1}}) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}\\ &\le \frac{4}{3}\left(0 + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m\xi_{m+1}}\right) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}\\ &\le 43.2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}\end{aligned} \tag{47}$$

Hence from (47), we have (37) holds with $C_3 = 43.2$. Therefore from Lemma 13 we have (48).

$$\begin{aligned}&\forall \pi \in \Pi,\\ &\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) \le \frac{7}{6}\mathrm{Reg}_{\Pi}(\pi) + \left(2l_1 + \frac{l_3 C_1}{C'}\right)\sqrt{\xi_{m+1}} + \left(2l_2 + \frac{l_3(C_2+C_3)}{C'}\right)\sqrt{\alpha_m\xi_{m+1}}\\ &= \frac{7}{6}\mathrm{Reg}_{\Pi}(\pi) + \left(2*2.2 + \frac{1.5*6.5}{20}\right)\sqrt{\xi_{m+1}} + \left(2*3.1 + \frac{1.5(12+43.2)}{20}\right)\sqrt{\alpha_m\xi_{m+1}}\\ &= \frac{7}{6}\mathrm{Reg}_{\Pi}(\pi) + 4.8875\sqrt{\xi_{m+1}} + 10.34\sqrt{\alpha_m\xi_{m+1}}\\ &\le \frac{4}{3}\mathrm{Reg}_{\Pi}(\pi) + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m\xi_{m+1}}\end{aligned} \tag{48}$$

From (46) and (48), we have (43) holds for epoch m. This completes our inductive argument. □

An immediate implication of Lemma 14 is that we have $\text{Reg}_{\hat{f}_m}(\pi^*) \leq U_m$ for all $m \in [\hat{m}]$. Hence, from Lemma 4 and Lemma 5, we have (49) holds.

$$V(p_m, \pi^*) \leq \frac{\mathbb{E}[\mu(C_m(x, \beta_{\max}, \eta_m))]}{1 - \beta_{\max}} + \frac{K}{\eta_m} \leq \alpha_m, \ \forall m \in [\hat{m}]. \tag{49}$$

B.6 Bounding Exploration and Cumulative Regret

This sub-section leverages the structure of the kernel p_{m+1}, and the guarantees in Lemmas 12 and 14 to bound the expected regret during exploration (Lemma 17). Then, summing up these exploration regret bounds, we get our cumulative regret bound in Theorem 1. We start with Lemma 15 which leverages structure in p_m to bound $\text{Reg}_{\hat{f}_{\bar{m}}}(p_m)$ for any $\bar{m} \in [m]$.

Lemma 15. *For any pair of epochs $m \in [\hat{m}+1]$ and $\bar{m} \in [m]$, we have that (50) holds.*

$$\textit{Reg}_{\hat{f}_{\bar{m}}}(p_m) \leq 15.2\sqrt{\xi_{\bar{m}}} + 28\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}} + 2\bar{m}^2\eta_m U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_m U_{\bar{m}}}\right) \tag{50}$$

Proof. We first make the following observation.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim D_x}\mathop{\mathbb{E}}_{a\sim p_m(a|x)}[I(a \notin \pi_{\hat{f}_m}(x)) \cdot (\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a))] \\
&= \mathop{\mathbb{E}}_{x\sim D_x}\left[\int_{a\in\mathcal{A}\setminus\pi_{\hat{f}_m}(x)} (\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)) p_m(a|x) d\mu(a)\right] \\
&\overset{(i)}{=} \mathop{\mathbb{E}}_{x\sim D_x}\left[\int_{a\in\mathcal{A}\setminus\pi_{\hat{f}_m}(x)}\int_{\beta\in[0,1]} (\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)) \frac{I[a \in C_m(x, \min(\beta, \beta_{\max})/\eta_m)]}{\mu\big(C_m(x, \min(\beta, \beta_{\max})/\eta_m)\big)} \mathrm{d}\beta \mathrm{d}\mu(a)\right] \\
&\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{x\sim D_x}\left[\int_{\beta\in[0,1]}\int_{a\in\mathcal{A}\setminus\pi_{\hat{f}_m}(x)} \min\left(1, \frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\min(\beta, \beta_{\max})}\right) \frac{I[a \in C_m(x, \min(\beta, \beta_{\max})/\eta_m)]}{\mu\big(C_m(x, \min(\beta, \beta_{\max})/\eta_m)\big)} \mathrm{d}\mu(a) \mathrm{d}\beta\right] \\
&\leq \int_{\beta\in[0,1]} \min\left(1, \frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\min(\beta, \beta_{\max})}\right)\mathrm{d}\beta \leq (1-\beta_{\max})\frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\beta_{\max}} + \int_0^{\beta_{\max}} \min\left(1, \frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\beta}\right)\mathrm{d}\beta
\end{aligned} \tag{51}$$

where (i) follows from the definition of p_m given in (4). (ii) follows from the fact that for any $\zeta \in (0,1)$ and $a \in C_m(x,\zeta) \setminus \pi_{\hat{f}_m}(x)$ we have $\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a) \leq \min(1, 2\bar{m}^2/\zeta)$, since $C_m(x,\zeta) \setminus \pi_{\hat{f}_m}(x) \subseteq \bar{C}_m(x,\zeta) \subseteq \tilde{C}_{\bar{m}}(x, \zeta/(2\bar{m}^2))$ by Definition 2, . We now bound $\text{Reg}_{\hat{f}_{\bar{m}}}(p_m)$.

$$\begin{aligned}
&\text{Reg}_{\hat{f}_{\bar{m}}}(p_m) = \mathop{\mathbb{E}}_{x\sim D_x}\mathop{\mathbb{E}}_{a\sim p_m(a|x)}[\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a)] \\
&= \mathop{\mathbb{E}}_{x\sim D_x}\mathop{\mathbb{E}}_{a\sim p_m(a|x)}[(I(a \in \pi_{\hat{f}_m}(x)) + I(a \notin \pi_{\hat{f}_m}(x))) \cdot (\hat{f}_{\bar{m}}(x, \pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a))] \\
&\overset{(i)}{\leq} \text{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_m}) + \left((1-\beta_{\max})\frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\beta_{\max}} + \int_0^{\beta_{\max}} \min\left(1, \frac{2\bar{m}^2\eta_m U_{\bar{m}}}{\beta}\right)\mathrm{d}\beta\right) \\
&\leq \text{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_m}) + \left(\frac{2\bar{m}^2\eta_m U_{\bar{m}}(1-\beta_{\max})}{\beta_{\max}} + 2\bar{m}^2\eta_m U_{\bar{m}} + 2\bar{m}^2\eta_m U_{\bar{m}}\int_{2\bar{m}^2\eta_m U_{\bar{m}}}^{\beta_{\max}} \frac{1}{\beta}\mathrm{d}\beta\right) \\
&= \text{Reg}_{\hat{f}_{\bar{m}}}(\pi_{\hat{f}_m}) + 2\bar{m}^2\eta_m U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_m U_{\bar{m}}}\right) \\
&\overset{(ii)}{\leq} \frac{4}{3}\text{Reg}_{\Pi}(\pi_{\hat{f}_m}) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}} + 2\bar{m}^2\eta_m U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_m U_{\bar{m}}}\right) \\
&\overset{(iii)}{\leq} \frac{4}{3}\big(6.5\sqrt{\xi_m} + 12\sqrt{\alpha_{m-1}\xi_m}\big) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}} \\
&\qquad + 2\bar{m}^2\eta_m U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_m U_{\bar{m}}}\right) \\
&\overset{(iv)}{\leq} 15.2\sqrt{\xi_{\bar{m}}} + 28\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}} + 2\bar{m}^2\eta_m U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_m U_{\bar{m}}}\right)
\end{aligned} \tag{52}$$

where (i) follows from (51), (ii) follows from Lemma 14, (iii) follows from Lemma 14 and $\text{Reg}_{\hat{f}_m}(\pi_{\hat{f}_m}) = 0$, and (iv) follows from $\bar{m} \leq m$. □

Now from the guarantees in Lemmas 12, 14, and 15 we get the following bound on $\text{Reg}_\Pi(p_{m+1})$.

Lemma 16. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Now for any epoch $m \in [\hat{m}-1]$, we have (53) holds.*

$$\textit{Reg}_\Pi(p_{m+1}) \leq 100(m+1)^2\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right) \tag{53}$$

Proof. Since $m \in [\hat{m}]$, from Lemma 12, we have (54) holds.

$$\begin{aligned}
&\forall \pi \in \Pi \cup \{p_{m+1}\},\\
&|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \leq \frac{22}{10}\sqrt{\xi_{m+1}} + \frac{31}{10}\sqrt{\alpha_m\xi_{m+1}} + \frac{3}{40}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}\text{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{2\bar{m}^2}
\end{aligned} \tag{54}$$

We will now bound $\text{Reg}_\Pi(p_{m+1})$ in terms of $\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1})$ for $\bar{m} \in [m+1]$.

$$\begin{aligned}
&\text{Reg}_\Pi(p_{m+1}) - \text{Reg}_{\hat{f}_{m+1}}(p_{m+1})\\
&= R(\pi^*) - R(p_{m+1}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) + R_{\hat{f}_{m+1}}(p_{m+1})\\
&\leq R(\pi^*) - R_{\hat{f}_{m+1}}(\pi^*) + (R_{\hat{f}_{m+1}}(p_{m+1}) - R(p_{m+1}))\\
&\overset{(i)}{\leq} \frac{44}{10}\sqrt{\xi_{m+1}} + \frac{62}{10}\sqrt{\alpha_m\xi_{m+1}} + \frac{3}{40}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}(\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1}) + \text{Reg}_{\hat{f}_{\bar{m}}}(\pi^*))\\
&\overset{(ii)}{\leq} \frac{44}{10}\sqrt{\xi_{m+1}} + \frac{62}{10}\sqrt{\alpha_m\xi_{m+1}}\\
&\qquad + \frac{3}{40}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}(\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1}) + 6.5\sqrt{\xi_{\bar{m}}} + 12\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}})\\
&\overset{(iii)}{\leq} \frac{44}{10}\sqrt{\xi_{m+1}} + \frac{62}{10}\sqrt{\alpha_m\xi_{m+1}}\\
&\qquad + \frac{3}{40}\sum_{\bar{m}\in[m]}\frac{1}{2\bar{m}^2}(z_{\bar{m},m+1}\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1}) + 6.5\sqrt{\xi_{m+1}} + 12\sqrt{\alpha_m\xi_{m+1}})\\
&\overset{(iv)}{\leq} 4.9\sqrt{\xi_{m+1}} + 7.1\sqrt{\alpha_m\xi_{m+1}} + \frac{3}{40}\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1}).
\end{aligned} \tag{55}$$

Where (i) follows from (54), (ii) follows from Lemma 14 and from $\text{Reg}_\Pi(\pi^*) = 0$, (iii) follows from $z_{\bar{m},m+1} := \sqrt{\frac{\alpha_m\xi_{m+1}}{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}$ and $\alpha_m \leq \alpha_{\bar{m}-1}$, finally (iv) follows from $\sum_{\bar{m}\in[m]} 1/(2\bar{m}^2) \leq 1$. We now simplify the last term in the upper bound of (55).

$$\begin{aligned}
&\sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}\text{Reg}_{\hat{f}_{\bar{m}}}(p_{m+1})\\
&\overset{(i)}{\leq} \sum_{\bar{m}\in[m]}\frac{z_{\bar{m},m+1}}{2\bar{m}^2}\left(15.2\sqrt{\xi_{\bar{m}}} + 28\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}} + 2\bar{m}^2\eta_{m+1}U_{\bar{m}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{2\bar{m}^2\eta_{m+1}U_{\bar{m}}}\right)\right)\\
&\overset{(ii)}{\leq} \sum_{\bar{m}\in[m]}\frac{1}{2\bar{m}^2}\Bigg(15.2\sqrt{\xi_{m+1}} + 28\sqrt{\alpha_m\xi_{m+1}}\\
&\qquad + 40\bar{m}^2\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right)\Bigg)\\
&\overset{(iii)}{\leq} 15.2\sqrt{\xi_{m+1}} + 28\sqrt{\alpha_m\xi_{m+1}} + 20m\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right)
\end{aligned} \tag{56}$$

Where (i) follows from Lemma 15, (ii) follows from $z_{\bar{m},m+1} := \sqrt{\frac{\alpha_m \xi_{m+1}}{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}$, choice of U_m, and $\alpha_m \le \alpha_{\bar{m}-1}$, finally (iii) follows from $\sum_{\bar{m}\in[m]} 1/(2\bar{m}^2) \le 1$. By combining (55), (56), and Lemma 15, we get our final result.

$$\begin{aligned}
&\mathrm{Reg}_{\Pi}(p_{m+1}) \\
&\overset{(i)}{\le} \mathrm{Reg}_{\hat{f}_{m+1}}(p_{m+1}) + 6.04\sqrt{\xi_{m+1}} + 9.2\sqrt{\alpha_m \xi_{m+1}} \\
&\qquad + 1.5m\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right) \qquad (57)\\
&\overset{(ii)}{\le} 21.3\sqrt{\xi_{m+1}} + 37.2\sqrt{\alpha_m\xi_{m+1}} \\
&\qquad + 41.5(m+1)^2\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right)
\end{aligned}$$

Where (i) follows from (55) and (56), and (ii) follows from Lemma 15. □

The earlier bound on $\mathrm{Reg}_{\Pi}(p_{m+1})$ now immediately gives us the following bound on $\mathrm{Reg}_{f^*}(p_{m+1})$.

Lemma 17. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Now for any epoch $m \in [\hat{m}-1]$, we have that* (58)

$$\mathit{Reg}_{f^*}(p_{m+1}) \le 2\sqrt{KB} + 100(m+1)^2\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}\left(\frac{1}{\beta_{\max}} + \ln\frac{\beta_{\max}}{40\eta_{m+1}\sqrt{\alpha_m\xi_{m+1}}}\right) \qquad (58)$$

Proof. From Assumption 1 (properties of EstOracle), we know the bias of the model class $\mathcal{F}$ is bounded by B. In particular, we know there exists $g \in \mathcal{F}$ such that $\mathbb{E}_{x\sim D_{\mathcal{X}}, a\sim\mathrm{Unif}(\mathcal{A})}\left[(g(x,a) - f^*(x,a))^2\right] \le B$. Hence, we have, the following.

$$\begin{aligned}
&\mathrm{Reg}_{f^*}(\pi^*) \overset{(i)}{\le} \mathrm{Reg}_{f^*}(\pi_g) = R(\pi_{f^*}) - R(\pi_g) \\
&= (R(\pi_{f^*}) - R_g(\pi_{f^*})) - \mathrm{Reg}_g(\pi_{f^*}) + (R_g(\pi_g) - R(\pi_g)) \\
&\overset{(ii)}{\le} |R(\pi_{f^*}) - R_g(\pi_{f^*})| + |R_g(\pi_g) - R(\pi_g)| \\
&\overset{(iii)}{\le} \left(\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim\pi_{f^*}}\left[\frac{\pi_{f^*}(a|x)}{1/K}\right]} + \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim\pi_g}\left[\frac{\pi_g(a|x)}{1/K}\right]}\right) \qquad (59)\\
&\qquad \cdot \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim\mathrm{Unif}(\mathcal{A})}\left[(g(x,a) - f^*(x,a))^2\right]} \\
&\overset{(iv)}{\le} 2\sqrt{KB}.
\end{aligned}$$

Here (i) follows from the fact that $\pi_g \in \Pi$ since $g \in \mathcal{F}$. (ii) follows from triangle inequality and the fact that $\mathrm{Reg}_g(\pi_{f^*}) \ge 0$. (iii) follows from the proof of Lemma 7. And (iv) follows from $\mathbb{E}_{x\sim D_{\mathcal{X}}, a\sim\mathrm{Unif}(\mathcal{A})}\left[(g(x,a) - f^*(x,a))^2\right] \le B$ and $\pi_f(a|x) = I(a \in \pi_f(x))$.

Since $\mathrm{Reg}_{f^*}(p_{m+1}) = R(\pi_{f^*}) - R(\pi^*) + R(\pi^*) - R(p_{m+1}) = \mathrm{Reg}_{f^*}(\pi^*) + \mathrm{Reg}_{\Pi}(p_{m+1})$, the result follows from combining the above with Lemma 16. □

We now get our final cumulative regret bound by summing up the exploration regret bounds in Lemma 17.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Then with probability $1-\delta$, ω-RAPR attains the following cumulative regret guarantee. Here $\xi_{m+1} = \xi((\tau_m - \tau_{m-1})/3, \delta/(16m^3)$ for all m.*

$$\mathit{CReg}_T \le \tilde{\mathcal{O}}\left(\sum_{t=\tau_1+1}^{T}\sqrt{\frac{K}{\alpha_{m(t)}}\frac{\alpha_{m(t)-1}}{\alpha_{m(t)}}}\left(\sqrt{KB} + \sqrt{K\xi_{m(t)}}\right)\right) \qquad (6a)$$

$$\le \tilde{\mathcal{O}}\left(\sqrt{\omega KB}T + \sum_{t=\tau_1+1}^{T}\sqrt{\omega K\xi_{m(t)}}\right). \qquad (6b)$$

Where we use $\tilde{\mathcal{O}}$ to hide terms logarithmic in $T, K, \xi(T,\delta)$.

Proof. From Appendix B.1, both $\mathcal{W}_1$ and $\mathcal{W}_2$ hold with probability $1-\delta$. We prove our cumulative regret bounds under these events. Under $\mathcal{W}_1$, from Lemma 11, we have $\hat{m} \geq m^* + 1$. Further, from conditions in Section 1.2, we have ξ_m is non-increasing in m. Since $\xi(n,\delta')$ scales polynomially in $1/n$ and $\log(1/\delta')$, there exists a constant $Q_0 > 1$ such that the doubling epoch structure ensures $\xi_m \leq Q_0 \xi_{m+1}$ for all m. Hence $\xi_{\hat{m}} \leq Q_0 \xi_{\hat{m}+1} \leq Q_0 \xi_{m^*+2} \leq 2Q_0 B$. Let $m'(t) = \min(m(t), \hat{m})$. Hence, $\xi_{m'(t)} \leq \max(\xi_{m(t)}, \xi_{\hat{m}}) \leq 2Q_0 B + \xi_{m(t)}$. Therefore, by summing up the bounds in Lemma 17, we have the following cumulative regret bound.

$$
\begin{aligned}
\mathrm{CReg}_T &\leq \sum_{t=1}^{T} \mathrm{Reg}_{f^*}(p_{m'(t)}) \\
&\leq \tau_1 + \sum_{t=\tau_1+1}^{T} \Bigg(2\sqrt{KB} \\
&\qquad + 100(m'(t))^2 \eta_{m'(t)} \sqrt{\alpha_{m'(t)-1}\xi_{m'(t)}} \left(\frac{1}{\beta_{\max}} + \ln \frac{\beta_{\max}}{40\eta_{m'(t)}\sqrt{\alpha_{m'(t)-1}\xi_{m'(t)}}} \right) \Bigg) \qquad (60) \\
&\leq \tilde{\mathcal{O}}\left(\sum_{t=\tau_1+1}^{T} \left(\eta_{m'(t)} \sqrt{\alpha_{m'(t)-1}\xi_{m'(t)}} \right) \right) = \tilde{\mathcal{O}}\left(\sum_{t=\tau_1+1}^{T} \eta_{m(t)} \sqrt{\frac{\alpha_{m(t)-1}}{K}} \left(\sqrt{K\xi_{m'(t)}} \right) \right) \\
&\leq \tilde{\mathcal{O}}\left(\sum_{t=\tau_1+1}^{T} \eta_{m(t)} \sqrt{\frac{\alpha_{m(t)-1}}{K}} \left(\sqrt{KB} + \sqrt{K\xi_{m(t)}} \right) \right)
\end{aligned}
$$

Now the theorem follows from the fact that we have:

$$
\eta_m \sqrt{\frac{\alpha_{m-1}}{K}} \overset{\text{Lemma 5}}{\leq} 3\sqrt{\frac{K}{\alpha_m}\frac{\alpha_{m-1}}{\alpha_m}} \overset{\text{Lemma 5}}{\leq} 3\eta_m \sqrt{\frac{\alpha_{m-1}}{K}} \overset{(5)}{\leq} 3\sqrt{\omega}. \qquad (61)
$$

□

C Bounding Simple Regret

In this section, we prove our simple regret bound (Theorem 2). Our analysis starts with Lemma 18, which provides instance dependent bounds on $\mathbb{E}_{x\sim D_{\mathcal{X}}}\left[\mu\Big(C_m(x,\beta,\eta)\Big)\right]$. We will later use Lemma 18 to derive instance-dependent bounds on α_m. This bound then helps us derive instance-dependant bounds on simple regret.

Lemma 18. *For some environment parameters $\lambda \in (0,1)$, $\Delta > 0$, and $A \in [1,K]$, consider an instance where (7) holds.*

$$
\mathbb{P}_{x\sim D_{\mathcal{X}}}\Big(\mu\big(\{a \in \mathcal{A} : f^*(x,\pi_{f^*}(x)) - f^*(x,a) \leq \Delta\}\big) \leq A\Big) \geq 1-\lambda. \qquad (62)
$$

Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. For all epochs m, suppose the action selection kernel is given by eq. (4), and suppose (2) holds for all $\bar{m} \in [m]$. Then for any epoch $m \in [m^*]$, we have (63) holds.

$$
\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\mu\Big(C_m(x,\beta,\eta)\Big)\right] \leq \Big(1 + A + K\lambda\Big) + 25\frac{K}{\Delta}\frac{\eta}{\beta}\sqrt{\alpha_{m-1}\xi_m}. \qquad (63)
$$

For any $\beta \in (0, 1/2]$ and $\eta \in [1,K]$.

Proof. Consider any epoch $m \in [m^*]$. In this proof, for short-hand, let $C := \mathbb{E}[\mu(C_m(x,\beta,\eta))]$. We then have,

$$
\begin{aligned}
C &= \mathbb{E}[\mu(C_m(x,\beta,\eta))] \\
&\leq (A+1)P(\mu(C_m(x,\beta,\eta)) \leq A+1) + KP(\mu(C_m(x,\beta,\eta)) > A+1) \\
&\leq A + 1 + KP(\mu(C_m(x,\beta,\eta)) > A+1) \\
&\leq A + 1 + K - KP(\mu(C_m(x,\beta,\eta)) \leq A+1)
\end{aligned}
$$

The above immediately implies (64).

$$P\big(\mu(C_m(x,\beta,\eta)) \le A+1\big) \le \frac{A+1+K-C}{K}. \tag{64}$$

Let $\pi_0 \in \tilde{\Pi}$ be defined by (65).

$$\forall x \in \mathcal{X},\ \pi_0(x) \in \underset{S\in\Sigma_1 | S\subseteq C_m(x,\beta,\eta)}{\arg\min} f^*(x,S). \tag{65}$$

Since π_0 only selects arms in $C_m(x,\beta,\eta)$, from Definition 2, we have (66).

$$\mathrm{Reg}_{\hat{f}_m}(\pi_0) \le \frac{\eta}{\beta} U_m. \tag{66}$$

We can lower bound the regret of π_0 as follows,

$$\begin{aligned}
&\mathrm{Reg}_{f^*}(\pi_0)\\
\ge& P(f^*(x,\pi_{f^*}(x)) - f^*(x,\pi_0(x)) > \Delta)\cdot\Delta\\
\overset{(i)}{=}& P(\exists\ S\in\Sigma_1 |\ S\subseteq C_m(x,\beta,\eta),\ \ f^*(x,\pi_{f^*}(x)) - f^*(x,S) > \Delta)\cdot\Delta\\
\overset{(ii)}{\ge}& P\Big(\mu\big(C_m(x,\beta,\eta)\big)\ge A+1 \text{ and } \mu\big(\{a : (f^*(x,\pi_{f^*}(x)) - f^*(x,a) > \Delta\}\big) \ge K-A\Big)\cdot\Delta\\
=& P\Big(\mu\big(C_m(x,\beta,\eta)\big)\ge A+1 \text{ and } \mu\big(\{a : (f^*(x,\pi_{f^*}(x)) - f^*(x,a) \le \Delta\}\big) \le A\Big)\cdot\Delta\\
=& \Big(1 - P\Big(\mu\big(C_m(x,\beta,\eta)\big) < A+1 \text{ or } \mu\big(\{a : (f^*(x,\pi_{f^*}(x)) - f^*(x,a) \le \Delta\}\big) > A\Big)\Big)\cdot\Delta\\
\overset{(iii)}{\ge}& \Big(1 - P\Big(\mu\big(C_m(x,\beta,\eta)\big) < A+1\Big) - P\Big(\mu\big(\{a : (f^*(x,\pi_{f^*}(x)) - f^*(x,a) \le \Delta\}\big) > A\Big)\Big)\cdot\Delta\\
=& \Big(P\Big(\mu\big(\{a : (f^*(x,\pi_{f^*}(x)) - f^*(x,a) \le \Delta\}\big) \le A\Big) - P\Big(\mu\big(C_m(x,\beta,\eta)\big) < A+1\Big)\Big)\cdot\Delta\\
\overset{(iv)}{\ge}& \left(1-\lambda-\frac{A+1+K-C}{K}\right)\Delta = \left(\frac{C-A-1}{K}-\lambda\right)\Delta.
\end{aligned} \tag{67}$$

where (i) is because by construction $\pi_0(x) \subseteq C_m(x,\beta,\eta)$ for all x, (ii) is by the fact that μ is a finite measure with $\mu(\mathcal{A}) =: K$, (iii) follows from union bound, and (iv) follows from (64) and (7).

We will now work towards upper bounding $\mathrm{Reg}_{f^*}(\pi_0)$, and use this bound in conjunction with (67) to obtain our desired bound on C. To upper bound $\mathrm{Reg}_{f^*}(\pi_0)$ using Lemma 14, we will upper bound $\mathrm{Reg}_{\hat{f}_{m-1}}(\pi_0)$ and $\mathrm{Reg}_{\hat{f}_{m-1}}(\pi_{f^*})$.

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_{m-1}}(\pi_0) &\overset{(i)}{\le} \frac{4}{3}\mathrm{Reg}_{\Pi}(\pi_0) + 12\sqrt{\alpha_{m-2}\xi_{m-1}} + 6.5\sqrt{\xi_{m-1}}\\
&\overset{(ii)}{\le} \frac{4}{3}\left(\frac{4}{3}\mathrm{Reg}_{\hat{f}_m}(\pi_0) + 12\sqrt{\alpha_{m-1}\xi_m} + 6.5\sqrt{\xi_m}\right) + 12\sqrt{\alpha_{m-2}\xi_{m-1}} + 6.5\sqrt{\xi_{m-1}}\\
&\overset{(iii)}{\le} \frac{16}{9}\mathrm{Reg}_{\hat{f}_m}(\pi_0) + 28\sqrt{\alpha_{m-2}\xi_{m-1}} + \frac{91}{6}\sqrt{\xi_{m-1}}\\
&\overset{(iv)}{\le} \frac{16}{9}\mathrm{Reg}_{\hat{f}_m}(\pi_0) + \frac{259}{6}\sqrt{\alpha_{m-2}\xi_{m-1}}.
\end{aligned} \tag{68}$$

Where (i) and (ii) follow from Lemma 14, (iii) follows from $z_{m-1} = \sqrt{\frac{\alpha_{m-1}\xi_m}{\alpha_{m-2}\xi_{m-1}}} \le 1$, and (iv) follows from $\alpha_{m-2} \ge 1$.

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_{m-1}}(\pi_{f^*}) &\overset{(i)}{\le} 12\sqrt{\alpha_{m-2}\xi_{m-1}} + 6.5\sqrt{\xi_{m-1}}\\
&\overset{(ii)}{\le} \frac{37}{2}\sqrt{\alpha_{m-2}\xi_{m-1}}.
\end{aligned} \tag{69}$$

Where (i) follows from Lemma 14, (ii) follows from $\alpha_{m-2} \ge 1$.

$$
\begin{aligned}
&\text{Reg}_{f^*}(\pi_0) \\
&= R(\pi_{f^*}) - R(\pi_0) \\
&= \big(R(\pi_{f^*}) - R_{\hat{f}_m}(\pi_{f^*})\big) - \big(R(\pi_0) - R_{\hat{f}_m}(\pi_0)\big) + \big(R_{\hat{f}_m}(\pi_{f^*}) - R_{\hat{f}_m}(\pi_0)\big) \\
&\overset{(i)}{\leq} 2\sqrt{\alpha_{m-1}\xi_m} + \frac{1}{2}\sqrt{\alpha_{m-1}\xi_m} \sum_{\bar{m}\in[m]} \frac{1}{2\bar{m}^2 U_{\bar{m}}} \Big(\text{Reg}_{\hat{f}_{m-1}}(\pi_{f^*}) + \text{Reg}_{\hat{f}_{m-1}}(\pi_0)\Big) + \text{Reg}_{\hat{f}_m}(\pi_0) \\
&\overset{(ii)}{\leq} 2\sqrt{\alpha_{m-1}\xi_m} + \frac{1}{40} \sum_{\bar{m}\in[m]} \frac{z_{\bar{m},m-1}}{2\bar{m}^2} \Big(\frac{16}{9}\text{Reg}_{\hat{f}_m}(\pi_0) + \Big(37 + 18.5 * \frac{4}{3}\Big)\sqrt{\alpha_{m-2}\xi_{m-1}}\Big) + \text{Reg}_{\hat{f}_m}(\pi_0) \\
&\overset{(iii)}{\leq} 3.6\sqrt{\alpha_{m-1}\xi_m} + \frac{47}{45}\text{Reg}_{\hat{f}_m}(\pi_0) \overset{(iv)}{\leq} \sqrt{\alpha_{m-1}\xi_m}\Big(3.6 + \frac{47}{45} * 20\frac{\eta}{\beta}\Big) \leq 25\frac{\eta}{\beta}\sqrt{\alpha_{m-1}\xi_m}
\end{aligned} \tag{70}
$$

Where (i) follows from Lemma 7. (ii) follows from (68), (69), and $U_{m-1} = 20\sqrt{\alpha_{m-2}\xi_{m-1}}$, (iii) follows from $z_{m-1} \leq 1$, and (iv) follows from (66) and $U_m = 20\sqrt{\alpha_{m-1}\xi_m}$. Finally, combining (67) and (70), we have,

$$
\begin{aligned}
&\Big(\frac{C - A - 1}{K} - \lambda\Big)\Delta \leq \text{Reg}_{f^*}(\pi_0) \leq 25\frac{\eta}{\beta}\sqrt{\alpha_{m-1}\xi_m} \\
&\implies C \leq A + 1 + K\lambda + 25\frac{K}{\Delta}\frac{\eta}{\beta}\sqrt{\alpha_{m-1}\xi_m}.
\end{aligned} \tag{71}
$$

□

In Lemma 19, we use the bound from Lemma 18 to derive instance-dependent bounds on α_m. Corollary 2 is an immediate implication of Lemma 19, and provides a bound on α_m that doesn't depend on α_{m-1}. Finally, Corollary 2 is used to derive our instance-dependant bound on simple regret.

Lemma 19. *For some environment parameters $\lambda \in (0,1)$, $\Delta > 0$, and $A \in [1, K]$, consider an instance where (7) holds. Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold, and η_m is chosen using (5). For all epochs m, suppose the action selection kernel is given by eq. (4), suppose eq. (17) holds, and suppose (2) holds for all $\bar{m} \in [m]$. Then for any epoch $m \in [m^*]$, we have (72) holds.*

$$
\alpha_m \leq \mathcal{O}\Bigg(\max\Bigg(\sqrt{\frac{K\alpha_{m-1}}{\omega}}, A + K\lambda + \frac{\sqrt{K^3\omega\xi_m}}{\Delta}\Bigg)\Bigg) \tag{72}
$$

Proof. Suppose $\eta_m \leq \sqrt{\frac{K\omega}{\alpha_{m-1}}} - 1/|S_{m-1,2}|$, we then have,

$$
\begin{aligned}
&\frac{K}{\eta_m} \overset{(i)}{\leq} \frac{|S_{m-1,2}| + 1}{|S_{m-1,2}|}\frac{K}{\eta_m + \frac{1}{|S_{m-1,2}|}} \\
&\overset{(ii)}{\leq} \frac{|S_{m-1,2}| + 1}{|S_{m-1,2}|}\lambda_m\Big(\eta_m + \frac{1}{|S_{m-1,2}|}\Big) \\
&\overset{(iii)}{\leq} \frac{|S_{m-1,2}| + 1}{|S_{m-1,2}|}\Bigg(1 + \mathbb{E}\Big[\mu\Big(C_m\Big(x_t, \beta_{\max}, \eta_m + \frac{1}{|S_{m-1,2}|}\Big)\Big)\Big] + \sqrt{\frac{2K^2\ln(8|S_{m-1,2}|m^2/\delta)}{|S_{m-1,2}|}}\Bigg) \\
&\overset{(iv)}{\leq} \frac{|S_{m-1,2}| + 1}{|S_{m-1,2}|}\Bigg((2 + A + K\lambda) + 25\frac{K}{\Delta}\frac{\eta_m + \frac{1}{|S_{m-1,2}|}}{\beta_{\max}}\sqrt{\alpha_{m-1}\xi_m} + \sqrt{\frac{2K^2\ln(8|S_{m-1,2}|m^2/\delta)}{|S_{m-1,2}|}}\Bigg) \\
&\overset{(v)}{\leq} \frac{|S_{m-1,2}| + 1}{|S_{m-1,2}|}\Bigg((1 + A + K\lambda) + \frac{50}{\Delta}\sqrt{K^3\omega\xi_m} + \sqrt{\frac{2K^2\ln(8|S_{m-1,2}|m^2/\delta)}{|S_{m-1,2}|}}\Bigg)
\end{aligned} \tag{73}
$$

Where (i) follows from $\eta_m \geq 1$, (ii) follows from (5), (iii) follows from $\mathcal{W}_2$, (iv) follows from Lemma 18, and (v) follows from (5) and the fact that $\beta_{\max} = 0.5$. Finally, the result now follows from Lemma 5. □

Corollary 2. *For some environment parameters $\lambda \in (0,1)$, $\Delta > 0$, and $A \in [1, K]$, consider an instance where (7) holds. Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. For all epochs m, suppose the action selection kernel is given by eq. (4), suppose eq. (17) holds, and suppose suppose (2) holds for all $\bar{m} \in [m]$. Then for any epoch $m \in [m^*]$, we have (74) holds.*

$$\alpha_m \leq \mathcal{O}\left(\frac{K}{\omega} + A + K\lambda + \frac{\sqrt{K^3\omega\xi_{m-\lceil \log_2 \log_2(K)\rceil}}}{\Delta}\right) \tag{74}$$

Where for notational convenience, we let $\xi_i = 1$ for $i \leq 0$.

Proof. By repeatedly applying Lemma 19, we have:

$$\begin{aligned}
&\alpha_m \\
&\leq \mathcal{O}\left(\max\left(\left(\frac{K}{\omega}\right)^{\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{\lceil \log_2 \log_2(K)\rceil}}} K^{0.5^{\lceil \log_2 \log_2(K)\rceil}}, A + K\lambda + \frac{\sqrt{K^3\omega\xi_{m-\lceil \log_2 \log_2(K)\rceil}}}{\Delta} \right)\right) \\
&\overset{(i)}{\leq} \mathcal{O}\left(\max\left(\left(\frac{K}{\omega}\right) K^{0.5^{\lceil \log_2 \log_2(K)\rceil}}, A + K\lambda + \frac{\sqrt{K^3\omega\xi_{m-\lceil \log_2 \log_2(K)\rceil}}}{\Delta} \right)\right) \\
&\overset{(ii)}{\leq} \mathcal{O}\left(\max\left(\left(\frac{K}{\omega}\right), A + K\lambda + \frac{\sqrt{K^3\omega\xi_{m-\lceil \log_2 \log_2(K)\rceil}}}{\Delta} \right)\right) \\
&\leq \mathcal{O}\left(\frac{K}{\omega} + A + K\lambda + \frac{\sqrt{K^3\omega\xi_{m-\lceil \log_2 \log_2(K)\rceil}}}{\Delta}\right)
\end{aligned} \tag{75}$$

where (i) follows from $\sum_{i=1}^{\infty} 1/2^i = 1$, and (ii) follows from $K^{1/2^{\lceil \log_2 \log_2(K)\rceil}} \leq K^{1/2^{\log_2 \log_2(K)}} = K^{1/\log_2 K} = K^{\log_K 2} = 2$. □

We now re-state and prove Theorem 2. As discussed earlier, this result relies on the bound in Corollary 2.

Theorem 2. *Suppose Assumptions 1 and 2 hold. For some $(\lambda, \Delta, A) \in [0,1] \times (0,1] \times [1, K]$, consider instances where for $1-\lambda$ fraction of contexts at most A arms are Δ optimal (i.e. (7) holds).*

$$\mathbb{P}_{x \sim D_{\mathcal{X}}}\left(\mu\big(\{a \in \mathcal{A} : f^*(x, \pi_{f^*}(x)) - f^*(x,a) \leq \Delta\}\big) \leq A\right) \geq 1 - \lambda. \tag{7}$$

Let $m' = \min(\hat{m}, m(T)) - 1$. Let the learned policy $\hat{\pi}$ be given by (8) (equivalent to variance penalized policy optimization).

$$\hat{\pi} \in \arg\max_{\pi \in \Pi} \hat{R}_{m(T)}(\pi) - \frac{1}{2}\sqrt{\alpha_{m'}\xi_{m(T)}} \sum_{\bar{m} \in [m']} \frac{\hat{R}_{m(T),\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m(T),\hat{f}_{\bar{m}}}(\pi)}{40\bar{m}^2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}}. \tag{8}$$

Then with probability $1-\delta$, ω-RAPR has the following simple regret bound when T samples.

$$\begin{aligned}
&\mathit{Reg}_{\Pi}(\hat{\pi}) \leq \mathcal{O}\left(\sqrt{\alpha_{m'}\xi_{m(T)}}\right) \\
&\leq \mathcal{O}\left(\sqrt{\xi_{m(T)} \min\left(K, A + K\lambda + \frac{K}{\omega} + \frac{K^{3/2}\omega^{1/2}}{\Delta}\sqrt{\xi_{\min(m^*, m(T)-1)-\lceil \log_2 \log_2(K)\rceil}}\right)}\right).
\end{aligned}$$

Proof. From Appendix B.1, both $\mathcal{W}_1$ and $\mathcal{W}_2$ hold with probability $1-\delta$. We prove our simple regret bounds under these events. Let $m = m(T)$, we then have the following bound.

$$
\begin{aligned}
R(\hat{\pi}) &\overset{(i)}{\geq} \hat{R}_m(\hat{\pi}) - \sqrt{\alpha_{m'}\xi_m} - \frac{1}{2}\sqrt{\alpha_{m'}\xi_m}\sum_{\bar{m}\in[m']}\frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\hat{\pi})}{2\bar{m}^2 U_{\bar{m}}} - \frac{K\xi_m}{\eta_{m'}\min_{\bar{m}\in[m']}U_{\bar{m}}} \\
&\overset{(ii)}{\geq} \hat{R}_m(\hat{\pi}) - \sqrt{\alpha_{m'}\xi_m} - \frac{1}{2}\sqrt{\alpha_{m'}\xi_m}\sum_{\bar{m}\in[m']}\frac{\widehat{\mathrm{Reg}}_{m,\hat{f}_{\bar{m}}}(\hat{\pi})}{2\bar{m}^2 U_{\bar{m}}} - \frac{2\sqrt{\alpha_{m'}}\xi_m}{U_{m'}} - \frac{\alpha_{m'}\xi_m}{U_{m'}} \\
&\overset{(iii)}{\geq} \hat{R}_m(\pi^*) - \sqrt{\alpha_{m'}\xi_m} - \frac{1}{2}\sqrt{\alpha_{m'}\xi_m}\sum_{\bar{m}\in[m']}\frac{\widehat{\mathrm{Reg}}_{m,\hat{f}_{\bar{m}}}(\pi^*)}{2\bar{m}^2 U_{\bar{m}}} - \frac{2\sqrt{\alpha_{m'}}\xi_m}{U_{m'}} - \frac{\alpha_{m'}\xi_m}{U_{m'}} \\
&\overset{(iv)}{\geq} \hat{R}_m(\pi^*) - \sqrt{\alpha_{m'}\xi_m} - \frac{1}{2}\sqrt{\alpha_{m'}\xi_m}\sum_{\bar{m}\in[m']}\frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi^*)}{2\bar{m}^2 U_{\bar{m}}} - \frac{4\sqrt{\alpha_{m'}}\xi_m}{U_{m'}} - \frac{\alpha_{m'}\xi_m}{U_{m'}} \\
&\overset{(v)}{\geq} R(\pi^*) - 2\sqrt{\alpha_{m'}\xi_m} - \sqrt{\alpha_{m'}\xi_m}\sum_{\bar{m}\in[m']}\frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi^*)}{2\bar{m}^2 U_{\bar{m}}} - \frac{4\sqrt{\alpha_{m'}}\xi_m}{U_{m'}} - \frac{2\alpha_{m'}\xi_m}{U_{m'}} \\
&\overset{(vi)}{\geq} R(\pi^*) - 3.3\sqrt{\alpha_{m'}\xi_m}.
\end{aligned}
\tag{76}
$$

Here (i) follows from Lemma 8. (ii) follows from Lemma 10, Lemma 5, and the fact that $U_{m'} \leq U_{\bar{m}}$ for any $\bar{m} \in [m']$. (iii) follows from (8). (iv) follows from Lemma 10. (v) follows from Lemma 8, Lemma 5, and the fact that $U_{m'} \leq U_{\bar{m}}$ for any $\bar{m} \in [m']$. Finally, (vi) follows from Lemma 14 and $U_{m'} \leq 20\sqrt{\alpha_{m'}\xi_m}$. Hence $\mathrm{Reg}_{\Pi}(\hat{\pi}) \leq \mathcal{O}(\sqrt{\alpha_{m'}\xi_m})$. Now the final bound follows from the fact that $\alpha_{m'} \leq \alpha_1 = 3K$, $\alpha_{m'} \leq \alpha_{\min(m^*, m(T)-1)}$, and Corollary 2. □

D Lower bound

Theorem 3. *Given parameters $K, F, T \in \mathbb{N}$ and $\phi \in [1,\infty)$. There exists a context space $\mathcal{X}$ and a function class $\mathcal{F} \subseteq (\mathcal{X}\times\mathcal{A}\to[0,1])$ with K actions such that $|\mathcal{F}| \leq F$ and the following lower bound on cumulative regret holds:*

$$
\inf_{\mathbf{A}\in\Psi_\phi}\sup_{D\in\mathcal{D}}\mathop{\mathbb{E}}_{D}\left[\sum_{t=1}^{T}\big(r_t(\pi^*(x_t)) - r_t(a_t)\big)\right] \geq \tilde{\Omega}\left(\sqrt{\frac{K}{\phi}}\sqrt{KT\log F}\right)
$$

Here $(a_1,\dots a_T)$ denotes the actions selected by an algorithm A. $\mathcal{D}$ denotes the set of environments such that $f^* \in \mathcal{F}$ and (7) hold with $(A,\lambda,\Delta) = (1,0,0.24)$. Π denotes policies induced by $\mathcal{F}$. Ψ_ϕ denotes the set of CB algorithms that run for T rounds and output a learned policy with a simple regret guarantee of $\sqrt{\phi\log F/T}$ for any instance in $\mathcal{D}$ with confidence at least 0.95, i.e., $\Psi_\phi := \{A : \mathbb{P}(\mathit{Reg}(\hat{\pi}_{\mathbf{A}}) \leq \sqrt{\phi\log F/T}) \geq 0.95$ for any instance in $\mathcal{D}\}$. Finally, $\tilde{\Omega}(\cdot)$ hides factors logarithmic in K and T.

We prove theorem 3 in the following sub-sections.

D.1 Basic Technical Results

The following result is established in Raginsky and Rakhlin [2011], with this version taken from the proof of Lemma D.2 in Foster et al. [2020b].

Lemma 20 (Fano's inequality with reverse KL-divergence)**.** *Let*

$$
\mathcal{H} = (x_1, a_1, r_1(a_1)), \dots, (x_T, a_T, r_T(a_T)),
$$

and let $\{\mathbb{P}^{(i)}\}_{i\in[M]}$ be a collection of measures over $\mathcal{H}$, where $M \geq 2$. Let $\mathcal{Q}$ be any reference measure over $\mathcal{H}$, and let $\mathbb{P}$ be the law of $(m^*, \mathcal{H})$ under the following process:

- *Sample m^* uniformly from $[M]$.*
- *Sample $\mathcal{H} \sim \mathbb{P}^{(m^*)}$.*

Then for any function $\hat{m}(\mathcal{H})$, if $\mathbb{P}(\hat{m} = m^*) \geq 1 - \delta$, then

$$\left(1 - \frac{1}{M}\right) \log(1/\delta) - \log 2 \leq \frac{1}{M} \sum_{i=1}^{M} D_{KL}(Q||\mathbb{P}^{(i)}). \tag{77}$$

D.2 Construction

If $K \leq 10$ or $T \leq 152^2 K \log F$ or $\phi \geq K$, our lower bound directly follows from the cumulative regret lower bound in Foster et al. [2020b]. Hence, without loss of generality, we can assume $K \geq 10, T \geq 152^2 K \log F$, and $\phi \leq K$.

The following construction closely follows lower bound arguments in Foster et al. [2020b]. Let $\mathcal{A} = \{a^{(1)}, a^{(2)}, \ldots, a^{(K)}\}$ be an arbitrary set of discrete actions. Let $k = \lfloor 1/\epsilon \rfloor$, and d be parameters that will be fixed later. With $\epsilon \in (0, 1)$, note that $1/(2\epsilon) \leq k \leq 1/\epsilon$. We will now define the context set $\mathcal{X}$ as the union of d disjoint partitions $\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(d)}$, where $\mathcal{X}^{(i)} = \{x^{(i,0)}, x^{(i,1)}, \ldots, x^{(i,k)}\}$ for all $i \in [d]$. Hence, we have $\mathcal{X} = \cup \mathcal{X}^{(i)}$ and $|\mathcal{X}| = d(k + 1)$.

For each partition index $i \in [d]$, we construct a policy class $\Pi^{(i)} \subseteq (\mathcal{X}^{(i)} \to \mathcal{A})$ as follows. First we let $\pi^{(i,0)} : \mathcal{X}^{(i)} \to \mathcal{A}$ be the policy that always selects arm $a^{(1)}$, and let $\pi^{(i,l,b)} : \mathcal{X}^{(i)} \to \mathcal{A}$ be defined as follows for all $l \in [k]$ and $b \in \mathcal{A}_0 := \mathcal{A} \setminus \{a^{(1)}\}$,

$$\forall x^{(i,j)} \in \mathcal{X}^{(i)},\ \pi^{(i,l,b)}(x^{(i,j)}) = \begin{cases} a^{(1)}, & \text{if } j \neq l, \\ b, & \text{if } j = l. \end{cases} \tag{78}$$

Construct $\Pi^{(i)} := \{\pi^{(i,l,b)} | l \in [k] \text{ and } b \in \mathcal{A}_0\} \cup \{\pi^{(i,0)}\}$.[20] Finally, we let $\Pi := \Pi^{(1)} \times \Pi^{(2)} \times \cdots \times \Pi^{(d)}$. We will now construct a reward model class $\mathcal{F}$ that induces Π.

Let $\Delta := 1/4$. For each partition index $i \in [d]$, we construct a reward model class $\mathcal{F}^{(i)} \subseteq (\mathcal{X}^{(i)} \times \mathcal{A} \to [0, 1])$ as follows. First we let $f^{(i,0)} : \mathcal{X}^{(i)} \times \mathcal{A} \to [0, 1]$ be defined as follows,

$$\forall (x^{(i,j)}, a) \in \mathcal{X}^{(i)} \times \mathcal{A},\ f^{(i,0)}(x^{(i,j)}, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a^{(0)}, \\ \frac{1}{2}, & \text{if } a \in \mathcal{A}_0. \end{cases} \tag{79}$$

For all $l \in [k]$ and $b \in \mathcal{A}_0$, we define $f^{(i,l,b)} : \mathcal{X}^{(i)} \times \mathcal{A} \to [0, 1]$ as follows,

$$\forall (x^{(i,j)}, a) \in \mathcal{X}^{(i)} \times \mathcal{A},\ f^{(i,l,b)}(x^{(i,j)}, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a^{(0)} \\ \frac{1}{2} + 2\Delta, & \text{if } j = l \text{ and } a = b, \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{80}$$

Note that $f^{(i,l,b)}$ differs from $f^{(i,0)}$ only at context $(x^{(i,l)}, b)$. Construct $\mathcal{F}^{(i)} := \{f^{(i,l,b)} | l \in [k] \text{ and } b \in \mathcal{A}_0\} \cup \{f^{(i,0)}\}$. Finally, we let $\mathcal{F} := \mathcal{F}^{(1)} \times \mathcal{F}^{(2)} \times \cdots \times \mathcal{F}^{(d)}$. Hence, we have,

$$|\mathcal{F}| = |\mathcal{F}^{(i)}|^d \leq (k \cdot K)^d \implies d \geq \frac{\log |\mathcal{F}|}{\log(K \cdot k)} \geq \frac{\log |\mathcal{F}|}{\log(K/\epsilon)}. \tag{81}$$

We choose d to be the largest value such that $F \geq (k \cdot K)^d$. Hence we choose $d = \lfloor \log F / \log(K \cdot k) \rfloor \geq \log F / (2 \log(K \cdot k))$.

To use lemma 20, we will describe a collection of environments that share a common distribution over contexts and only differ in the reward distribution. The context distribution $D_{\mathcal{X}}$ is given by $D_{\mathcal{X}} := \frac{1}{d} \sum_i D_{\mathcal{X}^{(i)}}$, where $D_{\mathcal{X}^{(i)}}$ is a distribution over $\mathcal{X}^{(i)}$, with ϵ probability of sampling each context in $\mathcal{X}^{(i)} \setminus \{x^{(i,0)}\}$, and $1 - k\epsilon$ probability of sampling the context $x^{(i,0)}$.

For each block $\mathcal{X}^{(i)}$, we let $\mathbb{P}^{(i,0)}$ denote the law given by the reward distribution $r(a) \sim \text{Ber}(f^{(i,0)}(x, a))$ for all $x \in \mathcal{X}^{(i)}$. Further, for any $l \in [k]$ and $b \in \mathcal{A}_0$, we let $\mathbb{P}^{(i,l,b)}$ denote the law given by the reward distribution $r(a) \sim \text{Ber}(f^{(i,l,b)}(x, a))$ for all $x \in \mathcal{X}^{(i)}$. For any policy $\pi \in \Pi^{(i)}$, we let $R^{(i,l,b)}(\pi) = \mathbb{E}_{\mathbb{P}^{(i,l,b)}}[r(\pi(x))]$ denote expected reward under $\mathbb{P}^{(i,l,b)}$, and let $\text{Reg}^{(i,l,b)}(\pi) = R^{(i,l,b)}(\pi^{(i,l,b)}) - R^{(i,l,b)}(\pi)$ denote expected simple regret under $\mathbb{P}^{(i,l,b)}$.

[20] Here $[k] = \{1, 2, \ldots, k\}$

We use ρ to index environments. Here $\rho = (\rho_1, \ldots, \rho_d)$, where $\rho_i = (l_i, b_i)$ for $l_i \in \{0, 1, \ldots, k\}$ and $b_i \in \mathcal{A}_0$. We let $\mathbb{P}_\rho$ denote an environment with the law $\mathbb{P}^{(i,l_i,b_i)}$ for contexts in $\mathcal{X}^{(i)}$.[21] Finally let π_ρ denote the optimal policy under $\mathbb{P}_\rho$, and let ${\pi_\rho}^{(i)}$ denote its restriction to $\mathcal{X}^{(i)}$. Let $\mathbb{E}_\rho[\cdot]$ denote the expectation under $\mathbb{P}_\rho$. Let $R_\rho(\pi) = \mathbb{E}_\rho[r(\pi(x))]$ denote the expected reward of π under $\mathbb{P}_\rho$, and let $\text{Reg}_\rho(\pi) = R_\rho(\pi_\rho) - R_\rho(\pi)$ denote the simple regret of π under $\mathbb{P}_\rho$.

D.3 Lower bound argument

We sample ρ from a distribution ν defined as follows. For each $i \in [d]$, set $l_i = 0$ with probability 0.5, otherwise l_i is selected uniformly from $[k]$. Select b_i uniformly from $\mathcal{A}_0$. Note that when $l_i = 0$, we disregard the value of b_i.

We let $\hat{\pi}_{\mathbf{A}} \in \Pi$ denotes the policy recommended by the contextual bandit algorithm $\mathbf{A}$ at the end of T rounds, and let $\hat{\pi}_{\mathbf{A}}^{(i)} \in \Pi^{(i)}$ be the restriction of $\hat{\pi}_{\mathbf{A}}$ to block $\mathcal{X}^{(i)}$. Note that the policy recommended by $\mathbf{A}$ will depend on the environment ρ.

Let $\mathcal{I} := \{i \in [d] | \text{Reg}^{(i,l_i,b_i)}({\pi_{\mathbf{A}}}^{(i)}) \leq 19\sqrt{\phi \log F/T}\}$. Since we only consider algorithms that guarantee the following with probability at least $19/20$,

$$\frac{1}{d}\sum_{i=1}^{d} \text{Reg}^{(i,l_i,b_i)}({\pi_{\mathbf{A}}}^{(i)}) = \text{Reg}_\rho(\hat{\pi}_{\mathbf{A}}) \leq \sqrt{\phi \log F/T}. \tag{82}$$

Under this event, we have that at most $d/19$ block indices satisfy $\text{Reg}^{(i,l_i,b_i)}({\pi_{\mathbf{A}}}^{(i)}) > 19\sqrt{\phi \log F/T}$. Therefore, we have $|\mathcal{I}| \geq 18d/19$.
Define event $M_i = \{i \in \mathcal{I}\}$. We have

$$\sum_{i=1}^{d} P(M_i) = \sum_{i=1}^{d} \mathbb{E}[1\{i \in \mathcal{I}\}] \geq \frac{19}{20}\,\mathbb{E}[|\mathcal{I}| | \text{Reg}_\rho(\hat{\pi}_{\mathbf{A}}) \leq \sqrt{\phi \log F/T}] \geq \frac{9d}{10}. \tag{83}$$

Consider any fixed index i, under the event M_i, we have the following. First observe for any $(l, b) \neq (l', b')$, we have $\text{Reg}^{(i,l,b)}(\pi^{(i,l',b')}) \geq \epsilon\Delta$. Let $(\hat{l}_i, \hat{b}_i)$ be indices such that ${\pi_{\mathbf{A}}}^{(i)} = \pi^{(i,\hat{l}_i,\hat{b}_i)}$. We now choose ϵ such that,

$$\epsilon = \frac{38}{\Delta}\sqrt{\frac{\phi \log F}{T}} \iff \frac{\epsilon\Delta}{2} = 19\sqrt{\frac{\phi \log F}{T}}. \tag{84}$$

Hence from definition of $\mathcal{I}$, we have $\text{Reg}^{(i,l_i,b_i)}({\pi_{\mathbf{A}}}^{(i)}) \leq \epsilon\Delta/2$. Further since $\text{Reg}^{(i,l_i,b_i)}(\pi) \geq \epsilon\Delta$ for all $\pi \in \Pi^{(i)} \setminus \{\pi^{(i,l_i,b_i)}\}$, we have $(\hat{l}_i, \hat{b}_i) = (l_i^*, b_i^*)$.

Restating the above result, we have the following. For any $i \in \mathcal{I}$, with probability $1 - 1/16$, we have $(\hat{l}_i, \hat{b}_i) = (l_i^*, b_i^*)$. Hence from lemma 20, we have,

$$\begin{aligned}
&\left(1 - \frac{1}{(K-1)k}\right)\log\left(1/P(\overline{M_1})\right) - \log 2 \\
&\leq \frac{1}{(K-1)k}\sum_{l=1}^{k}\sum_{b\in\mathcal{A}_0} D_{KL}(\mathbb{P}^{(i,0)} || \mathbb{P}^{(i,l,b)}) \\
&\overset{(i)}{=} \frac{1}{(K-1)k}\sum_{l=1}^{k}\sum_{b\in\mathcal{A}_0} D_{KL}(\text{Ber}(1/2) || \text{Ber}(1/2+2\Delta)) \mathop{\mathbb{E}}_{\mathbb{P}^{(i,0)}}[|\{t | x_t = x^{(i,l)}, a_t = b\}|] \\
&\overset{(ii)}{\leq} \frac{1}{(K-1)k}\sum_{l=1}^{k}\sum_{b\in\mathcal{A}_0} 4\Delta^2 \mathop{\mathbb{E}}_{\mathbb{P}^{(i,0)}}[|\{t | x_t = x^{(i,l)}, a_t = b\}|] \\
&= \frac{4\Delta^2}{(K-1)k} \mathop{\mathbb{E}}_{\mathbb{P}^{(i,0)}}[|\{t | x_t \in \mathcal{X}^{(i)} \setminus \{x^{(i,0)}\}, a_t \in \mathcal{A}_0\}|].
\end{aligned} \tag{85}$$

[21] Here $\mathbb{P}^{(i,0)} \equiv \mathbb{P}^{(i,0,b)}$ for all $b \in \mathcal{A}_0$.

Where (i) follows from the fact that $\mathbb{P}^{(i,0)}$ and $\mathbb{P}^{(i,l,b)}$ are identical unless $x_t = x^{(i,l)}$ and $a_t = b$, and (ii) follows from $\Delta \leq 1/4$. Clearly we have:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\rho\sim\nu}\mathop{\mathbb{E}}_{\rho}\left[\sum_{t=1}^{T}\big(r_t(\pi^*(x_t)) - r_t(a_t)\big)\right] \\
&\geq \Delta \mathop{\mathbb{E}}_{\rho\sim\nu}\mathop{\mathbb{E}}_{\rho}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\mathbb{I}\Big(\Big\{x_t \in \mathcal{X}^{(i)} \setminus \{x^{(i,0)}\}, a_t \in \mathcal{A}_0, l_i = 0\Big\}\Big)\right] \\
&\overset{(i)}{\geq} \frac{\Delta}{2}\sum_{i=1}^{d}\mathop{\mathbb{E}}_{\mathbb{P}^{(i,0)}}\left[|\Big\{t|\ x_t \in \mathcal{X}^{(i)} \setminus \{x^{(i,0)}\}, a_t \in \mathcal{A}_0\Big\}|\right] \\
&\overset{(ii)}{\geq} \frac{\Delta}{2}\sum_{i=1}^{d}\left(-\Big(1 - \frac{1}{k(K-1)}\Big)\log\big(P(\overline{M_i})\big) - \log 2\right)\cdot\frac{(K-1)k}{4\Delta^2} \\
&= \sum_{i=1}^{d}\left(-\Big(1 - \frac{1}{k(K-1)}\Big)\log\big(1 - P(M_i)\big) - \log 2\right)\cdot\frac{(K-1)k}{8\Delta} \qquad (86) \\
&\overset{(iii)}{\geq} \sum_{i=1}^{d}\left(\Big(1 - \frac{1}{k(K-1)}\Big)P(M_i) - \log 2\right)\cdot\frac{(K-1)k}{8\Delta} \\
&\overset{(iv)}{\geq} \Big\{\Big(1 - \frac{1}{k(K-1)}\Big)\frac{9}{10} - \log 2\Big\}\cdot\frac{(K-1)kd}{8\Delta} \\
&\overset{(v)}{\geq} \frac{Kkd}{100\Delta} \overset{(vi)}{\geq} \frac{dK}{200\Delta\epsilon} \overset{(vii)}{=} \frac{1}{7600}\sqrt{\frac{Td^2K^2}{\phi\log F}} \overset{(viii)}{\geq} \frac{1}{15200}\sqrt{\frac{K^2T\log F}{\phi\log^2(K\cdot k)}} \\
&\overset{(ix)}{\geq} \frac{1}{15200}\sqrt{\frac{K^2T\log F}{\phi\log^2(K\cdot T)}}.
\end{aligned}
$$

Where (i) follows from the fact that $\nu(l_i = 0) = 1/2$, (ii) follows from (85) and that $|\mathcal{I}| \geq d/2$, (iii) uses $\log(1+x) \leq x$, for $x > -1$, , (iv) uses (83), (v) follows from $k \geq 1$ and $K \geq 10$, (vi) follows from $k \geq 1/(2\epsilon)$, (vii) follows from choice of ϵ, (viii) follows from (81), and (ix) since $k \leq 1/\epsilon \overset{84}{=} \frac{1}{152}\sqrt{\frac{T}{\phi\log F}} \leq T$. This completes the proof of theorem 3.

E Additional Details

E.1 Conformal Arm Sets

The below lemma shows that, for any given policy π, the conformal arm sets given in definition 2 can be probabilistically relied on (over the distribution of contexts) to contain arms recommended by π, with low regret under the models estimated up to epoch m. Recall we earlier define $U_m = 20\sqrt{\alpha_{m-1}\xi_m}$.

Lemma 21 (Conformal Uncertainty)**.** *For any policy π and epoch m, we have:*

$$
\Pr_{x\sim D_{\mathcal{X}}, a\sim\pi(\cdot|x)}(a \in C_m(x,\zeta)) \geq 1 - \zeta\sum_{\bar{m}\in[m]}\frac{Reg_{\hat{f}_{\bar{m}}}(\pi)}{(2\bar{m}^2)U_{\bar{m}}} \qquad (87)
$$

Proof. For any policy π, we have (88) holds.

$$\begin{aligned}
&\Pr_{x\sim D_{\mathcal{X}},a\sim\pi(\cdot|x)}(a \notin C_m(x,\zeta)) \\
&\leq \Pr_{x\sim D_{\mathcal{X}},a\sim\pi(\cdot|x)}\Big(\bigcup_{\bar{m}\in[m]}\{a \notin \tilde{C}_{\bar{m}}(x,\zeta/(2\bar{m}^2))\}\Big) \\
&\overset{(i)}{\leq} \sum_{\bar{m}\in[m]} \Pr_{x\sim D_{\mathcal{X}},a\sim\pi(\cdot|x)}\Big(\hat{f}_{\bar{m}}(x,\pi_{\hat{f}_{\bar{m}}}(x)) - \hat{f}_{\bar{m}}(x,a) > \frac{(2\bar{m}^2)U_{\bar{m}}}{\zeta}\Big) \\
&\overset{(ii)}{\leq} \sum_{\bar{m}\in[m]} \frac{\mathrm{Reg}_{\hat{f}_{\bar{m}}}(\pi)}{(2\bar{m}^2)U_{\bar{m}}/\zeta}.
\end{aligned} \tag{88}$$

Where (i) follows from union bound and (ii) follows from Markov's inequality. □

Recall that right after Lemma 14, we show that $\mathrm{Reg}_{\hat{f}_{\bar{m}}} \leq U_{\bar{m}}$ with high-probability for any $\bar{m} \in [\hat{m}]$. Hence, Lemma 21 gives us that with high-probability we have $\Pr_{x\sim D_{\mathcal{X}},a\sim\pi^*(\cdot|x)}(a \in C_m(x,\zeta)) \geq 1-\zeta$. While we don't directly use Lemma 21, this lemma helps demonstrate the utility of CASs.

E.2 Argument for Surrogate Objective

Lemma 22 is a self-contained result proving that guarantying tighter bounds on the optimal cover leads to tighter simple regret bounds for any contextual bandit algorithm. Hence the optimal cover is a valid surrogate objective for simple regret. This lemma is not directly used in the analysis of ω-RAPR, however similar results (see Theorem 2) were proved and used. Note that the parameters below (including α) are not directly related to parameters maintained by ω-RAPR.

Lemma 22 (Valid Surrogate Objective)**.** *Suppose Π is a finite class and suppose a contextual bandit algorithm collects T samples using kernels $(p_t)_{t\in[T]}$ such that $p_t(\cdot|\cdot) \geq \sqrt{\frac{\ln(4|\Pi|/\delta)}{\alpha T}}$. Further suppose that the following condition holds with some $\alpha \in [1,\infty)$:*

$$\frac{1}{T}\sum_{t=1}^{T} V(p_t,\pi^*) \leq \alpha \tag{89}$$

Then we can estimate a policy $\hat{\pi} \in \Pi$ such that with probability at least $1-\delta$, we have:

$$|R(\pi^*) - R(\hat{\pi})| \leq \mathcal{O}\left(\sqrt{\frac{\alpha \ln(4|\Pi|/\delta)}{T}}\right). \tag{90}$$

Proof. WOLG we assume $T \geq \ln(4|\Pi|/\delta)$, since otherwise the result trivially holds. Now consider any policy π. Let $y_t := \frac{r_t\mathbb{I}(\pi(x_t)=a_t)}{p_t(\pi(x_t)|x_t)}$. Now note that:

$$\mathrm{Var}_t[y_t] \leq \mathop{\mathbb{E}}_{D(p_t)}[y_t^2] = \mathop{\mathbb{E}}_{(x_t,a_t,r_t)\sim D(p_t)}\left[\frac{r_t^2\mathbb{I}(\pi(x_t)=a_t)}{p_t^2(\pi(x_t)|x_t)}\right] \leq \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\frac{1}{p_t(\pi(x_t)|x_t)}\right] = V(p_t,\pi). \tag{91}$$

Then from from a Freedman-style inequality [See theorem 13 in Dudik et al., 2011], we have with probability at least $1-\delta/(2|\Pi|)$ that the following holds:

$$\begin{aligned}
&\left|\sum_{t=1}^{T}(y_t - R(\pi))\right| \leq 2\max\left\{\sqrt{\sum_{t=1}^{T}\mathrm{Var}(y_t)\ln(4|\Pi|/\delta)}, \frac{\ln(4|\Pi|/\delta)}{\sqrt{\frac{\ln(4|\Pi|/\delta)}{\alpha T}}}\right\} \\
&\overset{(i)}{\Longrightarrow} \left|\frac{1}{T}\sum_{t=1}^{T}\frac{r_t\mathbb{I}(\pi(x_t)=a_t)}{p_t(\pi(x_t)|x_t)} - R(\pi)\right| \leq 2\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\max\left\{\frac{1}{T}\sum_{t=1}^{T}V(p_t,\pi), \alpha\right\}}
\end{aligned} \tag{92}$$

Here (i) follows from (91). Similarly with probability at least $1-\delta/(2|\Pi|)$ the following holds:

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p_t(\pi(x_t)|x_t)}-\frac{1}{T}\sum_{t=1}^{T}V(p_t,\pi)\right| \\
&\overset{(i)}{\leq}\frac{2}{T}\max\left\{\sqrt{\sum_{t=1}^{T}\mathrm{Var}\Big(\frac{1}{p_t(\pi(x_t)|x_t)}\Big)\ln(4|\Pi|/\delta)},\frac{\ln(4|\Pi|/\delta)}{\sqrt{\frac{\ln(4|\Pi|/\delta)}{\alpha T}}}\right\} \\
&\overset{(ii)}{\leq}\frac{2}{T}\max\left\{\sqrt{T\frac{\alpha T}{\ln(4|\Pi|/\delta)}\ln(4|\Pi|/\delta)},\sqrt{\alpha T\ln(4|\Pi|/\delta)}\right\}\overset{(iii)}{\leq}2\sqrt{\alpha}\overset{(iv)}{\leq}2\alpha.
\end{aligned} \tag{93}$$

Here (i) follows from Freedman's inequality, (ii) follows from the lower bound on p_t, (iii) follows from $T\geq\ln(4|\Pi|/\delta)$, and (iv) follows from $\alpha\geq 1$. Hence the above events hold with probability at least $1-\delta$ for all policies $\pi\in\Pi$. Now let $\hat{\pi}$ be given as follows.

$$\hat{\pi}\in\arg\max_{\pi\in\Pi}\frac{1}{T}\sum_{t=1}^{T}\frac{r_t\mathbb{I}(\pi(x_t)=a_t)}{p_t(\pi(x_t)|x_t)}-2\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\left(2\alpha+\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p_t(\pi(x_t)|x_t)}\right)} \tag{94}$$

We then have the following lower bound on $R(\hat{\pi})$ using the definition of $\hat{\pi}$ and the above to high-probability events.

$$\begin{aligned}
R(\hat{\pi})&\overset{(i)}{\geq}\frac{1}{T}\sum_{t=1}^{T}\frac{r_t\mathbb{I}(\hat{\pi}(x_t)=a_t)}{p_t(\hat{\pi}(x_t)|x_t)}-2\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\max\left\{\frac{1}{T}\sum_{t=1}^{T}V(p_t,\hat{\pi}),\alpha\right\}} \\
&\overset{(ii)}{\geq}\frac{1}{T}\sum_{t=1}^{T}\frac{r_t\mathbb{I}(\hat{\pi}(x_t)=a_t)}{p_t(\hat{\pi}(x_t)|x_t)}-2\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\left(2\alpha+\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p_t(\hat{\pi}(x_t)|x_t)}\right)} \\
&\overset{(iii)}{\geq}\frac{1}{T}\sum_{t=1}^{T}\frac{r_t\mathbb{I}(\pi^*(x_t)=a_t)}{p_t(\pi^*(x_t)|x_t)}-2\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\left(2\alpha+\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p_t(\pi^*(x_t)|x_t)}\right)} \\
&\overset{(iv)}{\geq}R(\pi^*)-4\sqrt{\frac{\ln(4|\Pi|/\delta)}{T}\left(4\alpha+\frac{1}{T}\sum_{t=1}^{T}V(p_t,\pi^*)\right)}\geq R(\pi^*)-4\sqrt{\frac{5\alpha\ln(4|\Pi|/\delta)}{T}}
\end{aligned} \tag{95}$$

Here (i) follows from (92), (ii) follows from (93), (iii) follows from (94), and (iv) follows from (92) and (93). This completes the proof. □

E.3 Testing Misspecification via CSC

We restate the misspecification test that is used at the end of epoch m and argue how this test can be solved via two calls to a cost sensitive classification solver. First, let us restate the test in (96).

$$\begin{aligned}
&\max_{\pi\in\Pi\cup\{p_{m+1}\}}|\hat{R}_{m+1,\hat{f}_{m+1}}(\pi)-\hat{R}_{m+1}(\pi)|-\sqrt{\alpha_m\xi_{m+1}}\sum_{\bar{m}\in[m]}\frac{\hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}})-\hat{R}_{m+1,\hat{f}_{\bar{m}}}(\pi)}{40\bar{m}^2\sqrt{\alpha_{\bar{m}-1}\xi_{\bar{m}}}} \\
&\leq 2.05\sqrt{\alpha_m\xi_{m+1}}+1.1\sqrt{\xi_{m+1}},
\end{aligned} \tag{96}$$

We are interested in calculating the value of the maximization problem in (96). To calculate this maximum, we need to fix our estimators. Let $\hat{R}_{m+1,f}(\pi):=\frac{1}{|S_{m,3}|}\sum_{t\in S_{m,3}}f(x_t,\pi(x_t))=\frac{1}{|S_{m,3}|}\sum_{t\in S_{m,3}}\mathbb{E}_{a\sim\pi(\cdot|x_t)}f(x_t,a)$ for any policy π and reward model f, which is the only obvious estimator we could think off for $R_f(\pi)$. Also let us use IPS estimaton for policy evaluation (the same argument works for DR), $\hat{R}_{m+1}(\pi):=\frac{1}{|S_{m,3}|}\sum_{t\in S_{m,3}}\frac{\pi(a_t|x_t)r_t(a_t)}{p_m(a_t|x_t)}$. [22] [23] Note that the value of

[22] When evaluating a general kernel q, we use the natural extension of these estimators of policy value. In particular, simply replace $\pi(\cdot|x)$ with $q(\cdot|x)$ in their formulas.

[23] Up to constant factors, these estimators give us the best rates in Assumption 2 with finite classes. These estimators are also used in several contextual bandit papers [e.g., Agarwal et al., 2014, Li et al., 2022].

the maximization problem in (96) is equal to $\max(L_1, L_2, L_3)$, where $\{L_i | i \in [3]\}$ are defined as follows.

$$
\begin{aligned}
L_1 &:= \max_{\pi \in \Pi} \hat{R}_{m+1, \hat{f}_{m+1}}(\pi) - \hat{R}_{m+1}(\pi) - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{R}_{m+1, \hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m+1, \hat{f}_{\bar{m}}}(\pi)}{40 \bar{m}^2 \sqrt{\alpha_{\bar{m}-1} \xi_{\bar{m}}}} \\
L_2 &:= \max_{\pi \in \Pi} \hat{R}_{m+1}(\pi) - \hat{R}_{m+1, \hat{f}_{m+1}}(\pi) - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{R}_{m+1, \hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m+1, \hat{f}_{\bar{m}}}(\pi)}{40 \bar{m}^2 \sqrt{\alpha_{\bar{m}-1} \xi_{\bar{m}}}} \\
L_3 &:= |\hat{R}_{m+1, \hat{f}_{m+1}}(p_{m+1}) - \hat{R}_{m+1}(p_{m+1})| - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{R}_{m+1, \hat{f}_{\bar{m}}}(\pi_{\hat{f}_{\bar{m}}}) - \hat{R}_{m+1, \hat{f}_{\bar{m}}}(p_{m+1})}{40 \bar{m}^2 \sqrt{\alpha_{\bar{m}-1} \xi_{\bar{m}}}}
\end{aligned}
\tag{97}
$$

Note that L_3 doesn't involve any optimization and can be easily calculated. Substituting value of these estimators for L_1 and L_2, we get.

$$
\begin{aligned}
L_1 = \max_{\pi \in \Pi} \sum_{t \in S_{m,3}} \frac{1}{|S_{m,3}|} \Bigg(& \hat{f}_{m+1}(x_t, \pi(x_t)) - \frac{\pi(a_t|x_t) r_t(a_t)}{p_m(a_t|x_t)} \\
& - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{f}_{\bar{m}}(x_t, \pi_{\hat{f}_{\bar{m}}}(x_t)) - \hat{f}_{\bar{m}}(x_t, \pi(x_t))}{40 \bar{m}^2 \sqrt{\alpha_{\bar{m}-1} \xi_{\bar{m}}}} \Bigg) \\
L_2 = \max_{\pi \in \Pi} \sum_{t \in S_{m,3}} \frac{1}{|S_{m,3}|} \Bigg(& \frac{\pi(a_t|x_t) r_t(a_t)}{p_m(a_t|x_t)} - \hat{f}_{m+1}(x_t, \pi(x_t)) \\
& - \sqrt{\alpha_m \xi_{m+1}} \sum_{\bar{m} \in [m]} \frac{\hat{f}_{\bar{m}}(x_t, \pi_{\hat{f}_{\bar{m}}}(x_t)) - \hat{f}_{\bar{m}}(x_t, \pi(x_t))}{40 \bar{m}^2 \sqrt{\alpha_{\bar{m}-1} \xi_{\bar{m}}}} \Bigg)
\end{aligned}
\tag{98}
$$

Clearly, both L_1 and L_2 are cost-sensitive classification problems [see Krishnamurthy et al., 2017, for problem definition].In both, we need to find a policy (classifier) that maps contexts to arms (classes), incurring a score (cost) for each decision such that the total score (cost) is maximized (minimized). Hence the misspecification test we use only requires two calls to CSC solvers.

E.4 Simulation

We ran uniform RCT, LinUCB, LinTS, 1-RAPR, and 4-RAPR with linear function classes and an exploration horizon of 5000 on a synthetic data generating process (DGP).[24]

Data generating process. We consider four arms, i.e., $\mathcal{A} = [8]$. The context $x = (x_1, x_2)$ is uniformly sampled from four regions on the two-dimensional unit ball; and in specific, x is generated via the following distribution:

1. $\tilde{x}_1 \sim \text{Uniform}(0.8, 1.0)$
2. $\tilde{x}_2 = \sqrt{1 - \tilde{x}_1^2} \cdot z$, where $z \sim \text{Uniform}\{-1, 1\}$.
3. Sample region index $r \sim \text{Uniform}\{0, 1, 2, 3\}$:
   - if $r = 0$: $\{x_1, x_2\} = \{\tilde{x}_1, \tilde{x}_2\}$.
   - if $r = 1$: $\{x_1, x_2\} = \{\tilde{x}_2, \tilde{x}_1\}$.
   - if $r = 2$: $\{x_1, x_2\} = \{-\tilde{x}_1, -\tilde{x}_2\}$.
   - if $r = 3$: $\{x_1, x_2\} = \{-\tilde{x}_1, -\tilde{x}_2\}$.
4. The reward for each arm is 0.4 plus a linear function of the contexts. The linear parameters for the 8 arms are $\{(a, b) || a| + |b| = 1, |a|, |b| \in \{0, 0.4, 0.6, 1\}\}$. Hence, the conditional expected rewards lies in the range $[0.2, 0.6]$. Finally, the noise was sampled uniformly at random from $[-0.4, 0.4]$.

[24] The LinUCB scaling parameter was set to a default of 0.25, we similarly let $\sqrt{\xi(T, 0.5)} = 0.25 \times \sqrt{d/T}$. We also set the bloated constant of 20 in Definition 2 to be 1.

A simulation run takes less than 9 seconds for any of these algorithms on a laptop with 16GB RAM and an Apple M1 Pro chip, demonstrating the computational tractability of this approach. We provide a scatter plot (aggregating results from 50 runs) showing (i) the value of the average reward during exploration (as a proxy for cumulative regret, x-axis) and (ii) the value of the learned policy at the end of the experiment (as a proxy for simple regret of learned policy, y-axis). On simple regret, we see that 4-RAPR $\approx$ 1-RAPR $>$ RCT $>$ LinUCB $>$ LinTS. On cumulative regret, we see that LinUCB $>$ 1-RAPR $>$ 4-RAPR $>$ LinTS $>$ RCT. The RAPR algorithms achieve the best simple regret performance and achieve competitive performance on cumulative regret for this DGP. However, the fact that both RAPR algorithms learn policies of similar values suggests that our CASs are larger than necessary for at least some risk levels on this DGP. Further refining CAS is an important direction of future work.