Towards Costless Model Selection in Contextual Bandits: A Bias-Variance Perspective

Sanath Kumar Krishnamurthy* **Adrienne Margaret Propp*** **Susan Athey**
Stanford University

Abstract

Model selection in supervised learning provides costless guarantees as if the model that best balances bias and variance was known a priori. We study the feasibility of similar guarantees for cumulative regret minimization in the stochastic contextual bandit setting. Recent work (Marinov and Zimmert, 2021) identifies instances where no algorithm can guarantee costless regret bounds. Nevertheless, we identify benign conditions where costless model selection is feasible: gradually increasing class complexity, and diminishing marginal returns for best-in-class policy value with increasing class complexity. Our algorithm is based on a novel misspecification test, and our analysis demonstrates the benefits of using model selection for reward estimation. Unlike prior work on model selection in contextual bandits, our algorithm carefully adapts to the evolving bias-variance trade-off as more data is collected. In particular, our algorithm and analysis go beyond adapting to the complexity of the simplest realizable class and instead adapt to the complexity of the simplest class whose estimation variance dominates the bias. For short horizons, this provides improved regret guarantees that depend on the complexity of simpler classes.

1 INTRODUCTION

Contextual bandit algorithms are a fundamental tool for sequential decision making and have been the focus of an increasing amount of research in recent decades (Lattimore and Szepesvári, 2020). These algorithms have been used in a wide range of applications from recommendation systems (Agarwal et al., 2016) to mobile health (Tewari and Murphy, 2017).

We study the finite-armed, stochastic contextual bandit setting. In each round, the learner observes a feature vector, or context, drawn from a fixed distribution. The learner then selects an action and receives a reward that is a function of both the context and action. The data collected in each round is incorporated into the decision-making framework for the next round, with the goal of minimizing cumulative regret, i.e. maximizing the rewards received during the experiment.

A common approach to contextual bandits, which we call the regression-based approach, hinges on estimating the reward model. In each round, the data collected over prior rounds is used to estimate the true conditional expected reward for any context and action. When the next context is observed, the estimated reward is used to construct an action selection rule to balance two objectives: reduce uncertainty in the estimate for future rounds (exploration), and maximize the reward received in the current round (exploitation). This approach has led to the development of several contextual bandit algorithms (e.g. Agrawal and Goyal, 2013; Li et al., 2010; Foster and Rakhlin, 2020). In general, the analyst specifies a model class $\mathcal{F}$ for the true reward model, and as data is gathered, the algorithm updates its selection from the class. When we assume realizability – that is, that the true reward model lies in $\mathcal{F}$ – these algorithms ensure optimal minimax guarantees on regret.

However, these algorithms do not specify how the model class $\mathcal{F}$ should be chosen, motivating recent work on model selection in contextual bandits (Agarwal et al., 2017; Foster et al., 2019). In a COLT 2020 open problem, Foster et al. (2020b) pose a key question: given a set of M nested model classes $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M$, such that at least one of these classes is realizable, can a contextual bandit algorithm achieve the best regret guarantees ensured by regression-based algorithms for

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

*Denotes significant/equal contributors.

a single model class? We refer to this as a costless model selection guarantee.[1]

In this work, we present Mod-IGW, an algorithm that achieves costless model selection under mild structural assumptions of: 1) gradually increasing class complexity, and 2) diminishing best-in-class policy improvement with increasing class complexity (diminishing marginal returns). As we discuss in Section 1.2, our assumptions reflect a natural setting for model selection. Further, even without such assumptions, our algorithm still achieves state-of-the-art (SOTA) guarantees (though not costless).

Our algorithm also addresses the bias-variance tradeoff inherent in contextual bandits (Foster et al., 2020a; Krishnamurthy et al., 2021a,b), a topic that remains unexplored in the literature on model selection for contextual bandits. Existing work focuses on adapting to the complexity of the smallest realizable class, attempting to identify the single best-performing algorithm. We propose that attempting to find a single best-performing algorithm for all time horizons may not be the most effective strategy because simpler model classes provide better guarantees for shorter time horizons, while more expressive classes outperform in longer time horizons. We therefore argue that costless model selection should adapt to the simplest class where variance dominates bias. The difficulty in achieving such costless guarantees lies in detecting when the unknown bias of a class starts dominating its variance, and correcting for the potential under-exploration costs involved with delays in this detection.

We overcome this challenge to achieve our new definition of costless model selection through two main innovations:

1. We develop a new misspecification test based on the accuracy of estimated reward models in evaluating policies from different classes. A key property of this test is that it fails (with high probability) after reward model bias dominates variance but before policy class bias dominates variance, allowing us to smoothly navigate the bias-variance tradeoff. The misspecification test is of independent interest. Subsequent work has used this test to enable efficient pure exploration algorithms without assuming realizability (Krishnamurthy et al., 2023).
2. We quantify the cost of potential under-exploration due to delays in detecting when reward model bias dominates variance, and develop a method called "self-correction" to resolve any potential under-exploration.

[1] For brevity, we use the term costless guarantees to also capture near-costless guarantees, where we ignore terms logarithmic in number of model classes, model complexities, number of rounds, and confidence parameters.

1.1 Related Work

While Marinov and Zimmert (2021) have already responded to the COLT 2020 open problem in the negative, we argue that this result is too pessimistic. Their specific counterexample of one very simple and one very complex class is an unfavorable setting for model selection and unnecessary to enforce in practice. Building on recent work quantifying the bias-variance tradeoff in contextual bandits (Foster et al., 2020a; Krishnamurthy et al., 2021a,b) and literature reducing contextual bandit problems to supervised learning tasks (Langford and Zhang, 2007; Dudik et al., 2011; Agarwal et al., 2012, 2014; Foster et al., 2018; Foster and Rakhlin, 2020) we show that it is indeed possible to achieve costless regret bounds under mild structural assumptions.

Existing algorithms for model selection in contextual bandits can generally be described as adopting either sequential (e.g. Foster et al., 2019) or parallel (e.g. Agarwal et al., 2017) search strategies. These are two alternative approaches to addressing the main challenge of model selection in contextual bandits: balancing exploration and exploitation in classes of increasing complexity. Both strategies consider bandit algorithms corresponding to models from each class $\mathcal{F}_i$, and try to identify the "best" (simplest realizable) model class such that the corresponding algorithm minimizes regret.

Sequential search strategies have largely focused on model selection over a nested sequence of linear classes $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_M$ that are linear over a nested sequence of feature maps (e.g. Foster et al., 2019). In this strategy, contextual bandits are run in sequence with increasing class complexity. For each model class $\mathcal{F}_i$, some share of rounds are devoted to sampling arms uniformly at random and testing for misspecification, with the ultimate goal of identifying the smallest realizable class, $\mathcal{F}_{i^*}$, for $i^* \in [M]$. To achieve costless guarantees, these strategies rely on stringent distributional assumptions for model identification known as diversity conditions – i.e., they assume that the minimum eigenvalue of the covariance matrix for these feature maps is greater than some positive constant.[2] This is in contrast to our algorithm which has no such requirements on the feature distribution.

Parallel search strategies use master algorithms (e.g. Agarwal et al., 2017) to run M contextual bandit algorithms in parallel, one for each of the M classes. The master algorithm allocates rounds to the M base algo-

[2] This may not be easily satisfied for feature maps with many correlated features.

rithms, learns which algorithm maximizes expected cumulative reward, and ultimately allocates most rounds to this algorithm. Since the introduction of this approach, several master algorithms have been proposed (e.g. Arora et al., 2021; Pacchiano et al., 2020b,a). However, none achieve costless model selection.

1.2 Preliminaries

The stochastic contextual bandit setting is defined by a set of contexts $\mathcal{X}$, a finite set of arms $\mathcal{A} = \{1, ..., K\}$, and a distribution D over contexts and arm rewards. At every time-step $t \in [T]$, nature samples a context $x_t \in \mathcal{X}$ and reward vector $r_t \in [0,1]^K$ from the fixed but unknown distribution D. Upon observing context x_t, the learner chooses an arm a_t and receives a reward $r_t(a_t)$. Unless stated otherwise, all expectations are taken with respect to D.

We let $f^* : \mathcal{X} \times \mathcal{A} \to [0,1]$ denote the true conditional expected reward function given contexts and actions; i.e. $f^*(x,a) := \mathbb{E}[r_t(a)|x_t = x]$. A model f is a map from $\mathcal{X} \times \mathcal{A}$ to $[0,1]$, and a model class $\mathcal{F}$ is a set whose elements are models. A policy π is any function that maps contexts to a distribution over arms, and a policy class Π is a set of policies. For deterministic policies, $\pi(x)$ denotes the arm recommended by policy π at context x, and for randomized policies, $\pi(a|x)$ denotes the probability of sampling arm a at context x. For any model f, we let π_f denote the deterministic policy induced by the model f, that is $\pi_f(x) := \arg\max_a f(x,a)$ for every x.[3] We let π^* denote the policy that maximizes the conditional mean reward; i.e., $\pi^*(x) = \arg\max_a f^*(x,a)$.

We use the term exploration policy to refer to any randomized policy that our algorithm constructs for use in exploration. For any exploration policy p, 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 $a \sim p(\cdot|x)$. We let p_t denote the exploration policy for round t.

For any model f and policy π, we let $f(x,\pi(x)) := \mathbb{E}_{a\sim\pi(x)}[f(x,a)]$ at every context x, and we let $R_f(\pi)$ denote the expected instantaneous reward of the policy π with respect to model f:

$$R_f(\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}[f(x,\pi(x))]. \tag{1}$$

Similarly, we let $\mathrm{Reg}_f(\pi)$ denote the expected instantaneous regret for policy π with respect to model f:

$$\mathrm{Reg}_f(\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}[f(x,\pi_f(x)) - f(x,\pi(x))]. \tag{2}$$

When there is no possibility of confusion, we write $R(\pi)$ and $\mathrm{Reg}(\pi)$ to mean $R_{f^*}(\pi)$ and $\mathrm{Reg}_{f^*}(\pi)$ respectively.

[3] Where ties are broken with any fixed tie-breaking rule.

In this paper, we study contextual bandit algorithms that minimize expected cumulative regret CReg_T:

$$\mathrm{CReg}_T := \sum_{t=1}^{T} \mathrm{Reg}_{f^*}(p_t). \tag{3}$$

We consider M reward model classes $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M$. We let parameter d_i denote a bound on the complexity of class $\mathcal{F}_i$. Without loss of generality, for all $i \in [M]$, we require $d_i \in \{2^q | q \in \mathbb{N}\}$.[4] For notational convenience, we group the model classes $\mathcal{F}_1, \ldots, \mathcal{F}_M$ in terms of their complexities. Let M' be the number of unique parameters in the set $\{d_i | i \in [M]\}$, and let $\tilde{d}_i$ be the i-th smallest parameter in this set such that $\tilde{d}_1 \leq \tilde{d}_2 \leq \cdots \leq \tilde{d}_{M'}$. For all $i \in [M']$, we then define model class $\tilde{\mathcal{F}}_i$ and corresponding policy class $\tilde{\Pi}_i$:

$$\tilde{\mathcal{F}}_i := \bigcup_{\{j|d_j \leq \tilde{d}_i\}} \mathcal{F}_j, \quad \tilde{\Pi}_i := \{\pi_f | f \in \tilde{\mathcal{F}}_i\}. \tag{4}$$

We let π_i^* denote the policy that maximizes the conditional mean reward among those belonging to class $\tilde{\Pi}_i$, that is $\pi_i^*(x) = \mathrm{argmax}_{\pi\in\tilde{\Pi}_i} f^*(x,\pi(x))$. Similarly, we let $\mathrm{Reg}_i(\pi)$ denote the true expected instantaneous regret against the best policy in class $\tilde{\Pi}_i$:

$$\mathrm{Reg}_i(\pi) := \max_{\tilde{\pi}\in\tilde{\Pi}_i} R(\tilde{\pi}) - R(\pi). \tag{5}$$

We can also define the bias (misspecification error) of policy class $\tilde{\mathcal{F}}_i$ as:

$$\beta_i := R(\pi^*) - R(\pi_i^*). \tag{6}$$

To quantify how well class $\tilde{\mathcal{F}}_i$ can approximate f^*, we use the definition of average squared misspecification error studied in Krishnamurthy et al. (2021a). Similar definitions of misspecification were studied in Foster et al. (2020a) and Krishnamurthy et al. (2021b). We denote by B_i the average squared misspecification error for the class $\tilde{\mathcal{F}}_i$, that is:

$$B_i := \max_p \min_{f\in\tilde{\mathcal{F}}_i} \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p(\cdot|x)}[(f(x,a) - f^*(x,a))^2], \tag{7}$$

where $D_{\mathcal{X}}$ is the marginal distribution of D on the set of contexts $\mathcal{X}$. We label model class $\tilde{\mathcal{F}}_i$ as misspecified if $B_i > 0$, and as well-specified or realizable if $B_i = 0$. Note the difference in scales between our two measures of misspecification error: B_i captures squared error, while β_i captures non-squared error. For notational convenience, we let $\beta_0 = B_0 = 1$ and $\tilde{d}_0 = 0$.

Assumption 1 (Realizability)**.** *We assume that there exists a class index $i \in [M']$ such that $B_i = 0$, and we let i^* denote the smallest class index with zero squared misspecification error.*

[4] This is without loss of generality because we can always round d_i up to the nearest exponent of 2, only increasing excess risk bounds by a constant multiplicative factor.

Assumption 1 is standard and allows Mod-IGW to reliably use more complex classes upon detecting misspecification. Assumptions 2 and 3 formalize our structural conditions of gradually increasing class complexity and diminishing marginal returns from increasing class complexity.

Assumption 2 (Gradually Increasing Class Complexity). *For any $i \in [i^*]$, there exists a class index $j \in [M']$ such that $\tilde{d}_i < \tilde{d}_j \leq \omega \tilde{d}_i$ for some fixed but unknown constant $\omega > 1$.*

Assumption 3 (Diminishing Policy Improvement with Increasing Class Complexity). *Consider any $i \in [M']$ such that $B_i > 0$. Let $j \in [M']$ be the largest index such that $\tilde{d}_j \leq \omega \tilde{d}_i$. We denote by Δ_i the improvement in the best-in-class policy value obtained by moving from policy class i to policy class j, defined as follows:*

$$\Delta_i := \max_{\pi \in \tilde{\Pi}_j} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}}[f^*(x, \pi(x))] - \max_{\pi \in \tilde{\Pi}_i} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}}[f^*(x, \pi(x))].$$

We assume that Δ_i is non-increasing in class index i.

Both Assumptions 2 and 3 are parameterized by an unknown parameter $\omega > 1$ quantifying the statistical hardness of the instance. In particular, for $\omega = \tilde{d}_{i^*}/\tilde{d}_1$, the assumptions trivially hold and so do the negative results of Marinov and Zimmert (2021) – in this case we achieve SOTA (though not costless) guarantees. However, these assumptions are often satisfied for much smaller ω – in which case we achieve costless model selection guarantees.

Assumption 2 can be ensured with small ω by construction, as it is always possible to add additional classes for a small cost. Assumption 3 with small ω is standard in the fields of statistics and machine learning; evidence of this can be found in recent work on neural scaling laws, where loss scales as a power-law with model size (Kaplan et al., 2020; Hestness et al., 2017); decision trees and random forests, where it is well-known that larger complexity parameters and more terminal nodes offer diminishing returns (Boehmke and Greenwell, 2019; Kuhn, 2022); and empirical results of discrete event simulation (Robinson, 2023).

2 ORACLES

There is a long line of work that reduces contextual bandit algorithms to oracle subroutines (e.g., Agarwal et al., 2014; Foster et al., 2018; Foster and Rakhlin, 2020). We take a similar approach, and describe our key oracle subroutines and assumptions in this section.

To estimate models in class $\tilde{\mathcal{F}}_i$, we use a model selection oracle over the set $\{\mathcal{F}_k | d_k \leq \tilde{d}_i\}$. In Oracle Assumption 1, we state our requirements for this oracle.

Oracle Assumption 1 (Estimation Oracle). *For all $j \in [M']$, we assume access to an offline model selection oracle for estimation (EstOracle$_j$) over classes $\{\mathcal{F}_k | d_k \leq \tilde{d}_j\}$ satisfying the following property: There exists a constant $C_0 \geq 1$ such that for any exploration policy p, any natural number n, and any $\zeta \in (0, 1)$, the following holds with probability at least $1 - \zeta$:*

$$\begin{aligned} &\mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p(\cdot|x)}[(\hat{f}(x, a) - f^*(x, a))^2] \\ &\qquad \leq \min_{i \in [j]} \Big(C_0 B_i + \xi_i(n, \zeta) \Big). \end{aligned} \tag{8}$$

Here, $\hat{f}$ is the output of EstOracle$_j$ fitted on n independently and identically drawn samples from $D(p)$, B_i is defined in (7), and ξ_i is given by:

$$\xi_i(n, \zeta) := C_1 \left(\frac{\tilde{d}_i \ln(nM/\zeta)}{n} \right)^{\rho}, \tag{9}$$

for some known constant $C_1 > 0$ and $\rho \in (0, 1]$.[5]

We refer to $\xi_i(\cdot, \cdot)$ as the estimation rate for model class $\tilde{\mathcal{F}}_i$ as it can be used to bound the squared prediction error of a regression oracle on model class $\tilde{\mathcal{F}}_i$. In Appendix H.1, we outline one of many approaches to construct an oracle that achieves the "fast rates" of Oracle Assumption 1. The approach we describe there is based on validation with the holdout method. Other potential approaches include cross validation, aggregation algorithms (see Lecué et al., 2014, and references therein), and penalized regression (see relevant chapters in Koltchinskii, 2011; Wainwright, 2019).

An important component of Mod-IGW is the comparison of a given policy's true value with its value according to estimated reward models. These tests verify the accuracy of estimated reward models, allowing us to detect misspecification. Oracle Assumptions 2 and 3 provide rates for estimating these quantities.

Oracle Assumption 2 (Direct Method Policy Estimation Rate). *For any index $i \in [M']$, any set of $M' + 1$ policies $\{q_0, q_1, \ldots, q_{M'}\}$, any reward model f, any natural number n, and any $\zeta \in (0, 1)$, the following holds with probability at least $1 - \zeta$:*

$$\begin{aligned} &\left| \frac{1}{n} \sum_{x \in S} f(x, \pi(x)) \ - \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}}[f(x, \pi(x))] \right| \\ &\qquad \leq \sqrt{\xi_i(n, \zeta)}, \ \forall\, \pi \in \tilde{\Pi}_i \cup \{q_0, q_1, \ldots, q_{M'}\}, \end{aligned}$$

where S is a set of n independently and identically drawn samples from the distribution $D_{\mathcal{X}}$.

Oracle Assumption 2 often follows from uniform convergence arguments (see Shalev-Shwartz and Ben-David,

[5] That is, the function $\xi_i(\cdot, \cdot)$ is known and can be used to compute (exploitation) parameters of our algorithm.

2014; Koltchinskii, 2011; Wainwright, 2019). For example, for finite function classes with $\xi_i(n,\zeta) = \mathcal{O}(\ln(|\mathcal{F}_i|/\zeta)/n)$, Oracle Assumption 2 follows from Hoeffding's inequality with uniform convergence.

Oracle Assumption 3 (Policy Evaluation Oracle). *For any index $i \in [M']$, any set of $M'+1$ policies $\{q_0, q_1, \ldots, q_{M'}\}$, any natural number n, any exploration policy p with $p(a|x) \geq \eta$ for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$, and any $\zeta \in (0,1)$, the following holds with probability at least $1-\zeta$:*

$$\left|\widehat{R}(\pi) - \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}[f^*(x,\pi(x))]\right| \leq \frac{\xi_i(n,\zeta)}{\eta} + \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\frac{1}{p(\pi(x)|x)}\right]\xi_i(n,\zeta)}, \forall\, \pi \in \tilde{\Pi}_i \cup \{q_k\}_{k=0}^{M'},$$

where $\widehat{R}$ is the output of EvalOracle$_i$ fitted on n independently and identically drawn samples from $D(p)$.

When $\hat{R}$ is estimated via inverse propensity score estimation, Agarwal et al. (2014) show that Oracle Assumption 3 is satisfied for finite function classes with $\xi_i(n,\zeta) = \mathcal{O}(\ln(|\mathcal{F}_i|/\zeta)/n)$. The covering arguments in Maurer and Pontil (2009) can be used to show Oracle Assumption 3 holds for general function classes. [6]

3 ALGORITHM

We present our algorithm, Mod-IGW, in Algorithm 1. As its name suggests, Mod-IGW is based on an inverse gap weighting (IGW) approach to action selection, which provides a simple analytical handle on important quantities like the expected inverse probability weight for any policy at any round, and is often used to develop optimal algorithms (Abe and Long, 1999; Foster and Rakhlin, 2020; Foster et al., 2020c; Simchi-Levi and Xu, 2020). However, Mod-IGW deviates from the standard IGW approach in three major respects: 1) how the model is estimated – specifically, using EstOracle (e.g. classical model selection for supervised learning; see Section 2); 2) how the exploitation parameter is determined – using a novel misspecification test that selects among M' candidate exploitation parameters (Section 3.1); and 3) how the candidate exploitation parameters scale with the number of rounds – using a "self-correction" strategy that accounts for any potential under-exploration in earlier rounds and updates the exploitation parameter accordingly (Section 3.2).

Mod-IGW proceeds in epochs indexed by m, with epoch m spanning time-steps $t \in [\tau_{m-1}+1, \tau_m]$. At any such t, the algorithm observes context x_t and samples action a_t from the distribution of exploration policies p_m, defined as:

$$p_m(a|x) := \begin{cases} \frac{1}{K+\gamma_m\left(\hat{f}_m(x,\hat{a})-\hat{f}_m(x,a)\right)}, & a \neq \hat{a}, \\ 1 - \sum_{a'\neq\hat{a}} p_m(a'|x), & a = \hat{a}. \end{cases} \tag{10}$$

Here, $\hat{f}_m$ is an estimate of the reward model computed via EstOracle with data from previous epochs, and $\hat{a} = \arg\max_a \hat{f}_m(x,a)$ is the predicted best action. The exploitation parameter γ_m governs the balance between exploration and exploitation. The higher the value of γ_m, the greater the probability that the greedy action $\hat{a}$ is chosen. The remainder of this section focuses on how the exploitation parameter γ_m should be chosen.

To optimize cumulative regret, we want to exploit as much as possible while still allowing for estimation of useful reward models. Following the IGW approach (Foster and Rakhlin, 2020; Simchi-Levi and Xu, 2020), γ_m should be specified based on the reward model error to balance this tradeoff. A common metric for gauging reward model error is the mean squared prediction error, which decomposes into bias and variance. Bias (B_i) is unknown and can be challenging to estimate, but in early rounds when the variance ($\xi_i(\cdot,\cdot)$) dominates the bias, we can bound the squared error by $2\xi_i(\cdot,\cdot)$ (similar ideas were used to quantify the bias-variance tradeoff in contextual bandits in Krishnamurthy et al., 2021a). Following prior IGW approaches, we can use this candidate bound on the squared prediction error to set the candidate exploitation parameter $\gamma_{m+1,i}$:

$$\gamma_{m+1,i} := \max\left(K, \sqrt{\frac{K}{8\xi_i\left(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\right)}}\right), \tag{11}$$

and corresponding exploration policies $p_{m+1,i}$ (by substituting $\gamma_{m+1} = \gamma_{m+1,i}$ in the formula for p_{m+1}). Parameter $\gamma_{m+1,i}$ induces sufficient exploration so long as the variance of estimating a reward model from class i dominates the bias of that class.[7] We refer to the (unknown) last epoch where variance dominates bias as the "safe epoch," denoted by m_i^*:[8]

$$m_i^* := \max\left\{m \middle| \xi_i\left(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\right) \geq C_0 B_i\right\}.$$

We let $m_0^* = 1$ and note that $m_{i^*}^*$ is infinity.[9] Note that since ξ_i is increasing in i, $\gamma_{m+1,i}$ is non-increasing in i. Therefore, to maximize exploitation (while ensuring enough exploration to estimate a useful model), we

[6] For example, Jin (2023) show Oracle Assumption 3 holds for tree-based and neural network function classes with appropriate choices of $\xi(\cdot,\cdot)$.

[7] Estimation variance decreases with more data and is eventually dominated by bias.

[8] m_i^* is unknown because the bias B_i is unknown.

[9] That is, the exploitation parameter corresponding to the realizable class always induces sufficient exploration.

want to use the exploitation parameter corresponding to the simplest class $i \in [M']$ such that $m \in [m_i^*]$.

Since bias (B_i), and hence the safe epoch (m_i^*) are unknown, we don't know which of the candidate exploitation parameters to choose for a given epoch m. The challenge of estimating bias stems from the difficulty of estimating the prediction error of the estimated reward model.[10] Note that we can't extract measures of mean prediction error from mean squared error (or other similar measures of error that average absolute differences between outcomes and predicted values) due to unknown irreducible noise.[11]

To overcome this, we develop a new way of testing if these variance-based candidate error bounds actually bound the prediction error of the estimated model. In particular, we posit a shift in perspective from measuring error via the mean squared prediction error, and instead gauge the error of an estimated reward model by its accuracy in evaluating candidate policies. We do this by comparing an estimated reward model's direct method estimates (see Oracle Assumption 2) with consistent policy estimates (see Oracle Assumption 3). Importantly, by leveraging the consistent estimators available for policies (which average out reward noise), we are able to capture the real prediction error of the estimated model in evaluating the candidate policies.[12] This policy-based approach forms the foundation for our misspecification test, MTOracle (described in the next section).

3.1 Misspecification Test

As discussed above, a key challenge for our exploration strategy is the specification of the exploitation parameter γ_m given that the safe epochs m_i^* are unknown. To address this challenge, we introduce a new misspecification test: MTOracle. This section describes the test in detail. In subsequent work, this test enables efficient pure exploration in contextual bandits (Krishnamurthy et al., 2023).

MTOracle adopts a policy-based approach to assess estimated models. With high-probability, the test detects misspecification for class i after its reward model bias dominates variance (that is, after the corresponding safe epoch) but before its policy class bias dominates variance. This allows Mod-IGW to explore with the exploitation parameter corresponding to the simplest class whose variance dominates bias. There are three components to MTOracle, given in Oracle Assumption 4. Each of these is sufficient to rule out classes whose bias dominates variance.

Algorithm 1 Mod-IGW (Model Selection with Inverse Gap Weighting)

input: Initial epoch length $\tau_1 \geq 2$, horizon T, and confidence parameter δ.

Let $\hat{f}_1 \equiv 0$, $i_1 = 1$, $\gamma_1 = 1$, $\tau_0 = 0$, and $\hat{m} = 0$
for epoch $m = 1, 2, \ldots$ **do**
  Let p_m be given by (10)
  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. Split S_m into training ($S_{m,\mathbf{tr}}$) and holdout ($S_{m,\mathbf{ho}}$) sets of roughly equal size ($|S_{m,\mathbf{ho}}| = \lceil |S_m|/2 \rceil$)
  $\hat{f}_{m+1} \leftarrow \text{EstOracle}_{M'}(S_{m,\mathbf{tr}})$
  $\hat{f}_{m+1,i} \leftarrow \text{EstOracle}_i(S_{m,\mathbf{tr}}) \quad \forall i \geq i_m$
  $\hat{R}_{m+1} \leftarrow \text{EvalOracle}_i(S_{m,\mathbf{ho}})$
  $\hat{R}_{m+1,f}(\pi) := \frac{1}{|S_{m,\mathbf{ho}}|} \sum_{(x,a,r) \in S_{m,\mathbf{ho}}} f(x, \pi(x))$
  Let $i_{m+1} = \text{MTOracle}_{i_m}(S_{m,\mathbf{ho}})$
  if $i_{m+1} \neq i_m$ **then**
    $\hat{m} \leftarrow m + \lceil \log_2(\log_2(\gamma_{m,1}/\gamma_{m,i_{m+1}})) \rceil$
  end if
  $\tau_{m+1} \leftarrow \tau_m + (1 + \mathbb{1}\{m \geq \hat{m}\})(\tau_m - \tau_{m-1})$.
  $\gamma_{m+1} \leftarrow \gamma_{m+1,i_{m+1}}$.
end for

The main policy-based misspecification test in MTOracle checks whether the estimated reward models can be used to construct sufficiently accurate direct method estimates of policy values. We test this by comparing $\hat{R}_{m+1}(\pi)$, the estimate of a policy value obtained via EvalOracle(S_m), and $\hat{R}_{m+1,f}(\pi)$, the direct method estimate of a policy value under some estimated reward model f, defined by:

$$\hat{R}_{m+1,f}(\pi) := \frac{1}{|S_{m,\mathbf{ho}}|} \sum_{(x,a,r) \in S_{m,\mathbf{ho}}} f(x, \pi(x)). \quad (12)$$

If the difference between these two estimates surpasses the threshold given in Oracle Assumption 4 for some $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, this indicates the bias of class i dominates the variance of estimating from class i. In other words, we are underestimating the error of the estimated reward model, and so parameter $\gamma_{m+1,i}$ does not induce sufficient exploration.

Oracle Assumption 4. *(Misspecification Test Oracle) In each epoch m, our misspecification test MTOracle$_{i_m}$ identifies index i_{m+1}, which we define as the smallest*

[10] Asymptotically, prediction error of the estimated reward model converges to squared misspecification error.

[11] For example, while the mean squared error can be empirically estimated, it converges to the sum of mean squared prediction error and irreducible noise.

[12] We are unavoidably limited in the number of policies used to perform the comparison (due to issues related to multiple hypothesis testing), and using larger policy classes for the comparison requires more exploration data.

index such that $i_{m+1} \geq i_m$ and, for all $i \geq i_{m+1}$, $j \geq i$, $h \leq i_m$, and $\alpha > 0$, the following inequalities hold:

Policy-based misspecification test

$$\begin{aligned}
&|\hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi)| \\
&\leq \bigg(\frac{1+\theta_{i,j}}{\alpha} + \frac{(1+\theta_{i,j})\alpha}{16} \\
&\quad + \frac{(2\theta_{i,j}^2 + (1+\theta_{i,j})^2/\alpha)\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j} \bigg) \frac{K}{\gamma_{m+1,i}} \\
&\quad + \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{Reg}_{m+1,\hat{f}_m}(\pi) \\
&\quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}, \pi \in \tilde{\Pi}_j \cup \Pi_{0,m+1,i},
\end{aligned}$$

Reward model agreement

$$\begin{aligned}
&\widehat{Reg}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \\
&\quad \leq 26 \frac{\gamma_{m,h}}{\gamma_{m,i}} \left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}} \right)^{1/2^{m-\hat{m}_{h-1}}} \frac{K}{\gamma_{m+1,i}}, \\
&\widehat{Reg}_{m+1,\hat{f}_m}(\pi) \\
&\quad \leq 4\widehat{Reg}_{m+1,f}(\pi) \\
&\quad\quad + 34 \frac{\gamma_{m,h}}{\gamma_{m,i}} \left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}} \right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}} \frac{K}{\gamma_{m,i}}, \\
&\quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}, \pi \in \Pi_{0,m+1,i},
\end{aligned}$$

where $\Pi_{0,m+1,i} = \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, $\theta_{i,j} = (\tilde{d}_j/\tilde{d}_i)^{\rho/2}$, *and* $\hat{m}_i := \max\{m | i_m \leq i\}$ ($\hat{m}_0 := 0$)*. That is, $\hat{m}_i$ is the latest epoch such that model class i has not been labeled as misspecified. The above inequalities are derived in Lemmas 7, 13 and 16, respectively. Index i_{m+1} is the output of MTOracle$_{i_m}$.*

MTOracle also includes two tests to verify that the estimated reward model exhibits agreement across possibly well-specified classes and across epochs. The first component confirms that the policy induced by $\hat{f}_{m+1,i}$ (the model estimated for class i) is a good policy according to $\hat{f}_{m+1}$ (the model estimated across all classes). In this case, $\widehat{\mathrm{Reg}}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})$ should not exceed the threshold given in Oracle Assumption 4, where $\widehat{\mathrm{Reg}}_{m,f}(\pi)$ denotes the empirical regret for policy π with respect to model f:

$$\widehat{\mathrm{Reg}}_{m,f}(\pi) = \hat{R}_{m,f}(\pi_f) - \hat{R}_{m,f}(\pi). \tag{13}$$

This helps ensure that once $\hat{f}_{m+1}$ believes a notably better policy lies in a larger policy class, we use candidate exploitation parameters corresponding to larger classes to ensure sufficient exploration.

To ensure reward model agreement across epochs, the final component of MTOracle confirms that the candidate exploration policies ($\Pi_{0,m+1,i}$, defined in Oracle Assumption 4) have sufficiently low regret under the prior epoch's estimated reward model. Thus, $\widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi)$ should not exceed $\widehat{\mathrm{Reg}}_{m+1,f}(\pi)$ for $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$ for policies in $\Pi_{0,m+1,i}$ beyond the threshold given in Oracle Assumption 4. This test helps confirm that candidate exploration policies for epoch $m+1$ were sufficiently explored in epoch m.

In Appendix H.2, we describe one approach to computationally test the inequalities in Oracle Assumption 4 via cost-sensitive classification. Note that not only do the tests in MTOracle check whether reward model bias dominates variance, but they also verify the accuracy of the estimated reward models in evaluating policies from different classes.

3.2 Self-Correction

To recap, for any epoch m, the goal of Oracle Assumption 4 is to verify that the prediction error of model $\hat{f}_m$ can be bounded by the variance of estimating from the class $\tilde{\mathcal{F}}_{i_m}$. This verification involves testing if $\hat{f}_m$ can accurately evaluate policies from classes of various complexities up to this error bound. Unfortunately, this verification is loose up to a factor $(\tilde{d}_j/\tilde{d}_{i_m})^{\rho/2}$, for policy class $\tilde{\Pi}_j$ more complex than $\tilde{\Pi}_{i_m}$. The detection of misspecification indicates that in past epochs, the ability of $\hat{f}_m$ to evaluate policies from $\tilde{\Pi}_j$ may have been loose, up to a multiplicative factor of $(\tilde{d}_j/\tilde{d}_{i_m})^{\rho/2}$. As a result, we may have under-estimated the value of some policies up to this factor, leading to corresponding under-exploration in prior epochs. Note that under-exploration in prior rounds affects our ability to evaluate these policies well in future rounds. Hence, upon detecting misspecification, we want to correct for the effects of potential under-exploration on our estimated models.[13] Our analysis uncovers a self-correction mechanism to manage this challenge, which we describe in this section.

Upon detecting misspecification, we hold the epoch length fixed for a small number of epochs. Importantly, this results in the candidate exploitation parameters being held fixed while the algorithm continues to collect data and improve the reward model via EstOracle. As the estimated reward model improves, we better explore good policies that were previously not well-explored, leading to reward models that are better at estimating good policies. After this process continues for a small number of epochs, we will have sufficiently corrected for potential prior underexploration and can resume increasing our candidate exploitation parameters.

Mod-IGW is the first algorithm to leverage this strategy. The typical approach is to restart a bandit algorithm from scratch upon detecting misspecification

[13] The factors described here bound the potential extent of this under-exploration.

(e.g., Foster et al., 2019). However, this may lead to larger than necessary cumulative regret, particularly leading to worse bounds for shorter horizons. Self-correction, in contrast, is efficient and unintrusive, allowing Mod-IGW to recalibrate after detecting misspecification in just a few epochs.

Finally, to better understand how Assumptions 2 and 3 help with detecting misspecification sufficiently quickly, consider a misspecified class that is being used to set exploitation parameters at some round. Our assumptions ensure that there exists a slightly more complex class (Assumption 2) with a sufficiently higher best-in-class policy value (Assumption 3). Since this class is only slightly more complex, the data collected is sufficiently explorative to evaluate policies in this class up to the desired accuracy. This allows us to leverage this estimation accuracy and gap between optimum policy values to quickly detect misspecification in the considered class.

4 MAIN RESULT

We now present our main result in Theorem 1.

Theorem 1. *Suppose Assumptions 1 to 3 hold and the oracle subroutines perform as stated in Oracle Assumptions 1 to 4. With probability at least $1-\delta$: for any $i, j \in [M']$ such that $i \leq j$ and $\tilde{\mathcal{F}}_j$ not yet labelled as misspecified as of round T, Mod-IGW attains the following regret guarantee:*

$$CReg_T \leq \tilde{\mathcal{O}}\left((\omega^2 K^{1/\rho}) \frac{\tilde{d}_{i-1}}{\beta_{i-1}^{2/\rho}} + \beta_j T + \left(\frac{\tilde{d}_j}{\tilde{d}_i}\right)^{\rho/2} \sqrt{K \tilde{d}_j^{\rho} T^{2-\rho}} \right) \tag{14}$$

Here, $\tilde{\mathcal{O}}$ hides terms logarithmic in $T, M, 1/\delta, \tilde{d}_{i^*}$. Further, $\tilde{\mathcal{F}}_j$ is not determined to be misspecified for at least $\Omega(\tilde{d}_j / B_j^{1/\rho})$ rounds.

To simplify our discussion and provide more insight into Theorem 1, we focus on the implications for classes with $\rho = 1$ [14] and ignore constant factors, logarithmic factors, and ω from Assumption 3. Then we achieve a cumulative regret bound of the form $(K\tilde{d}_{i-1})/\beta_{i-1}^2 + \beta_j T + \sqrt{\tilde{d}_j/\tilde{d}_i}\sqrt{K\tilde{d}_j T}$, so long as $\tilde{\mathcal{F}}_j$ has not yet been determined to have larger bias than variance.

Let us understand these terms better. The first term, $(K\tilde{d}_{i-1})/\beta_{i-1}^2$, bounds the time to detect misspecification in class $\tilde{\mathcal{F}}_{i-1}$. This marks the number of rounds required for policy class bias (β_{i-1}) to dominate the corresponding variance for policy learning from class i under uniform sampling ($\sqrt{K\tilde{d}_{i-1}/T}$). The second term, $\beta_j T$, accounts for the bias of class $\tilde{\Pi}_j$. This term would not appear in our bound had we defined cumulative regret relative to π_j^* (the best policy in class $\tilde{\Pi}_j$). The third term, $\sqrt{\tilde{d}_j/\tilde{d}_i}\sqrt{K\tilde{d}_j T}$, is the product of two quantities. The quantity $\sqrt{K\tilde{d}_j T}$ accounts for the estimation variance for class $\tilde{\mathcal{F}}_j$. This is also the minimax regret bound for contextual bandits working with class $\tilde{\mathcal{F}}_j$, assuming realizability in this class. The quantity $\sqrt{\tilde{d}_j/\tilde{d}_i}$ accounts for potential under-exploration of policies in class $\tilde{\Pi}_j$ after self-correcting to the exploitation parameters induced by class i. Note that the multiplicative cost of $\sqrt{\tilde{d}_j/\tilde{d}_i} = 1$ for $i = j$. As stated earlier, the cumulative regret bound we discussed holds if $\tilde{\mathcal{F}}_j$ was not determined to be misspecified through round T. Note that we only detect misspecification for class $\tilde{\mathcal{F}}_j$ after the average squared misspecification error (B_j) is larger than the corresponding variance ($\tilde{d}_j/T$). That is, $\tilde{\mathcal{F}}_j$ is not determined to be misspecified for at least $\Omega(\tilde{d}_j/B_j)$ rounds. In this way, we always rely on the simplest class whose bias dominates the variance, achieving costless model selection guarantees under mild assumptions.

An important implication concerns how well we adapt to the realizable class $\tilde{\mathcal{F}}_{i^*}$. Choosing $i = j = i^*$, our cumulative regret in terms of $(K, T, \tilde{d}_{i^*})$ is given by $\tilde{\mathcal{O}}(\sqrt{K\tilde{d}_i T})$. This shows that, despite the negative result in Marinov and Zimmert (2021), under mild assumptions, it is possible to achieve the costless regret guarantees requested in the COLT 2020 open problem (Foster et al., 2020b).

As argued in Section 1, Assumptions 2 and 3 hold with small ω in many instances, allowing us to achieve the costless guarantees described in this section. However for those instances where Assumptions 2 and 3 do not hold with small ω, we show Mod-IGW can achieve any SOTA (although not costless) guarantee on the Pareto frontier of Marinov and Zimmert (2021).

To see this, note that Assumptions 2 and 3 capture instance hardness with parameter $\omega > 1$ and are always satisfied for the choice $\omega = \tilde{d}_{i^*}/\tilde{d}_1$. Recall that Mod-IGW does not need ω as an input, as it automatically adapts to instance hardness. By setting $i = 1$, we achieve regret bounds that are independent of ω (since $\tilde{d}_0 \equiv 0$), giving us our worst-case guarantees (free of Assumptions 2 and 3). Further, setting $j = i^*$, we achieve a regret bound of $\tilde{\mathcal{O}}(\sqrt{\tilde{d}_{i^*}/\tilde{d}_1}\sqrt{KT\tilde{d}_{i^*}})$. This recovers one point on the SOTA (although not costless)

[14] Estimation rates with $\rho = 1$ hold for a wide variety of popular classes, e.g. finite function classes, linear classes, and classes with finite VC-sub-graph dimension (see Koltchinskii, 2011).

Pareto-front from Marinov and Zimmert (2021).

To recover the other Pareto-front guarantees, we use looser class complexity bounds as input to Mod-IGW. Consider any $\mathcal{C} > 0$, and let $\max(\tilde{d}_i, \mathcal{C}^2)$ be the looser class complexity input to the algorithm for every class $\tilde{\mathcal{F}}_i$. This ensures the looser complexity for $\tilde{\mathcal{F}}_1$ is greater than or equal to $\mathcal{C}^2$, and the looser complexity for $\tilde{\mathcal{F}}_{i^*}$ is $\max(\tilde{d}_{i^*}, \mathcal{C}^2)$. Substituting this into the above bound that did not rely on Assumptions 2 and 3, we achieve a regret bound of $\tilde{\mathcal{O}}(\max(\mathcal{C}, \tilde{d}_{i^*}/\mathcal{C})\sqrt{KT})$. Adjusting $\mathcal{C}$ then allows us to navigate the SOTA Pareto front in Marinov and Zimmert (2021).

5 CONCLUSION

We study the feasibility of costless model selection in contextual bandits. First, we expanded the definition of costless model selection to not just adapt to the complexity of the simplest realizable class, but adapt to the complexity of the simplest class whose variance dominates the bias. This introduces the perspective of bias-variance trade-off to model selection in contextual bandits. Second, we identify mild assumptions under which costless model selection is feasible and can be achieved by our algorithm, Mod-IGW. If the unknown parameter ω in our assumptions is large enough, the assumptions we introduce are trivially satisfied – in this case, we can't achieve costless model selection guarantees, but still recover near-optimal guarantees. Our analysis is enabled by two key algorithmic insights: our policy-based misspecification test and self-correction.

The policy-based misspecification test we develop here, in particular, is of broad interest – allowing us to sidestep the challenge of unknown reward model class bias. It has been used to develop oracl-efficient assumption-free algorithms for pure exploration in contextual bandits (Krishnamurthy et al., 2023). The key insight is that, without the need for additional assumptions, we can gauge the error of an estimated reward model by its accuracy in evaluating candidate policies, and this error gauge is sufficient for decision-making in contextual bandits.

Acknowledgements

S.A. and S.K.K. are grateful for the generous support provided by Golub Capital Social Impact Lab and the Office of Naval Research grant N00014-22-1-2668. A.M.P. is grateful for the generous support of the Stanford Graduate Fellowship (SGF).

References

Abe, N. and Long, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer.

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., et al. (2016). Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. (2012). Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR.

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

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

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Arora, R., Marinov, T. V., and Mohri, M. (2021). Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR.

Boehmke, B. C. and Greenwell, B. M. (2019). Hands-on machine learning with r.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*.

Foster, D. J., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. E. (2018). Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088*.

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

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

Foster, D. J., Krishnamurthy, A., and Luo, H. (2020b). Open problem: Model selection for contextual bandits. In *Conference on Learning Theory*, pages 3842–3846. PMLR.

Foster, D. J. and Rakhlin, A. (2020). Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*.

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

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically.

Jin, Y. (2023). Upper bounds on the natarajan dimensions of some function classes. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1020–1025. IEEE.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

Koltchinskii, V. (2011). *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.

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

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

Krishnamurthy, S. K., Hadad, V., and Athey, S. (2021b). Tractable contextual bandits beyond realizability. In *International Conference on Artificial Intelligence and Statistics*, pages 1423–1431. PMLR.

Krishnamurthy, S. K., Zhan, R., Athey, S., and Brunskill, E. (2023). Proportional response: Contextual bandits for simple and cumulative regret minimization. *Advances in Neural Information Processing Systems*, 36.

Kuhn, M. (2022). Machine learning with caret in r.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer.

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

Lecué, G., Rigollet, P., et al. (2014). Optimal learning with q-aggregation. *Annals of Statistics*, 42(1):211–224.

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

Marinov, T. V. and Zimmert, J. (2021). The pareto frontier of model selection for general contextual bandits. *Advances in Neural Information Processing Systems*, 34.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Mitchell, C., van de Geer, S., et al. (2009). General oracle inequalities for model selection. *Electronic Journal of Statistics*, 3:176–204.

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

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

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

Robinson, S. (2023). Exploring the relationship between simulation model accuracy and complexity. *Journal of the Operational Research Society*, 74(9):1992–2011.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

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

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes – this is discussed throughout the paper.

The setting and assumptions are discussed in Sections 1 and 2, and the algorithm is discussed in Section 3.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes – this is discussed in Section 4.]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes – all assumptions are described clearly in Section 1 and labeled as **Assumption X**, and all black-box oracle assumptions are described clearly in Sections 2 and 3.1 and labeled as **Oracle Assumption X**.]

(b) Complete proofs of all theoretical results. [Yes – these are provided in the Appendix.]

(c) Clear explanations of any assumptions. [Yes – all assumptions are described clearly in Section 1 and labeled as **Assumption X**, and all black-box oracle assumptions are described clearly in Sections 2 and 3.1 and labeled as **Oracle Assumption X**.]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A ADDITIONAL PRELIMINARIES

We now work towards proving Theorem 1. We start with providing a proof outline, set up additional notation in Appendix A.1, and aggregate commonly used notation in Table 1.

Proof outline: Appendix A.2 provides basic well-known properties of the IGW exploration strategy, after which we get into the meat of our proof. Appendix B provides bounds on accuracy of the estimated reward model in evaluating policies via the direct method within various safe epochs (this argument is similar to the one provided in Krishnamurthy et al., 2021a). Appendix C designs and analyzes our main policy-based misspecification test, in order to test and verify the direct method implications of the estimated reward model.[15] Appendix D adds additional tests to ensure sufficient reward model agreement across epochs and classes. Appendix E then bounds the true regret of various policies with regret according to estimated models via an inductive argument. Appendix F upper bounds the time to detect misspecification for various classes under our assumptions. Finally, Appendix G uses these results to prove Theorem 1. Additional details of interest are discussed in Appendix H.

Symbol	Description
i_m	simplest possibly-well-specified model class
h, i, j	class indices such that $h \leq i \leq j$
$\tilde{d}_i$	complexity of model class $\tilde{\mathcal{F}}_i$
ξ_i	estimation rate for model class $\tilde{\mathcal{F}}_i$, defined in (9)
$\gamma_{m,i}$	exploration parameter for model class $\tilde{\mathcal{F}}_i$ in epoch m, defined in (11)
m	epoch index
m_i^*	"safe epoch" for model class i, up to which sufficient exploration is guaranteed, defined in (3)
$\hat{m}_i$	implicit estimate of safe epoch for model class i, defined in Oracle Assumption 4, where $\hat{m}_i = 0$ for $i \leq 0$
τ_m	final round of epoch m
p_m	exploration policy for epoch m
$V(p, \pi)$	expected inverse probability weight, defined in (16)
f	model mapping from contexts and actions to rewards
f^*	true conditional expectation reward function
$\hat{f}_m \in \tilde{\mathcal{F}}$	estimated reward model according to $\text{EstOracle}_M(S_{m-1,tr})$, fitted over classes $\tilde{\mathcal{F}}_1, ..., \tilde{\mathcal{F}}_M$
$\hat{f}_{m,i} \in \tilde{\mathcal{F}}$	estimated reward model according to $\text{EstOracle}_i(S_{m-1,tr})$, fitted over classes $\tilde{\mathcal{F}}_1, ..., \tilde{\mathcal{F}}_i$
π_f	optimal policy with respect to reward model f
$R_f(\pi)$	expected reward of policy π with respect to reward model f
$\hat{R}_{m+1}(\pi)$	estimated reward of policy π according to $\text{EvalOracle}(S_m)$
$\hat{R}_{m+1,f}(\pi)$	implicit estimated reward of policy π, defined in (12)
$\text{Reg}_f(\pi)$	regret of policy π with respect to reward model f: $\text{Reg}_f(\pi) = R_f(\pi_f) - R_f(\pi)$
$\text{Reg}_i(\pi)$	true expected instantaneous regret against the best policy in the class $\tilde{\Pi}_i$
$\widehat{\text{Reg}}_{m,f}(\pi)$	empirical regret of policy π with respect to model f, defined in (13)
S_m	data collected in epoch m
MTOracle	Misspecification test
EstOracle_i	Estimation oracle over model classes in $[i]$
EvalOracle	Policy evaluation oracle
Mod-IGW	Model selection with inverse gap weighting

Table 1: Table of notations

A.1 Additional Notation

The most commonly used notations in this paper are collected in Table 1. Let Γ_t denote the set of observed data points up to and including time t. That is

$$\Gamma_t := \{(x_s, a_s, r_s(a_s))\}_{s=1}^t \tag{15}$$

[15] Since safe epochs, which depend on model class bias, are unknown, we must test and verify these direct method bounds via our policy-based misspecification test.

For any randomized policy p and any policy π, we let $V(p, \pi)$ denote the expected inverse probability weight of covering π under p:

$$V(p,\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim\pi(x)}\left[\frac{\pi(a|x)}{p(a|x)}\right]. \tag{16}$$

The variance term for several policy evaluation estimators like IPW depends on this expected inverse probability weight (see e.g. Agarwal et al., 2014). We also let let $m(t)$ denote the epoch containing round t, so that $m(t) := \min\{m|t \leq \tau_m\}$.

A.2 Properties of the IGW Exploration Policy

We now state helpful properties of the exploration policy, and only include the proofs for completeness. Similar properties are explicitly stated and proven in Simchi-Levi and Xu (2020), but also show up in the analysis for Foster and Rakhlin (2020) (see section B.1 of their paper). Arguably, these properties characterize the key features of inverse gap weighting algorithms. Lemma 1 and Lemma 2 bound the estimated instantaneous regret and the expected inverse probability weight for the exploration policy constructed by inverse gap weighting.

Lemma 1. *For any epoch $m \geq 1$, we have:*

$$Reg_{\hat{f}_m}(p_m) \leq \frac{K}{\gamma_m}.$$

Proof. Note that:

$$\begin{aligned}\mathrm{Reg}_{\hat{f}_m}(p_m) &= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}} p_m(a|x)\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)\right] \\ &= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}} \frac{\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)}{K+\gamma_m\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)}\right] \leq \frac{K}{\gamma_m}.\end{aligned}$$

□

Lemma 2. *For all policy π and epochs $m \geq 1$, we have:*

$$V(p_m,\pi) \leq K + \gamma_m \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\left(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,\pi(x))\right)\right].$$

Proof. Consider any policy π and epoch $m \geq 1$. For any context $x \in \mathcal{X}$ and action $a \in \mathcal{A} \setminus \{\pi_{\hat{f}_m}(x)\}$, from our choice for p_m, we get:

$$\frac{1}{p_m(a|x)} = K + \gamma_m(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)).$$

For the action $a = \pi_{\hat{f}_m}(x)$, we have:

$$\frac{1}{p_m(a|x)} = \frac{1}{1-\sum_{a'\neq a}\frac{1}{K+\gamma_m\left(\hat{f}_m(x,\pi_{\hat{f}_m}(x))-\hat{f}_m(x,a')\right)}} \leq K$$

In particular, putting the above inequality together, we get:

$$\frac{\pi(a|x)}{p_m(a|x)} \leq \frac{1}{p_m(a|x)} \leq K + \gamma_m\left[\left(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\right)\right].$$

The lemma now follows by taking expectation over $x \sim D_{\mathcal{X}}, a \sim \pi(x)$. □

B DIRECT METHOD GUARANTEES FOR ESTIMATED MODELS

We judge our estimated reward model by its ability to evaluate policies via the direct method. In this section we prove direct method bounds that should hold with high-probability up to various safe epochs.

B.1 High Probability Events For Regression

In this section, we define an event $\mathcal{W}_1$ that holds with high probability under Oracle Assumption 1. At a high level, $\mathcal{W}_1$ defines the event where the prediction guarantees of EstOracle hold. That is, this event bounds the expected squared error difference between the true model (f^*) and the estimated model ($\hat{f}_{m+1}$).

$$\mathcal{W}_1 := \Bigg\{ \forall m \in [m_i^*], j \in [i, M'], \\ \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p_m(\cdot|x)} [(\hat{f}_{m+1,j}(x,a) - f^*(x,a))^2] \leq 2\xi_i\Big(\frac{\tau_m - \tau_{m-1}}{2}, \delta/(6TM'^2)\Big)\Bigg\}. \tag{17}$$

In Lemma 3, we use standard union bound arguments to show that the event $\mathcal{W}_1$ holds with high probability.

Lemma 3. *Suppose Oracle Assumption 1 holds. Then the event $\mathcal{W}_1$ holds with probability at least $1-\delta/2$.*

Proof. Consider any epoch m. Note that, conditional on $\Gamma_{\tau_{m-1}}$ the number of samples in epoch m are fixed and these samples are i.i.d. from the distribution $D(p_m)$.[16] Hence with probability $1-\delta/(6TM')$, from Oracle Assumption 1, for all $i \in [M']$ and $j \in [i, M']$ such that $m \in [m_i^*]$ we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p_m(\cdot|x)} [(\hat{f}_{m+1,j}(x,a) - f^*(x,a))^2] &\leq \min_{i'\in[j]}(C_0 B_{i'} + \xi_{i'}(\frac{\tau_m - \tau_{m-1}}{2}, \delta/(6TM'^2))) \\ &\leq C_0 B_i + \xi_i(\frac{\tau_m - \tau_{m-1}}{2}, \delta/(6TM'^2)) \\ &\leq 2\xi_i(\frac{\tau_m - \tau_{m-1}}{2}, \delta/(6TM'^2)),\end{aligned} \tag{18}$$

where the last inequality follows from the definition of m_i^* and the fact that $m \leq m_i^*$. Therefore, the probability that $\mathcal{W}_1$ does not hold can be bounded by:

$$\sum_{m=1}^{m(T)} \frac{\delta}{6TM'} \leq \frac{\delta}{6M'} \leq \frac{\delta}{2}.$$

□

B.2 Direct Method for Policy Optimization

Given any estimated model $\hat{f}$, $R_{\hat{f}}(\pi)$ gives us an implicit estimate for any policy π. Moreover, as discussed earlier, $\pi_{\hat{f}}$ is the policy that maximizes these implicitly estimated rewards. This approach to policy optimization is known as the direct method for policy optimization. Several papers have analyzed the direct method for policy evaluation (e.g. Qian and Murphy, 2011; Simchi-Levi and Xu, 2020; Krishnamurthy et al., 2021a).

In Lemma 4 we state a guarantee on the direct method via a model selection oracle for estimation. The proof is essentially the same as the proof of prior guarantees on the direct method.

Lemma 4. *Suppose the event $\mathcal{W}_1$ defined in (17) holds. Then, for all policies π, class indices $i \in [M']$ and $j \in [i, M']$, $\alpha > 0$, and epochs $m \in [m_i^*]$, we have:*

$$|R_{\hat{f}_{m+1,j}}(\pi) - R(\pi)| \leq \left(\frac{1}{\alpha} + \frac{\alpha}{16}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}} Reg_{\hat{f}_m}(\pi).$$

Proof. For any policy π, class indices $i \in [M']$ and $j \in [i, M']$, $\alpha > 0$, and epochs $m \in [m_i^*]$, note that:

$$\begin{aligned}&|R_{\hat{f}_{m+1,j}}(q) - R(q)| \\ =&|\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, a\sim q}[\hat{f}_{m+1,j}(x,a) - f^*(x,a)]|\end{aligned}$$

[16] $D(p_m)$ depends on $\Gamma_{\tau_{m-1}}$ because p_m is constructed using the data in $\Gamma_{\tau_{m-1}}$.

$$
\begin{aligned}
&\overset{(i)}{=}\left|\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\Big[\frac{q(a|x)}{p_m(a|x)}\big(\hat{f}_{m+1,j}(x,a)-f^*(x,a)\big)\Big]\right| \\
&\overset{(ii)}{\leq}\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\Big[\frac{q(a|x)}{p_m(a|x)}\big|\hat{f}_{m+1,j}(x,a)-f^*(x,a)\big|\Big] \\
&=\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\Big[\sqrt{\Big(\frac{q(a|x)}{p_m(a|x)}\Big)^2\big|\hat{f}_{m+1,j}(x,a)-f^*(x,a)\big|^2}\Big] \\
&\overset{(iii)}{\leq}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\Big[\Big(\frac{q(a|x)}{p_m(a|x)}\Big)^2\Big]}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\big[(\hat{f}_{m+1,j}(x,a)-f^*(x,a))^2\big]} \\
&\overset{(iv)}{=}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim q}\Big[\frac{q(a|x)}{p_m(a|x)}\Big]}\sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m}\big[(\hat{f}_{m+1,j}(x,a)-f^*(x,a))^2\big]} \\
&\overset{(v)}{\leq}\sqrt{V(p_m,q)}\sqrt{2\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2},\frac{\delta}{6TM'^2}\Big)}=\frac{\sqrt{V(p_m,q)}\sqrt{K}}{2\gamma_{m+1,i}} \\
&\overset{(vi)}{\leq}\frac{V(p_m,q)}{\alpha\gamma_{m+1,i}}+\frac{\alpha K}{16\gamma_{m+1,i}} \\
&\overset{(vii)}{\leq}\frac{K+\gamma_m\mathrm{Reg}_{\hat{f}_m}(q)}{\alpha\gamma_{m+1,i}}+\frac{\alpha K}{16\gamma_{m+1,i}}=\Big(\frac{1}{\alpha}+\frac{\alpha}{16}\Big)\frac{K}{\gamma_{m+1,i}}+\frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(q),
\end{aligned}
$$

where (i) and (iv) follow from change of measure arguments, (ii) follows from Jenson's inequality, (iii) follows from Cauchy-Schwartz inequality, (v) follows from $\mathcal{W}_1$, (vi) follows from AM-GM inequality, and (vii) follows from Lemma 2. □

The accuracy of the direct method for policy evaluation only depends on the prediction error of the underlying estimator. We therefore note that when the underlying estimator is constructed by a model selection oracle for estimation, the prediction error will decrease more rapidly in terms of sample size for small datasets. This allows us to accordingly increase the corresponding exploitation parameters more rapidly for earlier rounds.

C POLICY-BASED MISSPECIFICATION TEST

In this section, we establish the foundation for the main misspecification test. By the definition of $\hat{m}_i$, none of the tests corresponding to class i fail until this epoch.

Proof outline: Appendix C.1 provides a high-probability event for policy evaluation that holds under Oracle Assumptions 2 and 3. Moving forward, all our analysis relies on the high-probability events defined so far ($\mathcal{W}_1, \mathcal{W}_2$). Appendix C.2 provides refined policy evaluation guarantees. Appendix C.3 develops the policy-based misspecification test (MTOracle), and provides validated guarantees for the direct method estimates.

C.1 High Probability Events For Explicit Policy Evaluation

In this section, we define an event $\mathcal{W}_2$ that holds with high-probability under Oracle Assumption 2 and Oracle Assumption 3. At a high-level, $\mathcal{W}_2$ defines the event where the evaluation guarantees of consistent (e.g. IPS/DR) and direct method policy estimates hold.

$$
\begin{aligned}
\mathcal{W}_2 := \Bigg\{&\forall m,\ \forall i,j\in[M'],\ \forall\,\pi\in\tilde{\Pi}_i\cup\{\pi_{\hat{f}_{m+1}},p_{m+1,1},\ldots,p_{m+1,M'}\}, \\
&|\hat{R}_{m+1}(\pi)-R(\pi)|\leq\sqrt{V(p_m,\pi)\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2},\frac{\delta}{6TM'^2}\Big)}+2\gamma_m\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2},\frac{\delta}{6TM'^2}\Big), \\
&|\hat{R}_{m+1,\hat{f}_{m+1}}(\pi)-R_{\hat{f}_{m+1}}(\pi)|\leq\sqrt{\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2},\frac{\delta}{6TM'^2}\Big)}, \\
&|\hat{R}_{m+1,\hat{f}_{m+1,j}}(\pi)-R_{\hat{f}_{m+1,j}}(\pi)|\leq\sqrt{\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2},\frac{\delta}{6TM'^2}\Big)}\Bigg\}.
\end{aligned}
\tag{19}
$$

In Lemma 5, we use standard union bound arguments to show that the event $\mathcal{W}_2$ holds with high-probability.

Lemma 5. *Suppose Oracle Assumption 2 and Oracle Assumption 3 hold. The event $\mathcal{W}_2$ holds with probability at least $1-\delta/2$.*

Proof. Consider any epoch m. Note that, conditional on $\Gamma_{\tau_{m-1}}$ the number of samples in epoch m are fixed and these samples are i.i.d. from the distribution $D(p_m)$. Consider any pair of model indices i,j. Hence with probability $1-3\frac{\delta}{6TM'^2}$, from Oracle Assumption 2 and Oracle Assumption 3, for all policies $\pi \in \tilde{\Pi}_i \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$ we have:[17]

$$
\begin{aligned}
&|\hat{R}_{m+1}(\pi) - R(\pi)| \leq \sqrt{V(p_m,\pi)\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big)} + 2\gamma_m \xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big),\\
&|\hat{R}_{m+1,\hat{f}_{m+1}}(\pi) - R_{\hat{f}_{m+1}}(\pi)| \leq \sqrt{\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big)},\\
&|\hat{R}_{m+1,\hat{f}_{m+1,j}}(\pi) - R_{\hat{f}_{m+1,j}}(\pi)| \leq \sqrt{\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big)}.
\end{aligned}
$$

Hence, $\mathcal{W}_2$ holds with probability at least:

$$
1 - \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{m=1}^{m(T)} \frac{3\delta}{6TM'^2} \geq 1-\delta/2.
$$

□

C.2 Policy Evaluation

In this section, we bound the error of $\hat{R}_{m+1}(\pi)$, the estimate of a policy value obtained via EvalOracle.

Lemma 6. *Suppose the event $\mathcal{W}_2$ defined in (19) holds. Then, for all class indices $i \in [M']$, policies $\pi \in \tilde{\Pi}_i \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, $\alpha > 0$, and epochs $m \geq 1$, we have:*

$$
|\hat{R}_{m+1}(\pi) - R(\pi)| \leq \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,i}}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}} Reg_{\hat{f}_m}(\pi).
$$

Proof. For any class index i, policy $\pi \in \tilde{\Pi}_i \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, and epoch $m \geq 1$, we have:

$$
\begin{aligned}
&|\hat{R}_{m+1}(\pi) - R(\pi)|\\
&\overset{(i)}{\leq} \sqrt{V(p_m,\pi)\xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big)} + 2\gamma_m \xi_i\Big(\frac{\tau_m-\tau_{m-1}}{2}, \frac{\delta}{6TM'^2}\Big)\\
&\overset{(ii)}{\leq} \frac{\sqrt{V(p_m,\pi)}\sqrt{K}}{2\gamma_{m+1,i}} + \frac{2\gamma_m}{\gamma_{m+1,i}}\frac{K}{\gamma_{m+1,i}}\\
&\overset{(iii)}{\leq} \frac{V(p_m,\pi)}{\alpha\gamma_{m+1,i}} + \frac{\alpha K}{16\gamma_{m+1,i}} + \frac{2\gamma_m}{\gamma_{m+1,i}}\frac{K}{\gamma_{m+1,i}}\\
&\overset{(iv)}{\leq} \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,i}}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi),
\end{aligned}
$$

where (i) follows from $\mathcal{W}_2$, (ii) follows from the definition of $\gamma_{m+1,i}$, (iii) follows from the AM-GM inequality for any $\alpha > 0$, and (iv) follows from Lemma 2. □

C.3 Validating Direct Method Estimates

In this section, we design the main policy-based misspecification test and provide the implied guarantees when some conditions of the test hold. In Lemma 7 we develop the empirical test that must hold through epoch m_i^*. The implications of this test are captured in Lemma 9, which provides guarantees through $\hat{m}_i$ (by definition, the test corresponding to class i is satisfied until $\hat{m}_i$).

[17]Note that $K \leq \gamma_m$, hence $p_m(\cdot|\cdot) \geq 1/(K+\gamma_m) \geq 1/(2\gamma_m)$.

Lemma 7. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [m_i^*]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$. Then for any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:*

$$\begin{aligned} &|\hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi)| \\ &\leq \left(\frac{1+\theta_{i,j}}{\alpha} + \frac{(1+\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + (1+\theta_{i,j})^2/\alpha)\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j}\right)\frac{K}{\gamma_{m+1,i}} \\ &\quad + \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}\widehat{Reg}_{m+1,\hat{f}_m}(\pi). \end{aligned} \tag{20}$$

Proof. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [m_i^*]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. For any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:

$$\begin{aligned} &|\hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi)| \\ &\overset{(i)}{\leq} |\hat{R}_{m+1}(\pi) - R(\pi)| + |R(\pi) - R_f(\pi)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)| \\ &\overset{(ii)}{\leq} \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,j}}\right)\frac{K}{\gamma_{m+1,j}} + \frac{\gamma_m}{\alpha\gamma_{m+1,j}}\mathrm{Reg}_{\hat{f}_m}(\pi) \\ &\quad + \left(\frac{1}{\alpha} + \frac{\alpha}{16}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi) + \frac{K}{\gamma_{m+1,j}} \\ &\overset{(iii)}{\leq} \left(\frac{1+\theta_{i,j}}{\alpha} + \frac{(1+\theta_{i,j})\alpha}{16} + \frac{2\theta_{i,j}^2\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j}\right)\frac{K}{\gamma_{m+1,i}} + \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi), \end{aligned}$$

where (i) is an application of triangle inequality, and (ii) follows from Lemma 4, Lemma 6, and events $\mathcal{W}_1$ and $\mathcal{W}_2$. Then (iii) follows from applying the definition of parameter $\theta_{i,j}$. Lemma 7 now follows from noting that:

$$\begin{aligned} &\frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi) \\ &\overset{(i)}{\leq} \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}\left(\widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) + |R_{\hat{f}_m}(\pi_{\hat{f}_m}) - \hat{R}_{m+1,\hat{f}_m}(\pi_{\hat{f}_m})| + |R_{\hat{f}_m}(\pi) - \hat{R}_{m+1,\hat{f}_m}(\pi)|\right) \\ &\overset{(ii)}{\leq} \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}\left(\widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) + (1+\theta_{i,j})\frac{K}{\gamma_{m+1,i}}\right), \end{aligned}$$

where (i) is an application of triangle inequality, and (ii) follows from $\mathcal{W}_2$.

□

Lemma 8. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [\hat{m}_i]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Then for any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:*

$$\begin{aligned} &|R_f(\pi) - R(\pi)| - |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)| \\ &\leq \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,j}} + 1\right)\frac{K}{\gamma_{m+1,j}} + \frac{\gamma_m}{\alpha\gamma_{m+1,j}}Reg_{\hat{f}_m}(\pi). \end{aligned}$$

Proof. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [\hat{m}_i]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Then for any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:

$$\begin{aligned} &|R_f(\pi) - R(\pi)| - |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)| \\ &\overset{(i)}{\leq} |\hat{R}_{m+1}(\pi) - R(\pi)| + |R_f(\pi) - \hat{R}_{m+1,f}(\pi)| \\ &\overset{(ii)}{\leq} \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,j}} + 1\right)\frac{K}{\gamma_{m+1,j}} + \frac{\gamma_m}{\alpha\gamma_{m+1,j}}\mathrm{Reg}_{\hat{f}_m}(\pi), \end{aligned}$$

where (i) is an application of triangle inequality, and (ii) follows from Lemma 6 and $\mathcal{W}_2$. □

Lemma 9. *Suppose $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [\hat{m}_i]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$. Then for any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:*

$$|R_f(\pi) - R(\pi)| \leq \left(\frac{1 + 2\theta_{i,j}}{\alpha} + \frac{(1 + 2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1 + \theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j} \right) \frac{K}{\gamma_{m+1,i}} + \frac{(1 + 2\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} Reg_{\hat{f}_m}(\pi).$$

Proof. Consider any pair of class indices $i, j \in [M']$ such that $j \geq i$, any epoch $m \in [\hat{m}_i]$, $\alpha > 0$, and model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Then for any policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have from Lemma 8:

$$\begin{aligned} &|R_f(\pi) - R(\pi)| - |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)| \\ &\leq \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,j}} + 1 \right) \frac{K}{\gamma_{m+1,j}} + \frac{\gamma_m}{\alpha\gamma_{m+1,j}} \mathrm{Reg}_{\hat{f}_m}(\pi). \end{aligned}$$

From Lemma 7, we know that for any class index $i \in [M]$, epoch $m \in [m_i^*]$, model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, and policy $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, we have:

$$\begin{aligned} &|\hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi)| \\ \leq &\left(\frac{1 + \theta_{i,j}}{\alpha} + \frac{(1 + \theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + (1 + \theta_{i,j})^2/\alpha)\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j} \right) \frac{K}{\gamma_{m+1,i}} + \frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi). \end{aligned}$$

Combining the above results, we have:

$$\begin{aligned} &|R_f(\pi) - R(\pi)| \\ &\leq \left(\frac{1 + \theta_{i,j}}{\alpha} + \frac{(1 + \theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + (1 + \theta_{i,j})^2/\alpha)\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j} \right) \frac{K}{\gamma_{m+1,i}} + \frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) \\ &\quad + \left(\frac{1}{\alpha} + \frac{\alpha}{16} + \frac{2\gamma_m}{\gamma_{m+1,j}} + 1 \right) \frac{K}{\gamma_{m+1,j}} + \frac{\gamma_m}{\alpha\gamma_{m+1,j}} \mathrm{Reg}_{\hat{f}_m}(\pi) \\ &\overset{(i)}{\leq} \left(\frac{1 + 2\theta_{i,j}}{\alpha} + \frac{(1 + 2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + (1 + \theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j} \right) \frac{K}{\gamma_{m+1,i}} \\ &\quad + \frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) + \frac{\theta_{i,j}\gamma_m}{\alpha\gamma_{m+1,i}} \mathrm{Reg}_{\hat{f}_m}(\pi), \end{aligned}$$

where (i) follows from plugging in $\frac{1}{\gamma_{m+1,j}} \leq \frac{\theta_{i,j}}{\gamma_{m+1,i}}$. We can then combine the last two terms using the same approach used in the proof of Lemma 7:

$$\begin{aligned} &\frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) \\ &\quad \overset{(i)}{\leq} \frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \left(\mathrm{Reg}_{\hat{f}_m}(\pi) + |R_{\hat{f}_m}(\pi_{\hat{f}_m}) - \hat{R}_{m+1,\hat{f}_m}(\pi_{\hat{f}_m})| + |R_{\hat{f}_m}(\pi) - \hat{R}_{m+1,\hat{f}_m}(\pi)| \right) \\ &\quad \overset{(ii)}{\leq} \frac{(1 + \theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \left(\mathrm{Reg}_{\hat{f}_m}(\pi) + (1 + \theta_{i,j}) \frac{K}{\gamma_{m+1,i}} \right). \end{aligned}$$

Here, (i) is an application of triangle inequality, and (ii) follows from $\mathcal{W}_2$. Applying this to our expression above gives the final form for Lemma 9:

$$\begin{aligned} &|R_f(\pi) - R(\pi)| \\ &\leq \left(\frac{1 + 2\theta_{i,j}}{\alpha} + \frac{(1 + 2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1 + \theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j} \right) \frac{K}{\gamma_{m+1,i}} \\ &\quad + \frac{(1 + 2\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \mathrm{Reg}_{\hat{f}_m}(\pi). \end{aligned}$$

∎

D VERIFYING REWARD MODEL AGREEMENT

In this section, we design the remainder of the misspecification test. In particular, we ensure agreement in reward models estimated across classes and epochs. We first prove an inductive result that allows us to relate the true regret to the regret according to estimated models. Then, in Appendix D.1, we develop an empirical test for reward model agreement and prove verified guarantees that hold as long as the test doesn't fail. By the definition of $\hat{m}_i$, none of the tests corresponding to class i fail until this epoch.

Proof outline: Lemma 10 proves our key inductive step that holds within safe epochs. Lemmas 11 and 12 use this inductive step to relate the true regret to the regret according to estimated models (within safe epochs). Lemma 13 provides the expected reward model agreement across classes and describes the corresponding test. Lemma 14 provides implied guarantees as long this test holds. Lemmas 15 and 16 provide the expected reward model agreement across epochs and describe the corresponding test. Lemma 17 provides implied guarantees as long this test holds.

Lemma 10. *Suppose the event $\mathcal{W}_1$ defined in* (17) *holds. Consider any class index $i \in [M']$ and consider any epoch $m \in [m_i^*]$. Suppose there exists a constant ($\eta > 0$) such that for all policies π, we have:*

$$
\begin{aligned}
\mathit{Reg}(\pi) &\leq 2\mathit{Reg}_{\hat{f}_m}(\pi) + \frac{\eta K}{\gamma_{m,i}} \\
\mathit{Reg}_{\hat{f}_m}(\pi) &\leq 2\mathit{Reg}(\pi) + \frac{\eta K}{\gamma_{m,i}}.
\end{aligned}
$$

We then have that:

$$
\begin{aligned}
\mathit{Reg}(\pi) &\leq 2\mathit{Reg}_{f}(\pi) + \frac{\eta' K}{\gamma_{m+1,i}}, \quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\} \\
\mathit{Reg}_{\hat{f}_{m+1}}(\pi) &\leq 2\mathit{Reg}(\pi) + \frac{\eta' K}{\gamma_{m+1,i}}.
\end{aligned}
$$

where $\eta' = 2\max\left(\frac{\gamma_m}{\gamma_{m,i}}, \sqrt{1+\frac{\gamma_m}{\gamma_{m,i}}\eta}\right)$.

Proof. Let α be any positive constant, and let $\alpha' = \gamma_m/\gamma_{m,i}$. Note that for any $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, we have:

$$
\begin{aligned}
\mathrm{Reg}(\pi) - \mathrm{Reg}_f(\pi) =& \Big(R(\pi^*) - R(\pi)\Big) - \Big(R_f(\pi_f) - R_f(\pi)\Big) \\
\overset{(i)}{\leq}& \Big(R(\pi^*) - R(\pi)\Big) - \Big(R_f(\pi^*) - R_f(\pi)\Big) \\
\overset{(ii)}{\leq}& |R(\pi^*) - R_f(\pi^*)| + |R(\pi) - R_f(\pi))| \\
\overset{(iii)}{\leq}& \left(\frac{2}{\alpha} + \frac{\alpha}{8}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\Big(\mathrm{Reg}_{\hat{f}_m}(\pi) + \mathrm{Reg}_{\hat{f}_m}(\pi^*)\Big),
\end{aligned}
\tag{21}
$$

where (i) follows from the definition of π_f for $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, (ii) follows from the triangle inequality, and (iii) follows from Lemma 4. Now note that:

$$
\frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi) \overset{(i)}{\leq} \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\Big(2\mathrm{Reg}(\pi) + \frac{\eta K}{\gamma_{m,i}}\Big) \leq \frac{2\alpha'}{\alpha}\mathrm{Reg}(\pi) + \frac{\alpha'\eta K}{\alpha\gamma_{m+1,i}}, \tag{22}
$$

where (i) follows from the conditions stated in Lemma 10. Similarly note that:

$$
\frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi^*) \overset{(i)}{\leq} \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\Big(2\mathrm{Reg}(\pi^*) + \frac{\eta K}{\gamma_{m,i}}\Big) \overset{(ii)}{=} \frac{\alpha'\eta K}{\alpha\gamma_{m+1,i}}, \tag{23}
$$

where (i) follows from the conditions stated in Lemma 10, and (ii) follows from the fact that $\mathrm{Reg}(\pi^*) = 0$. Now from combining (21), (22), and (23), we get:

$$\begin{aligned}
\mathrm{Reg}(\pi) - \mathrm{Reg}_f(\pi) &\le \left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}} + \frac{2\alpha'}{\alpha}\mathrm{Reg}(\pi) \\
\frac{\alpha - 2\alpha'}{\alpha}\mathrm{Reg}(\pi) &\le \mathrm{Reg}_f(\pi) + \left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}} \\
\mathrm{Reg}(\pi) &\le \frac{\alpha}{\alpha - 2\alpha'}\mathrm{Reg}_f(\pi) + \frac{\alpha}{\alpha-2\alpha'}\left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}}.
\end{aligned} \tag{24}$$

Similar to (21), we get:

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) - \mathrm{Reg}(\pi) &= \Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi)\Big) - \Big(R(\pi^*) - R(\pi)\Big) \\
&\overset{(i)}{\le} \Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi)\Big) - \Big(R(\pi_{\hat{f}_{m+1}}) - R(\pi)\Big) \\
&\overset{(ii)}{\le} |R(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})| + |R(\pi) - R_{\hat{f}_{m+1}}(\pi))| \\
&\overset{(iii)}{\le} \left(\frac{2}{\alpha} + \frac{\alpha}{8}\right)\frac{K}{\gamma_{m+1,i}} + \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\Big(\mathrm{Reg}_{\hat{f}_m}(\pi) + \mathrm{Reg}_{\hat{f}_m}(\pi_{\hat{f}_{m+1}})\Big),
\end{aligned} \tag{25}$$

where (i) follows from the definition of π^*, (ii) follows from the triangle inequality, and (iii) follows from Lemma 4. Similar to (23), we get:

$$\begin{aligned}
\frac{\gamma_m}{\alpha\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi_{\hat{f}_{m+1}}) &\overset{(i)}{\le} \frac{\gamma_m}{\alpha\gamma_{m+1,i}}\left(2\mathrm{Reg}(\pi_{\hat{f}_{m+1}}) + \frac{\eta K}{\gamma_{m,i}}\right) \\
&\le \frac{2\alpha'}{\alpha}\mathrm{Reg}(\pi_{\hat{f}_{m+1}}) + \frac{\alpha'\eta K}{\alpha\gamma_{m+1,i}} \\
&\overset{(ii)}{\le} \frac{2\alpha'}{\alpha}\left(\frac{\alpha}{\alpha-2\alpha'}\left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}}\right) + \frac{\alpha'\eta K}{\alpha\gamma_{m+1,i}},
\end{aligned} \tag{26}$$

where (i) follows from the conditions stated in Lemma 10, and (ii) follows from (24). Combining (22), (24), (25), and (26), we get:

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) - \mathrm{Reg}(\pi) &\le \left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}}\left(1 + \frac{2\alpha'}{\alpha - 2\alpha'}\right) + \frac{2\alpha'\mathrm{Reg}(\pi)}{\alpha}. \\
\mathrm{Reg}_{\hat{f}_{m+1}}(\pi) &\le \frac{\alpha+2\alpha'}{\alpha}\mathrm{Reg}(\pi) + \frac{\alpha}{\alpha-2\alpha'}\left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right)\frac{K}{\gamma_{m+1,i}}.
\end{aligned} \tag{27}$$

If $\alpha \ge 4\alpha'$, we have that:

$$\frac{\alpha+2\alpha'}{\alpha} \le 2, \text{ and } \frac{\alpha}{\alpha-2\alpha'} \le 2. \tag{28}$$

Further, if it is also true that $\alpha \ge 4\sqrt{1+\alpha'\eta}$, we get:

$$\begin{aligned}
\frac{\alpha}{\alpha-2\alpha'}\left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right) &\le 2\left(\frac{2}{\alpha} + \frac{\alpha}{8} + \frac{2\alpha'\eta}{\alpha}\right) \\
&\le 2\left(\frac{\alpha}{4}\right).
\end{aligned} \tag{29}$$

We therefore choose $\alpha = 4\max(\alpha', \sqrt{1+\alpha'\eta})$. We finally get the required result by combining (24), (27), (28), and (29). $\square$

Lemma 11. *Suppose the event $\mathcal{W}_1$ holds. Consider any class index $i \in [M']$. For all policies π and epochs $m \le m_i^* + 1$ we have:*

$$\mathit{Reg}(\pi) \le 2\mathit{Reg}_f(\pi) + \frac{\eta_{i,m}K}{\gamma_{m,i}}, \quad \forall f \in \{\hat{f}_m, \hat{f}_{m,i}\},$$

$$Reg_{\hat{f}_m}(\pi) \leq 2Reg(\pi) + \frac{\eta_{i,m} K}{\gamma_{m,i}},$$

where $\eta_{i,m} = 2 + 4(\gamma_{m-1,1}/\gamma_{m-1,i})$.

Proof. We will prove this by induction. The base case follows from the fact that for all policies π, we have:

$$\text{Reg}(\pi) \leq 1 \leq \eta_1 K/\gamma_{1,i}$$
$$\text{Reg}_{\hat{f}_1}(\pi) \leq 1 \leq \eta_1 K/\gamma_{1,i}.$$

For the inductive step, fix some $m \leq m_i^*$. Assume for all policies π, we have:

$$\text{Reg}(\pi) \leq 2\text{Reg}_f(\pi) + \frac{\eta_{i,m} K}{\gamma_{m,i}}, \quad \forall f \in \{\hat{f}_m, \hat{f}_{m,i}\},$$
$$\text{Reg}_{\hat{f}_m}(\pi) \leq 2\text{Reg}(\pi) + \frac{\eta_{i,m} K}{\gamma_{m,i}}.$$

Therefore, from Lemma 10 we have:

$$\text{Reg}(\pi) \leq 2\text{Reg}_f(\pi) + \frac{\eta'_{i,m+1} K}{\gamma_{m+1,i}}, \quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\},$$
$$\text{Reg}_{\hat{f}_{m+1}}(\pi) \leq 2\text{Reg}(\pi) + \frac{\eta'_{i,m+1} K}{\gamma_{m+1,i}},$$

where $\eta'_{i,m+1} = 2\max\left(\frac{\gamma_m}{\gamma_{m,i}}, \sqrt{1 + \frac{\gamma_m}{\gamma_{m,i}}\eta_{i,m}}\right)$. Then we have:

$$\begin{aligned}
\eta'_{i,m+1} &= 2\max\left(\frac{\gamma_m}{\gamma_{m,i}}, \sqrt{1 + \frac{\gamma_m}{\gamma_{m,i}}\eta_{i,m}}\right) \\
&\leq 2\max\left(\frac{\gamma_{m,1}}{\gamma_{m,i}}, \sqrt{1 + \frac{\gamma_{m,1}}{\gamma_{m,i}}\eta_{i,m}}\right) \\
&\leq 2\max\left(\frac{\gamma_{m,1}}{\gamma_{m,i}}, \sqrt{1 + 2^2\frac{\gamma_{m,1}^2}{\gamma_{m,i}^2} + 2\frac{\gamma_{m,1}}{\gamma_{m,i}}}\right) \\
&\leq \max\left(2\frac{\gamma_{m,1}}{\gamma_{m,i}}, \eta_{i,m+1}\right) \\
&= \eta_{i,m+1}.
\end{aligned}$$

This completes the inductive argument. □

Lemma 12. *Suppose the event $\mathcal{W}_1$ holds. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. For all policies π and epochs $m \in [\hat{m}_{h-1} + 1, m_i^* + 1]$, we have:*

$$Reg(\pi) \leq 2Reg_f(\pi) + 8\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}\frac{K}{\gamma_{m,i}}, \quad \forall f \in \{\hat{f}_m, \hat{f}_{m,i}\},$$
$$Reg_{\hat{f}_m}(\pi) \leq 2Reg(\pi) + 8\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}\frac{K}{\gamma_{m,i}}.$$

Proof. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. We will prove the required bound by induction. The bound for the base case $m = \hat{m}_{h-1} + 1$ follows from Lemma 11. Suppose the bound in Lemma 12 holds for class indices i, h and for some epoch $m \in [\hat{m}_{h-1} + 1, m_i^*]$. From Lemma 10, we have:

$$\text{Reg}(\pi) \leq 2\text{Reg}_f(\pi) + \frac{\eta' K}{\gamma_{m+1,i}}, \quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\},$$
$$\text{Reg}_{\hat{f}_{m+1}}(\pi) \leq 2\text{Reg}(\pi) + \frac{\eta' K}{\gamma_{m+1,i}},$$

along with:

$$
\begin{aligned}
\eta' &= 2\max\left(\frac{\gamma_m}{\gamma_{m,i}}, \sqrt{1+\frac{\gamma_m}{\gamma_{m,i}} 8\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}}\right)\\
&\overset{(i)}{\leq} 2\max\left(\frac{\gamma_{m,h}}{\gamma_{m,i}}, \sqrt{1+\frac{\gamma_{m,h}}{\gamma_{m,i}} 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}}\right)\\
&\overset{(ii)}{\leq} 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}},
\end{aligned}
$$

where (i) follows from the fact that $\gamma_m \leq \gamma_{m,h}$ (since $m \geq \hat{m}_{h-1}+1$), and $\gamma_{m-1,h}/\gamma_{m-1,i} \leq \gamma_{m,h}/\gamma_{m,i}$ for $h \leq i$. Then (ii) follows from the fact that $1+z \leq 2z$ for $z \geq 1$. Hence we have shown the bound in Lemma 12 holds for class indices i, h and epoch $m+1$. This completes our inductive argument. □

D.1 Verifying Reward Model Agreement Across Classes

In Lemma 13, we develop a bound on $\widehat{\text{Reg}}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})$, which indicates whether the policy induced by the model predicted for class i is considered to be a good policy by the model we have estimated. When this bound is exceeded, it suggests that we should use the exploitation parameter corresponding to larger classes. The implications of this test are captured in Lemma 14, which provides a bound on $\text{Reg}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m}+1,i})$ through $\hat{m}_i$, provided the test is satisfied.

Lemma 13. *Suppose the events $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Consider $h \leq i$ and $m \in [\hat{m}_{h-1}, m_i^*]$, we then have:*

$$
\widehat{Reg}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \leq 26\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}}. \tag{30}
$$

Proof. From Lemma 12, we have the following for any policy π:

$$
\begin{aligned}
\text{Reg}(\pi) &\leq 2\text{Reg}_f(\pi) + 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}}, \quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\},\\
\text{Reg}_{\hat{f}_{m+1}}(\pi) &\leq 2\text{Reg}(\pi) + 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}}.
\end{aligned}
$$

Combining the above and plugging in $\pi = \pi_{\hat{f}_{m+1,i}}$, we have:

$$
\begin{aligned}
\text{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) &\leq 2\text{Reg}(\pi_{\hat{f}_{m+1,i}}) + 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}}\\
&\leq 4\text{Reg}_{\hat{f}_{m+1,i}}(\pi_{\hat{f}_{m+1,i}}) + 24\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}} \qquad (31)\\
&\leq 24\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\frac{K}{\gamma_{m+1,i}},
\end{aligned}
$$

where the last inequality follows from $\text{Reg}_{\hat{f}_{m+1,i}}(\pi_{\hat{f}_{m+1,i}}) = 0$. We then have:

$$
\begin{aligned}
&\widehat{\text{Reg}}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})\\
&= \hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - \hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})\\
&= \left(\hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})\right)\\
&\quad + \left(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) - \hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})\right) + \text{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})
\end{aligned}
$$

$$\overset{(i)}{\leq} \frac{2K}{\gamma_{m+1,i}} + \mathrm{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})$$
$$\overset{(ii)}{\leq} 26\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}} \frac{K}{\gamma_{m+1,i}},$$

where (i) follows from $\mathcal{W}_2$ and (ii) follows from (31). □

Lemma 14. *Suppose the event $\mathcal{W}_2$ holds. Consider $h \leq i$ and $m \in [\hat{m}_{h-1}, \hat{m}_i]$, we then have:*

$$Reg_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \leq 28\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}} \frac{K}{\gamma_{m+1,i}}.$$

Proof. Since the test in Lemma 13 holds up to $\hat{m}_i$, we have the following guarantee:

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) &= R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \\
&= \left(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - \hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})\right) \\
&\quad + \left(\hat{R}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}})\right) + \widehat{\mathrm{Reg}}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \\
&\leq \frac{2K}{\gamma_{m+1,i}} + \widehat{\mathrm{Reg}}_{m+1,\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \\
&\leq 28\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}} \frac{K}{\gamma_{m+1,i}}.
\end{aligned} \tag{32}$$

□

D.2 Verifying Reward Model Agreement Across Epochs

The goal of this section is to verify that potential new exploration policies had sufficiently low regret according to models in previous epochs. This helps ensure that these new exploration policies were well-explored in previous epochs and we can rely on our estimates for these policies. Lemma 15 provides the expected reward model agreement across epochs by bounding $\mathrm{Reg}_{\hat{f}_m}(\pi)$ in terms of regret according to $\hat{f}_{m+1}$ and $\hat{f}_{m+1,i}$. Lemma 16 describes the corresponding test. Lemma 17 provides implied guarantees as long this test holds.

Lemma 15. *Suppose the event $\mathcal{W}_1$ holds. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. For all policies π and epochs $m \in [\hat{m}_{h-1}, m_i^*]$, we have:*

$$Reg_{\hat{f}_m}(\pi) \leq 4Reg_f(\pi) + 24\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}} \frac{K}{\gamma_{m,i}}, \quad \forall f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}.$$

Proof. Consider $m \in [\hat{m}_{h-1}, m_i^*]$, policy π, and $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$.

$$\begin{aligned}
\mathrm{Reg}_{\hat{f}_m}(\pi) &\overset{(i)}{\leq} 2\mathrm{Reg}(\pi) + 8\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}} \frac{K}{\gamma_{m,i}} \\
&\overset{(ii)}{\leq} 4\mathrm{Reg}_f(\pi) + 24\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}} \frac{K}{\gamma_{m,i}}.
\end{aligned} \tag{33}$$

To show (i) we consider two cases. For the case $m = \hat{m}_{h-1}$, (i) follows from Lemma 11 and the fact that $\frac{\gamma_{m-1,1}}{\gamma_{m-1,i}} \leq \frac{\gamma_{m,1}}{\gamma_{m,i}}$. For the case $m > \hat{m}_{h-1}$, (i) follows from Lemma 12 and the fact that $\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}} \leq \frac{\gamma_{m,h}}{\gamma_{m,i}}$. Then (ii) follows from Lemma 12 and the fact that $\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}} \geq 1$, $\frac{K}{\gamma_{m,i}} \geq \frac{K}{\gamma_{m+1,i}}$. □

Lemma 16. *Suppose the event $\mathcal{W}_1$ holds. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. For all policies $\Pi_{0,m+1,i} = \pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, epochs $m \in [\hat{m}_{h-1}, m_i^*]$, and models $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, we have:*

$$\widehat{Reg}_{m+1,\hat{f}_m}(\pi) \leq 4\widehat{Reg}_{m+1,f}(\pi) + 34\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}} \frac{K}{\gamma_{m,i}}. \tag{34}$$

Proof. Consider any policy $\pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, and epoch $m \in [\hat{m}_{h-1}, m_i^*]$.

$$
\begin{aligned}
&\widehat{\text{Reg}}_{m+1,\hat{f}_m}(\pi) - 4\widehat{\text{Reg}}_{m+1,f}(\pi) \\
&= (\hat{R}_{m+1,\hat{f}_m}(\pi_{\hat{f}_m}) - \hat{R}_{m+1,\hat{f}_m}(\pi)) - 4(\hat{R}_{m+1,f}(\pi_f) - \hat{R}_{m+1,f}(\pi)) \\
&= \big(\hat{R}_{m+1,\hat{f}_m}(\pi_{\hat{f}_m}) - R_{\hat{f}_m}(\pi_{\hat{f}_m})\big) + \big(R_{\hat{f}_m}(\pi) - \hat{R}_{m+1,\hat{f}_m}(\pi)\big) + \text{Reg}_{\hat{f}_m}(\pi) \\
&\quad + 4\big(R_f(\pi_f) - \hat{R}_{m+1,f}(\pi_f)\big) + 4\big(\hat{R}_{m+1,f}(\pi) - R_f(\pi)\big) - 4\text{Reg}_f(\pi) \\
&\overset{(i)}{\leq} \frac{10K}{\gamma_{m+1,i}} + \text{Reg}_{\hat{f}_m}(\pi) - 4\text{Reg}_f(\pi) \\
&\overset{(ii)}{\leq} 34\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}}\frac{K}{\gamma_{m,i}},
\end{aligned}
$$

where (i) follows from $\mathcal{W}_2$ and (ii) follows from Lemma 15. ☐

Lemma 17. *Suppose the event $\mathcal{W}_1$ holds. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. For all policies $\pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, epochs $m \in [\hat{m}_{h-1}, \hat{m}_i]$, and models $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, we have:*

$$
Reg_{\hat{f}_m}(\pi) \leq 4Reg_f(\pi) + 44\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}}\frac{K}{\gamma_{m,i}}.
$$

Proof. Consider any policy $\pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}$, model $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, and epoch $m \in [\hat{m}_{h-1}, \hat{m}_i]$.

$$
\begin{aligned}
&\text{Reg}_{\hat{f}_m}(\pi) - 4\text{Reg}_f(\pi) \\
&= (R_{\hat{f}_m}(\pi_{\hat{f}_m}) - R_{\hat{f}_m}(\pi)) - 4(R_f(\pi_f) - R_f(\pi)) \\
&= \big(R_{\hat{f}_m}(\pi_{\hat{f}_m}) - \hat{R}_{m+1,\hat{f}_m}(\pi_{\hat{f}_m})\big) + \big(\hat{R}_{m+1,\hat{f}_m}(\pi) - R_{\hat{f}_m}(\pi)\big) + \widehat{\text{Reg}}_{m+1,\hat{f}_m}(\pi) \\
&\quad + 4\big(\hat{R}_{m+1,f}(\pi_f) - R_f(\pi_f)\big) + 4\big(R_f(\pi) - \hat{R}_{m+1,f}(\pi)\big) - 4\widehat{\text{Reg}}_{m+1,f}(\pi) \\
&\overset{(i)}{\leq} \frac{10K}{\gamma_{m+1,i}} + \widehat{\text{Reg}}_{m+1,\hat{f}_m}(\pi) - 4\widehat{\text{Reg}}_{m+1,f}(\pi) \\
&\overset{(ii)}{\leq} 44\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{\max(0,m-\hat{m}_{h-1}-1)}}\frac{K}{\gamma_{m,i}},
\end{aligned}
$$

where (i) follows from $\mathcal{W}_2$ and (ii) follows from Lemma 16. ☐

E INDUCTIVE ARGUMENT BASED ON TESTED GUARANTEES

In Appendices C and D, we developed several verified guarantees on our estimated reward models. In this section, we rely on these guarantees to relate the true regret (with respect to the best policies in different classes) to the regret according to estimated models. Our proof follows by induction and demonstrates the benefits of the self-correction step (holding candidate exploitation parameters fixed by not increasing epoch lengths for a few epochs) in our algorithm.

Proof outline: Lemma 18 is our main inductive step. Lemmas 19 and 20 apply this step in order to relate the true regret (with respect to the best policies in different classes) to the regret according to estimated models. Lemma 20 in particular demonstrates how holding candidate exploitation parameters fixed for some epochs helps with correcting for the effects of under-exploration on our estimated reward models.

E.1 Inductive Step Based on Tested Guarantees

Lemma 18 is our main inductive step that utilizes tested guarantees. We state and prove it in this section.

Lemma 18. *Suppose the event $\mathcal{W}_1$ defined in* (17) *holds. Consider any class indices $i, j \in [M']$ such that $j \geq i$, and consider any epoch $m \in [\hat{m}_i]$. Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$. Suppose there exist constants $(\eta, \tilde{\eta}, \tilde{\eta}' > 0)$ such that we have the following for any $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$:*

$$\begin{aligned}
&Reg_j(\pi) \leq 2Reg_{\hat{f}_m}(\pi) + \frac{\eta K}{\gamma_{m,i}}, \quad \forall \pi \in \tilde{\Pi}_j \\
&Reg_{\hat{f}_m}(\pi) \leq 2Reg_j(\pi) + \frac{\eta K}{\gamma_{m,i}}, \quad \forall \pi \in \tilde{\Pi}_j \\
&Reg_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \leq \frac{\tilde{\eta} K}{\gamma_{m+1,i}}, \\
&Reg_{\hat{f}_m}(\pi) \leq 4Reg_f(\pi) + \frac{\tilde{\eta}' K}{\gamma_{m,i}}, \quad \forall \pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \dots, p_{m+1,M'}\}.
\end{aligned} \tag{35}$$

We then have that for all polices $\pi \in \tilde{\Pi}_j \cup \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \dots, p_{m+1,M'}\}$ and models $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$:

$$\begin{aligned}
Reg_j(\pi) &\leq 2Reg_f(\pi) + \frac{\eta' K}{\gamma_{m+1,i}} \\
Reg_f(\pi) &\leq 2Reg_j(\pi) + \frac{\eta' K}{\gamma_{m+1,i}},
\end{aligned} \tag{36}$$

where $\eta' = 5\alpha'(1 + 2\theta_{i,j})^2 + 2\tilde{\eta} + 2(1 + 2\theta_{i,j})\sqrt{\alpha'(\eta + \tilde{\eta}')}$ and $\alpha' = \gamma_m/\gamma_{m,i}$.

Proof. To begin our proof, we first define a few quantities. Let $\alpha' = \gamma_m/\gamma_{m,i}$, and let $\alpha \geq 8\alpha'(1 + 2\theta_{i,j})$ be a positive constant (will be fixed later in the proof). Let f be any model in $\{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. Let $\tilde{\Pi}_{0,m+1} = \{\pi_{\hat{f}_{m+1}}, p_{m+1,1}, \dots, p_{m+1,M'}\}$. Our proof is broken into two parts.

Part 1: The first part of the proof works towards the first inequality in (36). We start with bounding the difference between $\text{Reg}_j(\pi)$ and $\text{Reg}_f(\pi)$ for all policies $\pi \in \tilde{\Pi}_j \cup \tilde{\Pi}_{0,m+1}$.

$$\begin{aligned}
\pi \in \tilde{\Pi}_j \cup \tilde{\Pi}_{0,m+1},\ &\text{Reg}_j(\pi) - \text{Reg}_f(\pi) \\
&= \Big(R(\pi_j^*) - R(\pi)\Big) - \Big(R_f(\pi_f) - R_f(\pi)\Big) \\
&\overset{(i)}{\leq} \Big(R(\pi_j^*) - R(\pi)\Big) - \Big(R_f(\pi_j^*) - R_f(\pi)\Big) \\
&\overset{(ii)}{\leq} |R(\pi_j^*) - R_f(\pi_j^*)| + |R(\pi) - R_f(\pi))| \\
&\overset{(iii)}{\leq} 2\bar{C} + \frac{(1 + 2\theta_{i,j})}{\alpha} \frac{\gamma_m}{\gamma_{m+1,i}} \Big(\text{Reg}_{\hat{f}_m}(\pi_j^*) + \text{Reg}_{\hat{f}_m}(\pi)\Big), \\
&\overset{(iv)}{\leq} 2\bar{C} + \frac{(1 + 2\theta_{i,j})\alpha'}{\alpha} \frac{\eta K}{\gamma_{m+1,i}} + \frac{(1 + 2\theta_{i,j})}{\alpha} \frac{\gamma_m}{\gamma_{m+1,i}} \text{Reg}_{\hat{f}_m}(\pi),
\end{aligned} \tag{37}$$

where (i) follows from the definition of π_f, (ii) follows from the triangle inequality, (iii) follows from Lemma 9, and (iv) follows from (35) and $\text{Reg}_j(\pi_j^*) = 0$. For brevity, in (iii), we have defined the quantity:

$$\bar{C} = \left(\frac{1 + 2\theta_{i,j}}{\alpha} + \frac{(1 + 2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1 + \theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j}\right) \frac{K}{\gamma_{m+1,i}},$$

which is the first term from the result of Lemma 9.

Part 1 (Case 1: $\pi \in \tilde{\Pi}_j$): This case only considers policies $\pi \in \tilde{\Pi}_j$, we refine (37) for such policies using (35).

Combining (37) and (35), we get:

$$
\begin{aligned}
&\mathrm{Reg}_j(\pi) - \mathrm{Reg}_f(\pi) \le 2\bar{C} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}\frac{\eta K}{\gamma_{m+1,i}} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\left(2\mathrm{Reg}_j(\pi) + \frac{\eta K}{\gamma_{m,i}}\right) \\
\implies &\frac{\alpha - 2\alpha'(1+2\theta_{i,j})}{\alpha}\mathrm{Reg}_j(\pi) \le \mathrm{Reg}_f(\pi) + 2\bar{C} + \frac{2\alpha'\eta K(1+2\theta_{i,j})}{\alpha\gamma_{m+1,i}} \\
\implies &\mathrm{Reg}_j(\pi) \le \frac{\alpha}{\alpha - 2\alpha'(1+2\theta_{i,j})}\left(\mathrm{Reg}_f(\pi) + 2\bar{C} + \frac{2\alpha'\eta K(1+2\theta_{i,j})}{\alpha\gamma_{m+1,i}}\right) \\
\implies &\mathrm{Reg}_j(\pi) \le 2\mathrm{Reg}_f(\pi) + 4\bar{C} + \frac{4\alpha'\eta K(1+2\theta_{i,j})}{\alpha\gamma_{m+1,i}},
\end{aligned} \tag{38}
$$

where the last implication follows from the fact that $\alpha \ge 4\alpha'(1+2\theta_{i,j})$ and hence $\frac{\alpha}{\alpha-2\alpha'(1+2\theta_{i,j})} \le 2$.

Part 1 (Case 2: $\pi \in \Pi_{0,m+1}$): This case only considers policies $\pi \in \Pi_{0,m+1}$, we refine (37) for such policies using (35). Combining (37) and (35), we get:

$$
\begin{aligned}
&\mathrm{Reg}_j(\pi) - \mathrm{Reg}_f(\pi) \le 2\bar{C} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}\frac{\eta K}{\gamma_{m+1,i}} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\left(4\mathrm{Reg}_f(\pi) + \frac{\tilde{\eta}' K}{\gamma_{m,i}}\right) \\
\implies &\mathrm{Reg}_j(\pi) \le \left(1 + \frac{4(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\right)\mathrm{Reg}_f(\pi) + 2\bar{C} + \frac{\alpha'(\eta+\tilde{\eta}')K(1+2\theta_{i,j})}{\alpha\gamma_{m+1,i}} \\
\implies &\mathrm{Reg}_j(\pi) \le 2\mathrm{Reg}_f(\pi) + 2\bar{C} + \frac{\alpha'(\eta+\tilde{\eta}')K(1+2\theta_{i,j})}{\alpha\gamma_{m+1,i}},
\end{aligned} \tag{39}
$$

where the last implication follows from the fact that $\alpha \ge 4\alpha'(1+2\theta_{i,j})$.

Part 2: The second part of the proof works towards the second inequality in (36). We start with bounding the difference between $\mathrm{Reg}_f(\pi)$ and $\mathrm{Reg}_j(\pi)$ for all policies $\pi \in \tilde{\Pi}_j \cup \tilde{\Pi}_{0,m+1}$.

$$
\begin{aligned}
\pi \in \tilde{\Pi}_j \cup \tilde{\Pi}_{0,m+1},\ &\mathrm{Reg}_f(\pi) - \mathrm{Reg}_j(\pi) \\
&= \Big(R_f(\pi_f) - R_f(\pi)\Big) - \Big(R(\pi_j^*) - R(\pi)\Big) \\
&\overset{(i)}{\le} \Big(R_f(\pi_f) - R_f(\pi)\Big) - \Big(R(\pi_{\hat{f}_{m+1,i}}) - R(\pi)\Big) \\
&\overset{(ii)}{\le} \mathrm{Reg}_f(\pi_{\hat{f}_{m+1,i}}) + \Big|R_f(\pi_{\hat{f}_{m+1,i}}) - R(\pi_{\hat{f}_{m+1,i}})\Big| + \Big|R(\pi) - R_f(\pi)\Big|, \\
&\overset{(iii)}{\le} \frac{\tilde{\eta}K}{\gamma_{m+1,i}} + \Big|R_f(\pi_{\hat{f}_{m+1,i}}) - R(\pi_{\hat{f}_{m+1,i}})\Big| + \Big|R(\pi) - R_f(\pi)\Big| \\
&\overset{(iv)}{\le} \frac{\tilde{\eta}K}{\gamma_{m+1,i}} + 2\bar{C} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\Big(\mathrm{Reg}_{\hat{f}_m}(\pi_{\hat{f}_{m+1,i}}) + \mathrm{Reg}_{\hat{f}_m}(\pi)\Big) \\
&\overset{(v)}{\le} \frac{\tilde{\eta}K}{\gamma_{m+1,i}} + 2\bar{C} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}\frac{\tilde{\eta}'K}{\gamma_{m+1,i}} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\mathrm{Reg}_{\hat{f}_m}(\pi),
\end{aligned} \tag{40}
$$

where (i) follows from the definition of π_j^* and $\pi_{\hat{f}_{m+1,i}} \in \tilde{\Pi}_j$, (ii) follows from triangle inequality, (iii) follows from (35), (iv) follows from Lemma 9, and (v) follows from (35) and $\mathrm{Reg}_{\hat{f}_{m+1,i}}(\pi_{\hat{f}_{m+1,i}}) = 0$.

Part 2 (Case 1 $\pi \in \tilde{\Pi}_j$): This case only considers policies $\pi \in \tilde{\Pi}_j$, we refine (40) for such policies using (35). Combining (40) and (35), we get:

$$
\begin{aligned}
&\mathrm{Reg}_f(\pi) \\
&\le \frac{\tilde{\eta}K}{\gamma_{m+1,i}} + 2\bar{C} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}\frac{\tilde{\eta}'K}{\gamma_{m+1,i}} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\left(2\mathrm{Reg}_j(\pi) + \frac{\eta K}{\gamma_{m,i}}\right) + \mathrm{Reg}_j(\pi) \\
&\le \left(1 + \frac{2(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\right)\mathrm{Reg}_j(\pi) + 2\bar{C} + \frac{K}{\gamma_{m+1,i}}\left(\tilde{\eta} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}(\eta+\tilde{\eta}')\right) \\
&\le 2\mathrm{Reg}_j(\pi) + 2\bar{C} + \frac{K}{\gamma_{m+1,i}}\left(\tilde{\eta} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}(\eta+\tilde{\eta}')\right),
\end{aligned} \tag{41}
$$

where the last inequality follows from the fact that $\alpha \geq 4\alpha'(1+2\theta_{i,j})$.

Part 2 (Case 2 $\pi \in \tilde{\Pi}_{0,m+1}$): This case only considers policies $\pi \in \tilde{\Pi}_{0,m+1}$, we refine (40) for such policies using (35). Combining (40) and (35), we get:

$$\begin{aligned}
&\mathrm{Reg}_f(\pi) - \mathrm{Reg}_j(\pi) \leq \frac{\tilde{\eta}K}{\gamma_{m+1,i}} + 2\bar{C} + \frac{(1+2\theta_{i,j})\alpha'}{\alpha}\frac{\tilde{\eta}'K}{\gamma_{m+1,i}} + \frac{(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\left(4\mathrm{Reg}_f(\pi) + \frac{\tilde{\eta}'K}{\gamma_{m,i}}\right) \\
&\implies \left(1 - \frac{4(1+2\theta_{i,j})}{\alpha}\frac{\gamma_m}{\gamma_{m+1,i}}\right)\mathrm{Reg}_f(\pi) \leq \mathrm{Reg}_j(\pi) + 2\bar{C} + \frac{K}{\gamma_{m+1,i}}\left(\tilde{\eta} + \frac{2(1+2\theta_{i,j})\alpha'\tilde{\eta}'}{\alpha}\right) \\
&\implies \mathrm{Reg}_f(\pi) \leq 2\mathrm{Reg}_j(\pi) + 4\bar{C} + \frac{K}{\gamma_{m+1,i}}\left(2\tilde{\eta} + \frac{4(1+2\theta_{i,j})\alpha'\tilde{\eta}'}{\alpha}\right),
\end{aligned} \tag{42}$$

where the last implication follows from the fact that $\alpha \geq 8\alpha'(1+2\theta_{i,j})$.

Note that, if (43) holds and if $\alpha \geq 8\alpha'(1+2\theta_{i,j})$, then (38),(39), (41) (42) imply (36) holds for all policies in $\tilde{\Pi}_j \cup \Pi_{0,m+1}$:

$$\begin{aligned}
\eta' \geq& 4\left(\frac{1+2\theta_{i,j}}{\alpha} + \frac{(1+2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1+\theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j}\right) \\
&+ 2\tilde{\eta} + \frac{4(1+2\theta_{i,j})\alpha'(\eta+\tilde{\eta}')}{\alpha}.
\end{aligned} \tag{43}$$

We now fix our choice of α in a way that ensures (43) and $\alpha \geq 8\alpha'(1+2\theta_{i,j})$ does infact hold. We choose $\alpha = 8\alpha'(1+2\theta_{i,j}) + 4\sqrt{\alpha'(\eta+\tilde{\eta}')}$.Clearly, $\alpha \geq 8\alpha'(1+2\theta_{i,j})$ holds. We will now show that (43) holds.

$$\begin{aligned}
&4\left(\frac{1+2\theta_{i,j}}{\alpha} + \frac{(1+2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1+\theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j}\right) \\
&+ 2\tilde{\eta} + \frac{4(1+2\theta_{i,j})\alpha'(\eta+\tilde{\eta}')}{\alpha} \\
&\overset{(i)}{\leq} 4\left(\frac{1}{8} + \frac{(1+2\theta_{i,j})\alpha}{16} + 2\theta_{i,j}^2\alpha' + \frac{1+\theta_{i,j}}{4} + 2\theta_{i,j}\alpha'\right) + 2\tilde{\eta} + (1+2\theta_{i,j})\sqrt{\alpha'(\eta+\tilde{\eta}')} \\
&\overset{(ii)}{=} 4\left(\frac{1}{8} + \frac{(1+2\theta_{i,j})^2}{2} + 2\theta_{i,j}^2\alpha' + \frac{1+\theta_{i,j}}{4} + 2\theta_{i,j}\alpha'\right) + 2\tilde{\eta} + 2(1+2\theta_{i,j})\sqrt{\alpha'(\eta+\tilde{\eta}')} \\
&\leq 5\alpha'(1+2\theta_{i,j})^2 + 2\tilde{\eta} + 2(1+2\theta_{i,j})\sqrt{\alpha'(\eta+\tilde{\eta}')} = \eta',
\end{aligned} \tag{44}$$

where (i) follows from $\alpha \geq \max\{8\alpha'(1+2\theta_{i,j}), 4\sqrt{\alpha'(\eta+\tilde{\eta}')}\}$, (ii) follows from our choice of α, and the last inequality follows from simple algebraic manipulations. This completes our proof.

□

E.2 Under-Exploration and Self Correction

Lemmas 19 and 20 apply the inductive step established in Appendix E.1 to relate the true regret (with respect to the best policies in different classes) to the regret according to estimated models. Lemma 20 in particular demonstrates the self-correction property of Mod-IGW.

Lemma 19. *Suppose the event $\mathcal{W}_1$ holds. Consider any class indices $i, j \in [M']$ such that $j \geq i$. Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$. For any policy $\pi \in \tilde{\Pi}_j \cup \Pi_{0,m}$, model $f_m \in \{\hat{f}_m, \hat{f}_{m,i}\}$, and epoch $m \leq \hat{m}_i + 1$, we have:*

$$\begin{aligned}
Reg_j(\pi) &\leq 2Reg_{f_m}(\pi) + \frac{\eta_{i,m}K}{\gamma_{m,i}} \\
Reg_{f_m}(\pi) &\leq 2Reg_j(\pi) + \frac{\eta_{i,m}K}{\gamma_{m,i}},
\end{aligned}$$

where $\eta_{i,m} = 100(1+2\theta_{i,j})^2(\gamma_{m-1,1}/\gamma_{m-1,i})$.

Proof. We will prove this by induction. The base case follows from the fact that for any policy $\pi \in \tilde{\Pi}_j \cup \Pi_{0,1}$ and model $f_1 \in \{\hat{f}_1, \hat{f}_{1,i}\}$, we have:

$$\begin{aligned} \mathrm{Reg}_j(\pi) &\leq 1 \leq -2 + \eta_1 K/\gamma_{1,i} \\ \mathrm{Reg}_{f_1}(\pi) &\leq 1 \leq -2 + \eta_1 K/\gamma_{1,i}. \end{aligned} \tag{45}$$

For the inductive step, fix some $m \leq \hat{m}_i$. Assume for any policy $\pi \in \tilde{\Pi}_j \cup \Pi_{0,m}$ and model $f_m \in \{\hat{f}_m, \hat{f}_{m,i}\}$, we have:

$$\begin{aligned} \mathrm{Reg}_j(\pi) &\leq 2\mathrm{Reg}_{f_m}(\pi) + \frac{\eta_{i,m} K}{\gamma_{m,i}} \\ \mathrm{Reg}_{f_m}(\pi) &\leq 2\mathrm{Reg}_j(\pi) + \frac{\eta_{i,m} K}{\gamma_{m,i}}. \end{aligned} \tag{46}$$

Let $f_{m+1} \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. From Lemma 14 and Lemma 17, we have that (47) holds:

$$\begin{aligned} &\mathrm{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \leq \frac{\tilde{\eta} K}{\gamma_{m+1,i}}, \\ &\mathrm{Reg}_{\hat{f}_m}(\pi) \leq 4\mathrm{Reg}_{f_{m+1}}(\pi) + \frac{\tilde{\eta}' K}{\gamma_{m,i}}, \quad \forall \pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \ldots, p_{m+1,M'}\}, \\ &\tilde{\eta} = 28\frac{\gamma_{m,1}}{\gamma_{m,i}}, \;\; \tilde{\eta}' = 44\frac{\gamma_{m,1}}{\gamma_{m,i}}. \end{aligned} \tag{47}$$

Now from (46), (47), and Lemma 18, we have that (48) holds:

$$\begin{aligned} \forall \pi \in \tilde{\Pi}_j \cup \Pi_{0,m+1}, \;\; \mathrm{Reg}_j(\pi) &\leq 2\mathrm{Reg}_{f_{m+1}}(\pi) + \frac{\eta'_{i,m+1} K}{\gamma_{m+1,i}}, \\ \mathrm{Reg}_{f_{m+1}}(\pi) &\leq 2\mathrm{Reg}_j(\pi) + \frac{\eta'_{i,m+1} K}{\gamma_{m+1,i}}, \end{aligned} \tag{48}$$

where $\eta'_{i,m+1} = 5(\gamma_m/\gamma_{m,i})(1+2\theta_{i,j})^2 + 2\tilde{\eta} + 2(1+2\theta_{i,j})\sqrt{(\gamma_m/\gamma_{m,i})(\eta_{i,m}+\tilde{\eta}')}$. To complete our inductive argument, we only need to argue that $\eta'_{i,m+1} \leq \eta_{i,m+1}$; we will now show this.

$$\begin{aligned} \eta'_{i,m+1} &\leq 5\frac{\gamma_{m,1}}{\gamma_{m,i}}(1+2\theta_{i,j})^2 + 56\frac{\gamma_{m,1}}{\gamma_{m,i}} + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,1}}{\gamma_{m,i}}(\eta_{i,m} + 44\frac{\gamma_{m,1}}{\gamma_{m,i}})} \\ &\overset{(i)}{\leq} 75\frac{\gamma_{m,1}}{\gamma_{m,i}}(1+2\theta_{i,j})^2 + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,1}}{\gamma_{m,i}}\eta_{i,m}} \\ &\overset{(ii)}{=} 75\frac{\gamma_{m,1}}{\gamma_{m,i}}(1+2\theta_{i,j})^2 + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,1}}{\gamma_{m,i}}100(1+2\theta_{i,j})^2\frac{\gamma_{m-1,1}}{\gamma_{m-1,i}}} \\ &\overset{(iii)}{\leq} 95\frac{\gamma_{m,1}}{\gamma_{m,i}}(1+2\theta_{i,j})^2 \leq \eta_{i,m+1}, \end{aligned} \tag{49}$$

where (i) follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, (ii) follows from substituting $\eta_{i,m} = 100(1+2\theta_{i,j})^2(\gamma_{m-1,1}/\gamma_{m-1,i})$, and (iii) follows from $\frac{\gamma_{m-1,1}}{\gamma_{m-1,i}} \leq \frac{\gamma_{m,1}}{\gamma_{m,i}}$. This completes the inductive argument. $\square$

In the following lemma, we derive the self-correction property of our algorithm. That is, after a small number of epochs, we correct for effects of potential past under-exploration. This is evident in the factor of $\gamma_{m-1,1}/\gamma_{m-1,i}$ in the bound of Lemma 19, which is improved to a factor of $\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}$ in the bound of Lemma 20. Increasing epochs without increasing the candidate exploitation parameters helps reduce the term with the exponent, which converges to a constant within a small number of rounds.

Lemma 20. *Suppose the event $\mathcal{W}_1$ holds. Consider any two class indices $i, h \in [M']$ such that $h \leq i$. Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$. For any policy $\pi \in \tilde{\Pi}_j \cup \Pi_{0,m}$, model $f_m \in \{\hat{f}_m, \hat{f}_{m,i}\}$, and epoch $m \in [\hat{m}_{h-1}+1, \hat{m}_i+1]$, we have:*

$$\mathit{Reg}_j(\pi) \leq 2\mathit{Reg}_{f_m}(\pi) + \frac{\eta_{i,h,m} K}{\gamma_{m,i}},$$

$$\mathit{Reg}_{f_m}(\pi) \leq 2\mathit{Reg}_j(\pi) + \frac{\eta_{i,h,m}K}{\gamma_{m,i}},$$

where $\eta_{i,h,m} = 100(1+2\theta_{i,j})^2\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}$. *Further for* $m \in [\hat{m}_{h-1} + \lceil\log_2(\log_2(\gamma_{\hat{m}_{h-1},1}/\gamma_{\hat{m}_{h-1},i}))\rceil, \hat{m}_i+1]$*, we have* $\eta_{i,h,m} \leq 400(1+2\theta_{i,j})^2\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}$.

Proof. Consider any two class indices $i,h \in [M']$ such that $h \leq i$. We will prove the required bound by induction. The bound for the base case $m = \hat{m}_{h-1}+1$ follows from Lemma 19. Suppose the bound in Lemma 20 holds for class indices i,h and for some epoch $m \in [\hat{m}_{h-1}+1, \hat{m}_i]$. Let $f_{m+1} \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$. To complete our inductive argument, we will show the bound in Lemma 20 holds for class indices i,h and epoch $m+1$. Now, from Lemma 14 and Lemma 17 we have (50) holds.

$$\begin{aligned}
&\mathrm{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1,i}}) \leq \frac{\tilde{\eta}K}{\gamma_{m+1,i}},\\
&\mathrm{Reg}_{\hat{f}_m}(\pi) \leq 4\mathrm{Reg}_{f_{m+1}}(\pi) + \frac{\tilde{\eta}'K}{\gamma_{m,i}}, \quad \forall \pi \in \{\pi_{\hat{f}_{m+1}}, \pi_{\hat{f}_{m+1,i}}, p_{m+1,1}, \dots, p_{m+1,M'}\},\\
&\tilde{\eta} = 28\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}, \quad \tilde{\eta}' = 44\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}.
\end{aligned} \tag{50}$$

Now from (50), inductive hypothesis (bounds in Lemma 20 hold for epoch m) and Lemma 18, we have (51) holds.

$$\begin{aligned}
\forall \pi \in \tilde{\Pi}_j \cup \Pi_{0,m+1},\ \mathrm{Reg}_j(\pi) &\leq 2\mathrm{Reg}_{f_{m+1}}(\pi) + \frac{\eta'_{i,h,m+1}K}{\gamma_{m+1,i}},\\
\mathrm{Reg}_{f_{m+1}}(\pi) &\leq 2\mathrm{Reg}_j(\pi) + \frac{\eta'_{i,h,m+1}K}{\gamma_{m+1,i}},
\end{aligned} \tag{51}$$

where $\eta'_{i,h,m+1} = 5(\gamma_m/\gamma_{m,i})(1+2\theta_{i,j})^2 + 2\tilde{\eta} + 2(1+2\theta_{i,j})\sqrt{(\gamma_m/\gamma_{m,i})(\eta_{i,h,m}+\tilde{\eta}')}$. Note that since $m \geq \hat{m}_{h-1}+1$, we have $\gamma_m \geq \gamma_{m,h}$. To complete our inductive argument, we only need to argue that $\eta'_{i,h,m+1} \leq \eta_{i,h,m+1}$; we will now show this.

$$\begin{aligned}
\eta'_{i,h,m+1} &\leq 5\frac{\gamma_{m,h}}{\gamma_{m,i}}(1+2\theta_{i,j})^2 + 56\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\\
&\quad + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\eta_{i,h,m} + 44\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}\right)}\\
&\overset{(i)}{\leq} 75(1+2\theta_{i,j})^2\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}} + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,h}}{\gamma_{m,i}}\eta_{i,h,m}}\\
&\overset{(ii)}{=} 75(1+2\theta_{i,j})^2\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}}\\
&\quad + 2(1+2\theta_{i,j})\sqrt{\frac{\gamma_{m,h}}{\gamma_{m,i}}100(1+2\theta_{i,j})^2\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}}\\
&\overset{(iii)}{\leq} 95(1+2\theta_{i,j})^2\frac{\gamma_{m,h}}{\gamma_{m,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}}} \leq \eta_{i,h,m+1},
\end{aligned} \tag{52}$$

where (i) follows from $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for any $a,b \geq 0$, (ii) follows from substituting $\eta_{i,h,m} = 100(1+2\theta_{i,j})^2\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}\left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}}\right)^{1/2^{m-\hat{m}_{h-1}-1}}$, and (iii) follows from $\frac{\gamma_{m-1,h}}{\gamma_{m-1,i}} \leq \frac{\gamma_{m,h}}{\gamma_{m,i}}$. This completes the inductive argument.

Finally for $m \in [\hat{m}_{h-1} + \lceil \log_2(\log_2(\gamma_{\hat{m}_{h-1},1}/\gamma_{\hat{m}_{h-1},i})) \rceil, \hat{m}_i + 1]$, we have (53):

$$\begin{aligned}\eta_{i,h,m} &\leq 100(1+2\theta_{i,j})^2 \frac{\gamma_{m-1,h}}{\gamma_{m-1,i}} \left(\frac{\gamma_{\hat{m}_{h-1},1}}{\gamma_{\hat{m}_{h-1},i}} \right)^{2/2^{\log_2(\log_2(\gamma_{\hat{m}_{h-1},1}/\gamma_{\hat{m}_{h-1},i}))}} \\ &\leq 400(1+2\theta_{i,j})^2 \frac{\gamma_{m-1,h}}{\gamma_{m-1,i}}.\end{aligned} \tag{53}$$

□

F BOUNDING TIME TO DETECTION OF MISSPECIFICATION

In this section, we bound the number of rounds to determine whether class $\tilde{\mathcal{F}}_i$ is misspecified. In particular, under Assumptions 2 and 3, we show that that misspecification for class $\tilde{\mathcal{F}}_i$ is detected before the corresponding policy class bias dominates the corresponding variance. Unlike previous sections, the analysis in this section relies on Assumptions 2 and 3.

Proof outline: Lemma 21 first derives a minimum direct method evaluation error for models in $\tilde{\mathcal{F}}_i$ in terms of Δ_i. Lemma 22 allows us to re-write this error in terms of policy class bias β_i. The rest of our analysis bounds the number of rounds required to detect this error. Lemma 23 bounds the length of the epoch where misspecification is detected for class $\tilde{\mathcal{F}}_i$ in terms of the length of the epoch where misspecification is detected for class $\tilde{\mathcal{F}}_{i-1}$. Lemma 24 bounds the last round in an epoch in terms of its epoch length. Hence, Corollary 1 uses these results to bound the time to detect misspecification for class $\tilde{\mathcal{F}}_i$.

Lemma 21. *Suppose Assumption 3 holds. Consider some $i < j \in [M']$ such that $B_i > 0$ and $d_j \leq \omega d_i$. Further, consider any reward model $f \in \tilde{\mathcal{F}}_i$. We then have:*

$$\exists \pi \in \{\pi_f, \pi^*_j\}, \quad \Delta_i/2 \leq |R(\pi) - R_f(\pi)|. \tag{54}$$

Proof. From Assumption 3, we have (55) holds.

$$\Delta_i := R(\pi^*_j) - R(\pi^*_i). \tag{55}$$

Suppose for contradiction, assume that (56) holds.

$$\Delta_i/2 > |R(\pi) - R_f(\pi)| \quad \forall \quad \pi \in \{\pi_f, \pi^*_j\}. \tag{56}$$

We can decompose (56) to obtain the following two relations:

$$\begin{aligned}R(\pi_f) + \Delta_i/2 > R_f(\pi_f) \\ R(\pi^*_j) - \tilde{\Delta}_i/2 < R_f(\pi^*_j).\end{aligned} \tag{57}$$

By the definitions of π_f, we have $R_f(\pi_f) \geq R_f(\pi^*_j)$. Together with (57), this gives us:

$$\begin{aligned}&R(\pi^*_i) + \Delta_i/2 \geq R(\pi_f) + \Delta_i/2 > R_f(\pi_f) \geq R_f(\pi^*_j) > R(\pi^*_j) - \Delta_i/2 \\ &\implies R(\pi^*_j) - R(\pi^*_i) < \Delta_i.\end{aligned} \tag{58}$$

This contradicts (55), hence (56) must be false. Therefore (54) holds. □

Lemma 22. *Suppose Assumption 1 and Assumption 3 hold. Then for any $i < i^*$, we have $\Delta_i \geq \beta_i/(i^* - i) \geq \beta_i/\log_2(\tilde{d}_{i^*})$.*

Proof. The proof is fairly straightforward:

$$\begin{aligned}\beta_i =& R(\pi^*_{i^*}) - R(\pi_i) \\ =& \sum_{j=i}^{i^*-1} \Big(R(\pi^*_{j+1}) - R(\pi^*_j) \Big) \leq (i^* - i)\Delta_i.\end{aligned} \tag{59}$$

The first equality follows from definition of β_i and the last inequality follows from Assumption 3. Hence, we have $\Delta_i \geq \beta_i/(i^* - i)$. Now to complete the proof, note that $(i^* - i) \leq i^* \leq \log_2(\tilde{d}_{i^*})$ □

Lemma 23. *Suppose the events $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. Suppose also that Assumption 3 holds. Consider any $i \in [M']$ such that $B_i > 0$. There exists a constant C_2 such that the following holds:*

$$\tau_{\hat{m}_i} - \tau_{\hat{m}_i - 1} \leq (\tau_{\hat{m}_{i-1}} - \tau_{\hat{m}_{i-1}-1}) + C_2 \left(\frac{K}{\Delta_i^2}\right)^{1/\rho} \omega^2 \tilde{d}_i \ln(6M^3T^2/\delta). \tag{60}$$

Proof. We will prove Lemma 23 via induction. Note that the base case is trivially satisfied by defining $\tau_{\hat{m}_l} = 0$ for any $l \leq 0$. For our inductive hypothesis, suppose the statement in Lemma 23 holds for class index $i-1 \in [M'-1]$. To complete our inductive argument, we will show that the statement in Lemma 23 holds for class index $i \in [M']$.

We split our analysis into two cases, a trivial case and a more involved case. The first case is $\hat{m}_i \leq \hat{m}_{i-1} + l$, where $l = \log_2 \log_2(\gamma_{m,1}/\gamma_{m,i_{m+1}})$. In this case, the algorithm is still undergoing self-correction of the learning rates following detection of misspecification in model class $\hat{m}_{i-1}$, and so the epoch lengths are not yet doubling. Therefore, we have that $\tau_{\hat{m}_i} - \tau_{\hat{m}_i-1} = \tau_{\hat{m}_{i-1}} - \tau_{\hat{m}_{i-1}-1}$. Hence, the inequality we want to show (60) is trivially satisfied. The second case is $\hat{m}_i > \hat{m}_{i-1} + l$. Let $m = \hat{m}_i - 1$, and let j be the largest index such that $\tilde{d}_j \leq \omega \tilde{d}_i$. By Assumption 3 and Lemma 21, we know that (61) holds:

$$\exists \quad \pi \in \{\pi_{\hat{f}_{m+1,i}}, \pi_j^*\}, \quad \Delta_i/2 \leq |R(\pi) - R_{\hat{f}_{m+1,i}}(\pi)|. \tag{61}$$

Let $\theta_{i,j} := \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}}$, and let $\alpha > 0$ be a positive constant that we will fix later. Since $m \leq \hat{m}_i$, from Lemma 9 and (61), we have (62) holds:

$$\begin{aligned}\frac{\Delta_i}{2} \leq &\left(\frac{1+2\theta_{i,j}}{\alpha} + \frac{(1+2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1+\theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j}\right)\frac{K}{\gamma_{m+1,i}} \\ &+ \frac{(1+2\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \max_{\pi \in \{\pi_{\hat{f}_{m+1,i}}, \pi_j^*\}} \mathrm{Reg}_{\hat{f}_m}(\pi).\end{aligned} \tag{62}$$

We will now bound $\max_{\pi \in \{\pi_{\hat{f}_{m+1,i}}, \pi_j^*\}} \mathrm{Reg}_{\hat{f}_m}(\pi)$. From Lemma 20, the fact that $\mathrm{Reg}_j(\pi_j^*) = 0 = \mathrm{Reg}_{\hat{f}_{m+1,i}}(\pi_{\hat{f}_{m+1,i}})$, $m > \hat{m}_{i-1} + l$, and $\gamma_{m+1,i} \geq \gamma_{m,i}$ – we have (63) holds:

$$\begin{aligned}&\mathrm{Reg}_{\hat{f}_m}(\pi_j^*) \leq 400(1+2\theta_{i,j})^2 \frac{K}{\gamma_{m,i}}, \\ &\mathrm{Reg}_{\hat{f}_m}(\pi_{\hat{f}_{m+1,i}}) \leq 2\mathrm{Reg}_j(\pi_{\hat{f}_{m+1,i}}) + 400(1+2\theta_{i,j})^2 \frac{K}{\gamma_{m,i}} \leq 1200(1+2\theta_{i,j})^2 \frac{K}{\gamma_{m,i}}.\end{aligned} \tag{63}$$

Hence, combining (62) and (63), we have (64) holds:

$$\begin{aligned}\frac{\Delta_i}{2} \leq &\left(\frac{1+2\theta_{i,j}}{\alpha} + \frac{(1+2\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + 2(1+\theta_{i,j})^2/\alpha + 2\theta_{i,j})\gamma_m}{\gamma_{m+1,i}} + 2\theta_{i,j}\right)\frac{K}{\gamma_{m+1,i}} \\ &+ \frac{(1+2\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} 1200(1+2\theta_{i,j})^2 \frac{K}{\gamma_{m,i}}.\end{aligned} \tag{64}$$

We will now simplify (64). Since $m \in (\hat{m}_{i-1}, \hat{m}_i)$, we have $\gamma_m = \gamma_{m,i}$. Also note that $\gamma_{m,i} \leq \gamma_{m+1,i}$. Now by choosing $\alpha = 128(1+2\theta_{i,j})$, we get the following simplification of (64):

$$\begin{aligned}&\frac{\Delta_i}{2} \leq \left(\frac{1}{128} + 8(1+2\theta_{i,j})^2 + 2\theta_{i,j}^2 + (1+\theta_{i,j})/64 + 4\theta_{i,j} + 10(1+2\theta_{i,j})^2\right)\frac{K}{\gamma_{m,i}} \\ &\implies \Delta_i \leq 40(1+2\theta_{i,j})^2 \frac{K}{\gamma_{m,i}} \leq 360\ \theta_{i,j}^2 \sqrt{8C_1 K \left(\frac{\tilde{d}_i \ln(6M^3T^2/\delta)}{(\tau_{m-1} - \tau_{m-2})/2}\right)^\rho} \\ &\implies \tau_{\hat{m}_i} - \tau_{\hat{m}_i - 1} \leq 4(\tau_{m-1} - \tau_{m-2}) \leq 8 \cdot (8C_1 \cdot 360^2)^{1/\rho} \frac{K^{1/\rho}\theta_{i,j}^{4/\rho}\tilde{d}_i \ln(6M^3T^2/\delta)}{\Delta_i^{2/\rho}}.\end{aligned} \tag{65}$$

By substituting $\theta_{i,j} = \frac{\gamma_{m+1,i}}{\gamma_{m+1,j}} = \sqrt{\frac{d_j^\rho}{d_i^\rho}} \leq \omega^{\rho/2}$, we completes the proof of the inductive step for the second case. Hence this completes the proof of the inductive argument. □

Lemma 24. *For any epoch m, we have $\tau_m \leq 2l(\tau_m - \tau_{m-1})$, where $l = \lceil \log_2 \log_2(\gamma_{m,1}/\gamma_{m,i_m}) \rceil$.*

Proof. In Mod-IGW, at any epoch, the length of the next epoch is either equal to the length of the current epoch (if misspecification was recently detected) or two times the length of the current epoch (if it has been a while since misspecification was detected). By definition, within the first m epochs, misspecification is not detected for any class index in $[i_m, M']$. Hence for any epoch upto epoch m, at most l consecutive epochs have the same length. Since doubling more frequently would enable larger epoch length, given a bound on the length of epoch m, τ_m is the largest when epoch lengths only double once every l epochs. Hence $\tau_m \leq l \cdot 2(\tau_m - \tau_{m-1})$.

□

Corollary 1. *There exists a constant C_2 such that the following holds with probability at least $1-\delta$. Suppose Assumption 3 holds. Consider any $i \in [M']$ such that $B_i > 0$. We then have that:*

$$\tau_{\hat{m}_i} \leq 4lC_2 \left(\frac{K}{\Delta_i^2}\right)^{1/\rho} \omega^2 \tilde{d}_i \ln(6M^3T^2/\delta), \tag{66}$$

where $l = \lceil \log_2 \log_2(\gamma_{\hat{m}_i,1}/\gamma_{\hat{m}_i,i}) \rceil$.

Proof. For any $i \in [M']$, we have (67) holds.

$$\begin{aligned}
&\tau_{\hat{m}_i} - \tau_{\hat{m}_i - 1} \\
&\overset{(i)}{=} \sum_{h=1}^{i} ((\tau_{\hat{m}_h} - \tau_{\hat{m}_h - 1}) - (\tau_{\hat{m}_{h-1}} - \tau_{\hat{m}_{h-1}-1})) \\
&\overset{(ii)}{\leq} \sum_{h=1}^{i} C_2 \left(\frac{K}{\Delta_h^2}\right)^{1/\rho} \omega^2 \tilde{d}_h \ln(6M^3T^2/\delta) \\
&\overset{(ii)}{\leq} C_2 \left(\frac{K}{\Delta_i^2}\right)^{1/\rho} \omega^2 \tilde{d}_i \ln(6M^3T^2/\delta) \sum_{h=1}^{i} \frac{1}{2^{(i-h)}} \leq 2C_2 \left(\frac{K}{\Delta_i^2}\right)^{1/\rho} \omega^2 \tilde{d}_i \ln(6M^3T^2/\delta),
\end{aligned} \tag{67}$$

where (i) follows from the fact that $\tau_{\hat{m}_h} = 0$ for all $h \leq 0$, (ii) follows from Lemma 23, and (iii) follows from the fact that $\tilde{d}_{h+1} \geq 2\tilde{d}_h$ by construction and that $\Delta_{h+1} \leq \Delta_h$ (Assumption 3). The result now follows by combining (67) and Lemma 24.

□

G FINAL REGRET GUARANTEES

In this section, we derive our final regret bounds by utilizing the analysis in Lemma 20 and Corollary 1. Lemma 20 allows us to bound exploration regret until $\hat{m}_i$ (the epoch where misspecification is detected with respect to class $\tilde{\mathcal{F}}_i$). Corollary 1 bounds the time to detect misspecfication for various classes.

Theorem 1. *Suppose Assumptions 1 to 3 hold and the oracle subroutines perform as stated in Oracle Assumptions 1 to 4. With probability at least $1-\delta$: for any $i, j \in [M']$ such that $i \leq j$ and $\tilde{\mathcal{F}}_j$ not yet labelled as misspecified as of round T, Mod-IGW attains the following regret guarantee:*

$$\begin{aligned}
CReg_T \leq \tilde{\mathcal{O}}\Bigg(&(\omega^2 K^{1/\rho}) \frac{\tilde{d}_{i-1}}{\beta_{i-1}^{2/\rho}} \\
&+ \beta_j T + \left(\frac{\tilde{d}_j}{\tilde{d}_i}\right)^{\rho/2} \sqrt{K \tilde{d}_j^{\rho} T^{2-\rho}} \Bigg)
\end{aligned} \tag{14}$$

Here, $\tilde{\mathcal{O}}$ hides terms logarithmic in $T, M, 1/\delta, \tilde{d}_{i^*}$. Further, $\tilde{\mathcal{F}}_j$ is not determined to be misspecified for at least $\Omega(\tilde{d}_j/B_j^{1/\rho})$ rounds.

Proof. From Lemma 3 and Lemma 5, we have that both $\mathcal{W}_1$ and $\mathcal{W}_2$ hold with probability at least $1-\delta$. We now bound the expected cumulative regret up to round T while assuming that this high-probability event holds.

Let $\theta_{i,j} = (\tilde{d}_j/\tilde{d}_i)^{\rho/2}$ and let $m' \geq 1$ be the first epoch after detecting misspecification with respect to class $\tilde{\mathcal{F}}_i$ when we are guaranteed to self-correct (see Lemma 20) for possibly under-exploring with respect to the class i. That is, if $i=1$ let $m'=1$ and if $i>1$ let $m' = \hat{m}_{i-1} + \lceil \log_2(\log_2(\gamma_{\hat{m}_{i-1},1}/\gamma_{\hat{m}_{i-1},i}))\rceil)$. Note that for $i>1$, $\gamma_{\hat{m}_{i-1},1}/\gamma_{\hat{m}_{i-1},i} = (\tilde{d}_i/\tilde{d}_1)^{\rho/2} \leq \sqrt{\tilde{d}_i}$. Hence, $\tau_{m'} \leq \max(\tau_1, \tau_{\hat{m}_{i-1}} \log_2(\tilde{d}_i))$.

Now, from Lemma 20, we have that:

$$
\begin{aligned}
\mathrm{CReg}_T &:= \sum_{t=1}^{T} \mathrm{Reg}_{f^*}(p_{m(t)}) \overset{(i)}{=} \beta_j T + \sum_{t=1}^{T} \mathrm{Reg}_j(p_{m(t)}) \leq \beta_j T + \tau_{m'} + \sum_{t=\tau_{m'}+1}^{T} \mathrm{Reg}_j(p_{m(t)}) \\
&\overset{(ii)}{\leq} \beta_j T + \tau_{m'} + \sum_{t=\tau_{m'}+1}^{T} \left(2\mathrm{Reg}_{\hat{f}_{m(t)}}(p_{m(t)}) + \frac{400(1+2\theta_{i,j})^2 K}{\gamma_{m(t),i}} \right) \\
&\overset{(iii)}{\leq} \beta_j T + \tau_{m'} + \sum_{t=\tau_{m'}+1}^{T} \left(\frac{2K}{\gamma_{m(t),j}} + \frac{3600\theta_{i,j}^2 K}{\gamma_{m(t),i}} \right) \overset{(iv)}{\leq} \beta_j T + \tau_{m'} + \sum_{t=\tau_{m'}+1}^{T} \frac{3602\theta_{i,j} K}{\gamma_{m(t),j}} \\
&\overset{(v)}{\leq} \beta_j T + \tau_{m'} + \sum_{t=\tau_{m'}+1}^{T} 3602\theta_{i,j} \sqrt{8KC_1} \left(\frac{\tilde{d}_j \ln(6M^3T^2/\delta)}{(\tau_{m(t)-1} - \tau_{m(t)-2})/2} \right)^{\rho/2} \\
&\overset{(vi)}{\leq} \beta_j T + \tau_1 + \tau_{\hat{m}_{i-1}} \log_2(\tilde{d}_i) + (3602 \cdot 2^{\rho} \sqrt{8C_1}) \cdot \theta_{i,j} \sqrt{K} (\tilde{d}_j \ln(6M^3T^2/\delta))^{\rho/2} \sum_{m=m'+1}^{m(T)} \frac{\tau_{m(t)} - \tau_{m(t)-1}}{(\tau_{m(t)} - \tau_{m(t)-1})^{\rho/2}} \\
&\overset{(vii)}{\leq} \beta_j T + \tau_1 + 4 \log_2 \log_2(\tilde{d}_{i-1}) C_2 \left(\frac{K}{\Delta_{i-1}^2} \right)^{1/\rho} \omega^2 \tilde{d}_{i-1} \ln(6M^3T^2/\delta) \log_2(\tilde{d}_i) \\
&\quad + (3602 \cdot 2^{\rho} \sqrt{8C_1}) \cdot \theta_{i,j} \sqrt{K} (\tilde{d}_j \ln(6M^3T^2/\delta))^{\rho/2} T^{1-\rho/2} \log_2 \log_2(\tilde{d}_j) \log_2 T,
\end{aligned}
\tag{68}
$$

where (i) follows from $\beta_j := R(\pi_{f^*}) - R(\pi_j^*)$; (ii) follows from Lemma 20; (iii) follows from Lemma 1, the fact that $\gamma_m \geq \gamma_{m,j}$ (misspecification is not detected for class $\tilde{\mathcal{F}}_j$), and $(1+2\theta_{i,j})^2 \leq 9\theta_{i,j}^2$; (iv) follows from $\theta_{i,j} = \sqrt{\gamma_{m(t),i}/\gamma_{m(t),j}}$; (v) follows from our choice of $\gamma_{m(t),j}$; (vi) follows from $\tau_{m'} \leq \max(\tau_1, \tau_{\hat{m}_{i-1}} \log_2(\tilde{d}_i))$ and length of epoch $m(t)$ is at most double the size of length for epoch $m(t)-1$; (vii) follows from the bound on $\tau_{\hat{m}_{i-1}}$ from Corollary 1, the fact that $(\tau_{m(t)} - \tau_{m(t)-1})^{1-\rho/2} \leq T^{1-\rho/2}$ for any $t \leq T$, and the fact that $m(T) \leq \log_2 \log_2(\tilde{d}_j) \log_2 T$ (since misspecification is not detected for $\tilde{\mathcal{F}}_j$ and hence fraction of non-doubling rounds is at most $\log_2 \log_2(\tilde{d}_j)$). Finally, the regret guarantee follows from additionally noting that $\Delta_{i-1} \geq \beta_{i-1}/\log_2(\tilde{d}_{i^*})$ (Lemma 22).

Also, since misspecification with respect to class $\tilde{\mathcal{F}}_j$ is not detected until epoch m_j^* (see Lemma 7, Lemma 13, and Lemma 16), we know misspecification is not detected for at least $\Omega(\tilde{d}_j/B_j^{1/\rho})$ rounds.

□

H ADDITIONAL DETAILS

H.1 Constructing An Estimation Oracle

For completeness, we outline one of many approaches to construct an oracle that achieves the "fast rates" of Oracle Assumption 1. Consider a sequence of classes $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_i$ with VC subgraph dimensions of $d_1, d_2, \ldots, d_i$ respectively. Consider a probability kernel p and a natural number n. Consider n independently and identically drawn samples from the distribution $D(p)$. Let $\hat{f}_j$ be an estimator in $\mathcal{F}_j$ that minimizes empirical squared error loss over the first $\lceil n/2 \rceil$ samples. For any $\zeta \in (0,1)$, from fairly standard arguments based on local Rademacher complexities (see Theorem 5.2 and example 3 in chapter 5 of Koltchinskii, 2011), with probability $1-\zeta/(2i)$ we have:

$$
\mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p(\cdot|x)} [(\hat{f}_j(x,a) - f^*(x,a))^2] \leq (1+\epsilon) b_j(p) + \mathcal{O}\left(\frac{d_j \ln(ni/\zeta)}{n} \right),
\tag{69}
$$

where $\epsilon > 0$ is any fixed constant.[18] Now let $\hat{f}$ be an estimator in the set $\{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_i\}$ that minimizes empirical squared error loss over the remaining $\lfloor n/2 \rfloor$ samples. Again from using the same arguments based on localization (e.g. Mitchell et al., 2009; Koltchinskii, 2011), with probability $1 - \zeta/2$ we have:

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p(\cdot|x)}[(\hat{f}(x,a) - f^*(x,a))^2] \\ &\leq (1+\epsilon') \min_{j\in[i]} \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p(\cdot|x)}[(f_j(x,a) - f^*(x,a))^2] + \mathcal{O}\left(\frac{\ln(i/\zeta)}{n}\right),\end{aligned} \tag{70}$$

where $\epsilon' > 0$ is any fixed constant. By combining (69) and (70), with probability $1-\zeta$, we have:

$$\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim p(\cdot|x)}[(\hat{f}(x,a) - f^*(x,a))^2] \leq (1+\epsilon)(1+\epsilon')b_j(p) + \mathcal{O}\left(\frac{d_j \ln(ni/\zeta)}{n}\right). \tag{71}$$

This completes our outline for the construction of an oracle that satisfies Oracle Assumption 1. The approach described here is based on using empirical risk minimization on training and validation sets. Other approaches one could use include aggregation algorithms (see Lecué et al., 2014, and references therein), penalized regression (see relevant chapters in Koltchinskii, 2011; Wainwright, 2019), cross validation, etc.

H.2 Constructing an Implementation of a Misspecification Test Oracle

Oracle Assumption 4 describes a computational oracle to test/verify several inequalities. The test relies on several parameters, we can search over $\alpha > 0$ via single variable optimization methods and search over $i, j \in [M']$ and $f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$ via enumeration. The number of policies in $\Pi_{0,m+1,i}$ are few and the corresponding inequalities can be easily verified. Hence, we primarily need to argue that the inequalities corresponding to $\tilde{\Pi}_j$ in the "policy-based misspecification test" can be verified computationally.

For any choice of $\alpha > 0, i, j \in [M'], f \in \{\hat{f}_{m+1}, \hat{f}_{m+1,i}\}$, we restate the "policy-based misspecification test" that is used at the end of epoch m and argue how this test can be verified via two calls to a cost sensitive classification solver. First, let us restate the test as a maximization problem for a given set of parameters (here $\lambda_{i,j,\alpha}$ serves as a short-hand for the policy independent terms):

$$\begin{aligned}&\max_{\pi\in\Pi_j} |\hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi)| - \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) \\ &\leq \left(\frac{1+\theta_{i,j}}{\alpha} + \frac{(1+\theta_{i,j})\alpha}{16} + \frac{(2\theta_{i,j}^2 + (1+\theta_{i,j})^2/\alpha)\gamma_m}{\gamma_{m+1,i}} + \theta_{i,j}\right)\frac{K}{\gamma_{m+1,i}} =: \lambda_{i,j,\alpha}\end{aligned} \tag{72}$$

We are interested in calculating the value of the maximization problem in (72). To calculate this maximum, we need to fix our estimators. Let $\hat{R}_{m+1,f}(\pi) := \frac{1}{|S_{m,\mathbf{ho}}|}\sum_{t\in S_{m,\mathbf{ho}}} f(x_t, \pi(x_t)) = \frac{1}{|S_{m,\mathbf{ho}}|}\sum_{t\in S_{m,\mathbf{ho}}} \mathbb{E}_{a\sim\pi(\cdot|x_t)} f(x_t, a)$ 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), $\hat{R}_{m+1}(\pi) := \frac{1}{|S_{m,\mathbf{ho}}|}\sum_{t\in S_{m,\mathbf{ho}}} \frac{\pi(a_t|x_t) r_t(a_t)}{p_m(a_t|x_t)}$. [19] Note that the value of the maximization problem in (72) is equal to $\max(L_1, L_2)$, where $\{L_i | i \in [2]\}$ are defined as follows:

$$\begin{aligned}L_1 &:= \max_{\pi\in\Pi_j} \hat{R}_{m+1,f}(\pi) - \hat{R}_{m+1}(\pi) - \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi) \\ L_2 &:= \max_{\pi\in\Pi_j} \hat{R}_{m+1}(\pi) - \hat{R}_{m+1,f}(\pi) - \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}} \widehat{\mathrm{Reg}}_{m+1,\hat{f}_m}(\pi).\end{aligned} \tag{73}$$

Substituting value of these estimators for L_1 and L_2, we get:

$$\begin{aligned}L_1 &= \max_{\pi\in\Pi_j} \sum_{t\in S_{m,\mathbf{ho}}} \frac{1}{|S_{m,\mathbf{ho}}|}\left(f(x_t,\pi(x_t)) - \frac{\pi(a_t|x_t) r_t(a_t)}{p_m(a_t|x_t)} - \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}(\hat{f}_m(x_t, \pi_{\hat{f}_m}(x_t)) - \hat{f}_m(x_t, \pi(x_t)))\right) \\ L_2 &= \max_{\pi\in\Pi_j} \sum_{t\in S_{m,\mathbf{ho}}} \frac{1}{|S_{m,\mathbf{ho}}|}\left(\frac{\pi(a_t|x_t) r_t(a_t)}{p_m(a_t|x_t)} - f(x_t,\pi(x_t)) - \frac{(1+\theta_{i,j})\gamma_m}{\alpha\gamma_{m+1,i}}(\hat{f}_m(x_t, \pi_{\hat{f}_m}(x_t)) - \hat{f}_m(x_t, \pi(x_t)))\right).\end{aligned} \tag{74}$$

[18] Note that ϵ is zero when $\mathcal{F}_j$'s are convex or well-specified.

[19] Up to constant factors, IPS estimators give us the best rates in Oracle Assumption 3 with finite classes. These estimators are also used in several contextual bandit papers (e.g., Agarwal et al., 2014).

Clearly, both L_1 and L_2 are cost-sensitive classification problems (see Krishnamurthy et al., 2017, for problem definition).[20] Hence we propose an approach to implement Oracle Assumption 4.

H.3 General Estimation Rates

Recall that Mod-IGW uses estimation rates ξ_i defined in Oracle Assumption 1. Apart from Appendix F (bounding time to detect misspecification), our analysis allows for more flexible rates and does not rely on Assumptions 2 and 3. Hence, Mod-IGW can be used with more general rates and settings. In particular, we weaken the need for ξ_i's to share the same rate in n. We now describe the rates that allows for the rest of our analysis to go through (except Appendix F).

These more general rates can be described by two fairly benign conditions. First, we require ξ_i to be a non-increasing function of n. In particular, we require:[21]

$$\text{For all } i \in [M'] \text{ and } \zeta \in (0,1),\ \xi_i(n,\zeta) \text{ is non-increasing in } n. \tag{75}$$

The second condition helps us simplify notation. At a high-level, it requires larger classes indices to correspond to more complex classes and have slower estimation rates:[22]

$$\text{For all } i \in [M'] \text{ and } \zeta \in (0,1),\ \frac{\xi_i(n,\zeta)}{\xi_{i-1}(n,\zeta)} \text{ is non-increasing in } n \text{ and is} \geq 1, \tag{76}$$

where we define $\xi_0(n,\zeta) := \ln(1/\zeta)/n$, which is the estimation rate for estimating the mean of a one-dimensional bounded random variable.[23]

[20] 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).

[21] We require the first condition to ensure that $\gamma_{m,i}$ is non-decreasing in m.

[22] We use the second condition to ensure that $\gamma_{m,j}/\gamma_{m,i}$ is greater than or equal to one and is non-decreasing in m for $j \leq i$. We only require this condition to simplify notation and our results can easily be generalized.

[23] In general, estimation rates are never faster than ξ_0. So, this is not a strong condition to have and helps simplify notation when stating guarantees for some misspecification tests.