PESSIMISTIC MINIMAX LEARNING FOR PUBLIC-PRIVATE INFORMATION GAMES UNDER UNILATERAL COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study offline learning in two-player zero-sum contextual games with public and private information, motivated by strategic settings such as auctions and negotiations with private valuations. We introduce unilateral prescriptive concentrability and show that asymmetric information can change offline coverage through its effect on equilibrium behavior. For finite state-action spaces, we develop a pessimistic algorithm with an $\tilde{O}(1/\sqrt{n})$ exploitability rate, matching the standard sample-size dependence for fully observed minimax games. We further develop a pessimistic policy mirror descent framework, PPA-PMD, for general function approximation and obtain a unified $\tilde{O}(1/\sqrt{n} + 1/\sqrt{T})$ exploitability rate with no-regret actor updates. Together, these results provide the first theoretical framework for offline equilibrium learning under public-private information constraints.

1 INTRODUCTION

Offline reinforcement learning (RL) has emerged as a critical paradigm for developing autonomous agents in domains where real-time exploration is either prohibitively expensive, unethical, or safety-critical (Fujimoto et al., 2019; Levine et al., 2020). By leveraging large-scale, pre-collected datasets, offline RL aims to recover optimal policies without further environmental interaction. However, the transition from logged data to optimal behavior is fundamentally challenged by distribution shift—the mismatch between the behavior policy that generated the data and the target policy being learned (Fujimoto et al., 2019; Levine et al., 2020; Kumar et al., 2020). This shift often leads to extrapolation errors, where the learner overestimates the value of actions poorly represented in the dataset (Kumar et al., 2020; Wu et al., 2019). To address this, a significant body of literature has established the principle of pessimism as a provably efficient mechanism for handling distribution shift, typically through the use of lower confidence bounds (LCB) or conservative value estimates (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021).

While single-agent offline RL is well-studied, many high-stakes applications are best modeled as two-player zero-sum games (Shapley, 1953; Littman, 1994). In these competitive environments, the challenge of distribution shift is exacerbated by the need to identify a global minimax solution. Without suitable coverage, the equilibrium value can be unidentifiable, as the opponent's best response can pivot toward unknown out-of-distribution regions. Recent advancements have addressed this by extending the principle of pessimism to two-player zero-sum games, utilizing lower confidence bounds to ensure that the learned policies are at least as good as the behavior policy across the supported regions (Zhong et al., 2022; Zhang et al., 2023). However, these works typically assume a fully observed state space where both players possess a perfect view of the environment. In practice, strategic interactions—such as auctions or negotiations—often involve asymmetric observations where players hold private information hidden from their rivals (Vickrey, 1961; Myerson, 1981).

Strategic interactions with hidden information are formally modeled as Bayesian Games (Harsanyi, 1967) or games with Incomplete Information (Aumann et al., 1995; Kamien et al., 1990). In these settings, players observe a shared public context alongside their own private information, and they must act based on "prescriptions"—policies that condition on their local observations. This structure is a cornerstone of modern Operations Research (OR), particularly in the design of auctions (Milgrom & Weber, 1982; Roughgarden, 2010) and automated negotiation agents (Kraus, 2001; Jennings et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2001). Despite its prevalence, the problem of learning optimal strategies for such games from offline data remains unexplored. The unique informational constraints of public-private games introduce a novel challenge: the learner must not only handle distribution shift but also manage the public belief—the joint distribution of private information given the public context, which players use to estimate the hidden state of their opponent.

To the best of our knowledge, this work provides the first theoretical analysis of offline two-player zero-sum contextual games under a public-private information structure. Our model builds on the classical contextual bandit setting (Dudík et al., 2011; Swaminathan & Joachims, 2015) and captures applications such as auctions and negotiations with private valuations or costs. Information constraints change how players value actions under beliefs over hidden types, and therefore change which policy is an equilibrium. The coverage a dataset must provide is determined by the equilibrium the information structure induces, not by the size of the opponent's deviation class: at a common reference policy the prescriptive and fully observed coefficients coincide. We formalize this via *unilateral prescriptive concentrability* and construct a family in which the induced equilibrium shift changes the coefficient of Cui & Du (2022a) by an arbitrarily large factor, in either direction. Our contributions are summarized as follows:

- **Theoretical Foundations of Public-Private Coverage:** We introduce *unilateral prescriptive concentrability* and show that public-private information can substantially change offline coverage by altering equilibrium behavior.
- **Optimal Learning in Finite Settings:** We propose `PPA-Exact`, a pessimistic algorithm for finite state-action spaces. We prove that it achieves a statistical exploitability rate of $\tilde{\mathcal{O}}(1/\sqrt{n})$, matching the sample-size dependence for fully observed minimax games (Cui & Du, 2022a; Zhong et al., 2022).
- **Scalable Optimization and Unified Complexity:** We propose `PPA-PMD`, an iterative pessimistic policy mirror descent framework for general function approximation that respects the game's inherent information constraints. We establish a unified exploitability gap of $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/\sqrt{T})$, characterizing how statistical estimation from n samples and optimization regret over T iterations jointly determine the final performance in asymmetric strategic environments.

2 Preliminaries

We study a two-player zero-sum contextual game with a public-private information structure. The environment is defined by the tuple $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}_1, \mathcal{X}_2, \mathcal{A}_1, \mathcal{A}_2, r, \rho \rangle$, where the public context space $\mathcal{C}$, the private type spaces $\mathcal{X}_1, \mathcal{X}_2$, and the action spaces $\mathcal{A}_1, \mathcal{A}_2$ are finite. In each episode, the environment generates a latent state $s \in \mathcal{S}$ drawn from a fixed, unknown prior distribution $\rho \in \Delta(\mathcal{S})$. The latent state factors into a public context and two private types: $s = (c, x^1, x^2) \in \mathcal{S} \stackrel{\text{def}}{=} \mathcal{C} \times \mathcal{X}_1 \times \mathcal{X}_2$, where $c \in \mathcal{C}$ is a shared public context, and $x^i \in \mathcal{X}_i$ is the private context belonging exclusively to player $i \in \{1, 2\}$. Upon the realization of the state, players choose actions $a^1 \in \mathcal{A}_1$ and $a^2 \in \mathcal{A}_2$. Let $\mathcal{A} \stackrel{\text{def}}{=} \mathcal{A}_1 \times \mathcal{A}_2$ denote the joint action space. Player 1 receives a reward given by the function $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$. Because the game is zero-sum, Player 2 receives reward $-r(s, a)$.

2.1 Information Constraints and Prescriptive Policies

During execution, the latent state s is partially observed. Player 1 observes $o^1 = (c, x^1)$ and Player 2 observes $o^2 = (c, x^2)$. Consequently, the players cannot execute standard Markov policies $\pi(a \mid s)$. Instead, they are restricted to *prescriptive policies* that condition exclusively on their local observations. We define the feasible policy spaces Π_1 and Π_2 as the set of all such mappings:

$$\Pi_i \stackrel{\text{def}}{=} \{\pi_i : \mathcal{C} \times \mathcal{X}_i \to \Delta(\mathcal{A}_i)\}, \quad \text{for } i \in \{1, 2\}.$$

Any deployed strategy must reside within these asymmetric information constraints.

2.2 The Public Belief and Contextual Value

Because neither player has full access to the latent state, strategic evaluation requires a common informational baseline. We define the **public belief** as the conditional probability distribution over

the private types given the shared public context:

$$b(x^1, x^2 \mid c) \stackrel{\text{def}}{=} \mathbb{P}_{s\sim\rho}\left(X^1 = x^1, X^2 = x^2 \mid C = c\right).$$

For any fixed public context $c \in \mathcal{C}$, the interaction reduces to a Bayesian game. If the players commit to prescriptive policies $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, the expected reward conditioned on the public context is evaluated by marginalizing over the public belief:

$$r_c(\pi_1, \pi_2) \stackrel{\text{def}}{=} \sum_{x^1\in\mathcal{X}_1}\sum_{x^2\in\mathcal{X}_2} b(x^1, x^2 \mid c) \sum_{a^1\in\mathcal{A}_1}\sum_{a^2\in\mathcal{A}_2} \pi_1(a^1 \mid c, x^1)\pi_2(a^2 \mid c, x^2) r(c, x^1, x^2, a^1, a^2).$$

The global expected value of the game for Player 1 is the expectation of these contextual values over the marginal distribution of the public context, $\rho_{\mathcal{C}}(c)$: $V(\pi_1, \pi_2) \stackrel{\text{def}}{=} \mathbb{E}_{c\sim\rho_{\mathcal{C}}}\left[r_c(\pi_1, \pi_2)\right]$.

2.3 Offline Data and the CTDE Paradigm

We operate under the paradigm of Centralized Training with Decentralized Execution (CTDE) (Lowe et al., 2017; Foerster et al., 2018). We assume the offline learner (centralized trainer) has access to a *full-state logged* dataset containing n i.i.d. interactions generated by an unknown behavior policy pair (μ_1, μ_2):

$$\mathcal{D} = \left\{\left(c^{(k)}, x^{1(k)}, x^{2(k)}, a^{1(k)}, a^{2(k)}, r^{(k)}\right)\right\}_{k=1}^{n}.$$

This full-state logging assumption is standard in the multi-agent offline RL literature (Lyu et al., 2021; Meng et al., 2023; Pan et al., 2022; Tseng et al., 2022; Formanek et al., 2023). It also models platform-mediated markets, such as auctions, procurement, and logistics, where centralized historical logs are available for training while deployed agents act using only local information (Lowe et al., 2017).

2.4 Minimax Objective and the Exploitability Gap

The canonical objective of the zero-sum game is the identification of the minimax value:

$$V^\star \stackrel{\text{def}}{=} \max_{\pi_1\in\Pi_1}\min_{\pi_2\in\Pi_2} V(\pi_1, \pi_2).$$

The existence of this unique value and the well-posedness of the game under asymmetric information constraints are established in Section 3.1.

Given the offline dataset $\mathcal{D}$, the learner's goal is to output a decentralized policy $\widehat{\pi}_1 \in \Pi_1$ for Player 1 that minimizes exploitability. We evaluate the performance of $\widehat{\pi}_1$ using the **exploitability gap**:

$$Gap(\widehat{\pi}_1) \stackrel{\text{def}}{=} V^\star - \min_{\pi_2\in\Pi_2} V(\widehat{\pi}_1, \pi_2).$$

This measures the maximum exploitability of the learner's policy by an opponent equipped with perfect knowledge of $\widehat{\pi}_1$ but restricted to the feasible class Π_2.

3 Information Asymmetry and Coverage Analysis

In this section, we introduce unilateral prescriptive concentrability and compare it with its fully observed counterpart. We show that asymmetric information changes equilibrium behavior and therefore the state-action regions that require coverage; depending on the logged behavior, the resulting coverage requirement can be smaller or larger.

3.1 Existence of Contextual Value

Because the game lacks temporal transitions, it decomposes into a collection of independent Bayesian games—one for each public context $c \in \mathcal{C}$. For a fixed public context c and belief $b(\cdot, \cdot \mid c)$, the expected reward $r_c(\pi_1, \pi_2)$ is bilinear in the prescriptive policies $\pi_i \in \Pi_i(c) \stackrel{\text{def}}{=} \Delta(\mathcal{A}_i)^{\mathcal{X}_i}$.

Lemma 3.1 (Existence of Contextual Value (Kartik & Nayyar, 2021)). *For any public context $c \in \mathcal{C}$, there exists a unique value $V^\star(c)$ and a NE pair $(\pi_{1,\star}, \pi_{2,\star}) \in \Pi_1(c) \times \Pi_2(c)$ such that:*

$$V^\star(c) = \max_{\pi_1 \in \Pi_1(c)} \min_{\pi_2 \in \Pi_2(c)} r_c(\pi_1, \pi_2) = \min_{\pi_2 \in \Pi_2(c)} \max_{\pi_1 \in \Pi_1(c)} r_c(\pi_1, \pi_2).$$

Since the public context is observed by both players, policies can be chosen independently across contexts, so $\Pi_i = \prod_{c \in \mathcal{C}} \Pi_i(c)$ and $V^\star = \mathbb{E}_{c \sim \rho_\mathcal{C}}[V^\star(c)]$. We write $Gap_c(\pi_1) \stackrel{\text{def}}{=} V^\star(c) - \min_{\pi_2 \in \Pi_2(c)} r_c(\pi_1, \pi_2)$ for the contextual exploitability gap, so that $Gap(\pi_1) = \mathbb{E}_{c \sim \rho_\mathcal{C}}[Gap_c(\pi_1)]$.

3.2 Unilateral Prescriptive Coverage

The feasibility of offline learning depends on the extent to which the logged data $\mathcal{D}$ covers the state-action regions relevant to the optimal strategy. In two-player games, a learner's policy $\widehat{\pi}_1$ must be robust against any possible adversarial response within the feasible policy class (Cui & Du, 2022a). We formalize this via the *unilateral prescriptive concentrability* coefficient.

Assumption 3.2 (Unilateral Prescriptive Coverage). Write $\mathcal{E}_{\text{Pres}} \stackrel{\text{def}}{=} \text{NE}(\Pi_1, \Pi_2)$ for the set of Nash equilibria of the prescriptive game and $\mathcal{E}_{\text{Full}} \stackrel{\text{def}}{=} \text{NE}(\Pi_1^{\text{full}}, \Pi_2^{\text{full}})$ for those of the fully observed game. We define the *unilateral prescriptive concentrability* $C^\star_{\text{Pres}}$ by minimising over $\pi_\star = (\pi_{1,\star}, \pi_{2,\star}) \in \mathcal{E}_{\text{Pres}}$:

$$C^\star_{\text{Pres}} \stackrel{\text{def}}{=} \min_{\pi_\star \in \mathcal{E}_{\text{Pres}}} \max \left\{ \sup_{\nu \in \Pi_2} \max_{s \in \mathcal{S}, a \in \mathcal{A}} \frac{d^{\pi_{1,\star}, \nu}(s,a)}{d^\rho(s,a)}, \sup_{\mu \in \Pi_1} \max_{s \in \mathcal{S}, a \in \mathcal{A}} \frac{d^{\mu, \pi_{2,\star}}(s,a)}{d^\rho(s,a)} \right\}, \tag{1}$$

where $d^{\pi_1, \pi_2}(s,a) = \rho(c, x^1, x^2)\pi_1(a^1 \mid c, x^1)\pi_2(a^2 \mid c, x^2)$ is the density induced by information-constrained strategies, and $d^\rho(s,a)$ is the latent state-action density in the logged dataset $\mathcal{D}$. We assume $C^\star_{\text{Pres}} < \infty$.

We denote $d^\rho_{\min} = \min_{(s,a) \in \text{supp}(d^\rho)} d^\rho(s,a)$ as the minimum density of the supported state-action pairs in the behavior distribution.

3.3 Information Structure and Offline Coverage

A central question in multi-agent offline RL is how information constraints affect the "data-hardness" of the game. To quantify this effect, we compare our prescriptive concentrability to the standard *full-information concentrability* used in fully observed interactions (Cui & Du, 2022a).

In a fully observed baseline, players are not restricted by private information and can choose any policy $\pi_i \in \Pi_i^{\text{full}}$ that conditions on the entire latent state $s \in \mathcal{S}$. Following Cui & Du (2022a), the coefficient is defined as:

$$C^\star_{\text{Full}} \stackrel{\text{def}}{=} \min_{\pi_\star \in \mathcal{E}_{\text{Full}}} \max \left\{ \sup_{\nu \in \Pi_2^{\text{full}}} \max_{s,a} \frac{d^{\pi_{1,\star}, \nu}(s,a)}{d^\rho(s,a)}, \sup_{\mu \in \Pi_1^{\text{full}}} \max_{s,a} \frac{d^{\mu, \pi_{2,\star}}(s,a)}{d^\rho(s,a)} \right\}. \tag{2}$$

In $C^\star_{\text{Pres}}$, policies condition only on local observations, whereas in $C^\star_{\text{Full}}$ they condition on the full state. The two coefficients are also evaluated at equilibria of their respective games. Their difference can therefore arise from two sources: the equilibrium induced by the information structure and the information available to a unilateral deviator.

We first establish that the equilibrium effect can produce a coverage gap that scales with the number of private types.

Proposition 3.3 (Equilibrium-Induced Coverage Separation). *For every integer $K \geq 3$ there is a public–private zero-sum game with K private types per player and a full-support behavior policy for which $C^\star_{Pres} = 4K$ and $C^\star_{Full} = 4K^2$.*

Its proof is in Appendix B.1.3, which also shows that under other logging policies the comparison can reverse. The proposition shows that asymmetric information can change the coverage that offline learning requires through its effect on equilibrium behavior. We next ask whether the larger deviation class available under full information creates an additional difference. To isolate this effect, we hold the same reference policy fixed in both comparisons.

Lemma 3.4 (Equality at a Common Reference Policy). *For any reference policy $\bar{\pi}_1 : \mathcal{S} \to \Delta(\mathcal{A}_1)$ and any logged density d^ρ,*

$$\sup_{\nu \in \Pi_2} \max_{s,a} \frac{d^{\bar{\pi}_1,\nu}(s,a)}{d^\rho(s,a)} = \sup_{\nu \in \Pi_2^{full}} \max_{s,a} \frac{d^{\bar{\pi}_1,\nu}(s,a)}{d^\rho(s,a)},$$

where $d^{\bar{\pi}_1,\nu}(s,a) = \rho(s)\,\bar{\pi}_1(a^1 \mid s)\,\nu(a^2 \mid \cdot)$ with each policy conditioned on what it observes; symmetrically for Player 1's deviations against a fixed $\bar{\pi}_2$.

Its proof is in Appendix B.1.1. Because the criterion maximizes over a single state-action tuple, a prescriptive deviator can place all of its mass on the maximizing action at the corresponding local observation without observing the rival's private type. We also show in Appendix B.1 that the same equality holds for an integrated coverage criterion under decentralized logging.

Together, these results show that information structure affects offline coverage through the equilibrium behavior it induces. The comparison with fully observed games depends on how the logged data cover the respective equilibria, while at a fixed reference policy the two deviation classes yield the same pointwise coefficient. This motivates the unilateral prescriptive coverage condition used in the learning results below.

4 Offline Learning for Finite Asymmetric Games

In this section, we develop a pessimistic algorithm for two-player zero-sum public–private contextual games and prove an $\tilde{\mathcal{O}}(1/\sqrt{n})$ exploitability rate, matching the standard sample-size dependence for fully observed minimax games.

4.1 Empirical Estimators and Pessimistic Objective

The learner utilizes the full-state logged dataset $\mathcal{D} = \{(s^{(k)}, a^{(k)}, r^{(k)})\}_{k=1}^n$ to estimate the game parameters. For each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define the visit count $N(s,a) \stackrel{\text{def}}{=} \sum_{k=1}^n \mathbb{I}\{(s,a)^{(k)} = (s,a)\}$ and the corresponding empirical reward:

$$\widehat{r}(s,a) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^n r^{(k)} \mathbb{I}\{(s,a)^{(k)} = (s,a)\}}{N(s,a) \vee 1}.$$

We further define the marginal state counts $N(s) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} N(s,a)$ and the public contextual counts $N(c) \stackrel{\text{def}}{=} \sum_{x^1 \in \mathcal{X}_1, x^2 \in \mathcal{X}_2} N(c, x^1, x^2)$. To protect against distribution shift, we construct the *clipped pessimistic reward* $\underline{r}(s,a)$ using an uncertainty bonus $\beta(s,a)$:

$$\underline{r}(s,a) = \max\{\widehat{r}(s,a) - \beta(s,a), -1\}, \quad \text{where } \beta(s,a) = \sqrt{2}\sqrt{\frac{\iota}{N(s,a) \vee 1}}.$$

Here, $\iota \stackrel{\text{def}}{=} \log(4|\mathcal{S}||\mathcal{A}|/\delta)$ is the concentration parameter. For each public context c, let

$$\widehat{b}(x^1, x^2 \mid c) \stackrel{\text{def}}{=} \begin{cases} N(c, x^1, x^2)/N(c), & N(c) > 0, \\ b_0(x^1, x^2 \mid c), & N(c) = 0, \end{cases}$$

where $b_0(\cdot \mid c)$ is any fixed distribution on $\mathcal{X}_1 \times \mathcal{X}_2$. We use the conditionals $\widehat{b}(x^{-i} \mid c, x^i)$ induced by $\widehat{b}$, with arbitrary values on zero-probability conditioning events. We define the **pessimistic contextual objective** $L_c(\pi_1, \pi_2)$ as the expected value of the conservative reward under the empirical belief:

$$L_c(\pi_1, \pi_2) \stackrel{\text{def}}{=} \sum_{x^1, x^2} \widehat{b}(x^1, x^2 \mid c) \sum_{a^1, a^2} \pi_1(a^1 \mid c, x^1) \pi_2(a^2 \mid c, x^2) \underline{r}(c, x^1, x^2, a^1, a^2).$$

To address the challenge of learning from asymmetric offline data, we propose the Public-Private Asymmetric Exact Pessimism (PPA-Exact) algorithm. The algorithm proceeds in two main stages: first, it leverages the "global oracle" properties of the logged dataset to perform joint statistical estimation of the rewards and the public belief (Kartik & Nayyar, 2021). Second, it decomposes the global interaction into a collection of independent Bayesian games, identifying the minimax equilibrium over the space of prescriptive policies for each public context. By constructing conservative value

estimates, the learner ensures that the resulting decentralized policy remains robust against adversarial exploitation in regions with sparse data coverage.

Algorithm 1 Public-Private Asymmetric Exact Pessimism (PPA-Exact)

Input: Offline dataset $\mathcal{D} = \{(c^{(k)}, x^{1,(k)}, x^{2,(k)}, a^{1,(k)}, a^{2,(k)}, r^{(k)})\}_{k=1}^{n}$, confidence parameter $\delta \in (0,1)$.

1: **Initialize:** For all $s \in \mathcal{S}, a \in \mathcal{A}$, calculate visit counts $N(s,a) = \sum_{k=1}^{n} \mathbb{I}\{(s,a)^{(k)} = (s,a)\}$.

2: **Estimate Beliefs:** For each public context $c \in \mathcal{C}$, compute the empirical public belief:

$$\widehat{b}(x^1, x^2 \mid c) = \frac{\sum_{a^1,a^2} N(c, x^1, x^2, a^1, a^2)}{N(c)} \text{ if } N(c) > 0, \quad \text{and } b_0(\cdot \mid c) \text{ otherwise.}$$

3: **Construct Pessimistic Rewards:** Let $\iota = \log(4|\mathcal{S}||\mathcal{A}|/\delta)$. For all $s \in \mathcal{S}, a \in \mathcal{A}$:

4: Compute empirical reward: $\widehat{r}(s,a) = \frac{\sum_{k=1}^{n} r^{(k)} \mathbb{I}\{(s,a)^{(k)}=(s,a)\}}{N(s,a) \vee 1}$

5: Compute uncertainty bonus: $\beta(s,a) = \sqrt{2}\sqrt{\frac{\iota}{N(s,a)\vee 1}}$

6: Set clipped pessimistic reward: $\underline{r}(s,a) = \max\{\widehat{r}(s,a) - \beta(s,a), -1\}$

7: **for** each public context $c \in \mathcal{C}$ **do**

8: Define the expected pessimistic objective over prescriptions $\pi_1 \in \Pi_1(c)$ and $\pi_2 \in \Pi_2(c)$:

$$L_c(\pi_1, \pi_2) = \sum_{x^{1:2}, a^{1:2}} \widehat{b}(x^1, x^2 \mid c) \pi_1(a^1 \mid c, x^1) \pi_2(a^2 \mid c, x^2) \underline{r}(s,a)$$

9: Solve the resulting bilinear minimax optimization via Linear Programming:

$$\widehat{\pi}_{1,c} \in \operatorname*{argmax}_{\pi_1 \in \Pi_1(c)} \min_{\pi_2 \in \Pi_2(c)} L_c(\pi_1, \pi_2)$$

10: **end for**

Output: The information-constrained policy for player 1: $\widehat{\pi}_1 = \{\widehat{\pi}_{1,c}\}_{c \in \mathcal{C}}$.

4.2 Sample Complexity Analysis

Our main result characterizes the performance of PPA-Exact in terms of the exploitability gap.

Theorem 4.1. *Suppose the dataset size satisfies* $n \geq \frac{8}{d_{\min}^{p}} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)$. *Under Assumption 3.2, it holds with probability at least* $1-\delta$ *that the policy* $\widehat{\pi}_1$ *obtained from Algorithm 1 satisfies:*

$$Gap(\widehat{\pi}_1) \leq 2(2 + 2\kappa_b)\sqrt{\frac{C^{\star}_{Pres}|\mathcal{S}||\mathcal{A}|\iota}{n}},$$

where $\kappa_b \overset{\text{def}}{=} \frac{\sqrt{|\mathcal{X}_1||\mathcal{X}_2|}+\sqrt{2}}{\sqrt{2}}$ *and* $\iota \overset{\text{def}}{=} \log(4|\mathcal{S}||\mathcal{A}|/\delta)$. *Its full proof is in Appendix B.2.*

Proof Sketch. The proof proceeds in three stages. First, we establish *approximate pessimism*: since $\underline{r} \leq r$ (Lemma B.6), $L_c(\pi_1,\pi_2) \leq r_c(\pi_1,\pi_2) + \|b - \widehat{b}\|_1$, so L_c lower-bounds the true value up to the belief-estimation error. We also bound the underestimation, $r_c(\pi_1,\pi_2) - L_c(\pi_1,\pi_2) \leq \mathbb{E}_{s,a}[2\beta] + \|b - \widehat{b}\|_1$ (11).

Second, we utilize the minimax optimality of $\widehat{\pi}_1$ to bound the contextual gap by the pessimism of the equilibrium policy: $Gap_c(\widehat{\pi}_1) \leq \mathbb{E}_{d^{\pi_{1,\star},\widehat{\pi}_2}|c}[(2+2\kappa_b)\beta(s,a)]$ (12), where $\widehat{\pi}_2$ is the empirical best response.

Third, we integrate this bound over the public contexts (13). By applying a count concentration event and the Cauchy-Schwarz inequality, we bound the sum of uncertainty bonuses by the unilateral prescriptive concentrability $C^{\star}_{\text{Pres}}$ (14). This yields the stated $\tilde{\mathcal{O}}(1/\sqrt{n})$ sample-size dependence.

While `PPA-Exact` (Algorithm 1) leverages Linear Programming to identify the minimax equilibrium exactly, this centralized approach faces a "memory wall" in high-dimensional environments where the private information space $|\mathcal{X}|$ is massive. To address this, we provide a decentralized iterative

alternative, `PPA-Learning` (Algorithm 2), in Appendix A. By decomposing the global minimax objective into local decision problems through no-regret dynamics, we establish a unified exploitability gap of $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/\sqrt{T})$, characterizing the joint impact of statistical estimation and optimization regret. This iterative framework serves as the theoretical foundation for our extension to general function approximation in Section 5.

5 General Function Approximation

The previous analysis of finite games establishes the fundamental sample complexity of the public–private setting; however, the polynomial dependence on state-action cardinality limits applicability to large-scale environments. To address this, we extend the framework to general function approximation by replacing counting-based estimators with supervised learning and utilizing Policy Mirror Descent for decentralized optimization.

5.1 Offline Estimation and Version Space Pessimism

Instead of point-wise estimation, we utilize function generalization across the latent state $s = (c, x^1, x^2)$ and joint actions $a = (a^1, a^2)$.

Belief and Reward Estimation. We approximate the public belief using a conditional density class $\mathcal{B}$. The learner identifies the empirical best belief $\widehat{b} \in \mathcal{B}$ via Maximum Likelihood Estimation (MLE):

$$\widehat{b} = \operatorname{argmax}_{b \in \mathcal{B}} \frac{1}{n} \sum_{k=1}^{n} \log b(x^{1,(k)}, x^{2,(k)} \mid c^{(k)}). \tag{3}$$

Similarly, we model the joint reward using a regressor class $\mathcal{F}$ and compute the Empirical Risk Minimizer (ERM) $\widehat{r} \in \mathcal{F}$ by minimizing the mean squared error over $\mathcal{D}$: $\mathcal{L}_{\mathcal{D}}(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=1}^{n} \left(f(s^{(k)}, a^{(k)}) - r^{(k)}\right)^2$, and $\widehat{r} = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(f)$. To quantify statistical uncertainty in the FA regime, we adopt the principle of version space pessimism (Xie et al., 2021). Because our setting is a contextual bandit ($H = 1$), the Bellman error collapses to the standard regression error. We construct the version space $\mathcal{F}(\beta) \subseteq \mathcal{F}$ as the set of statistically plausible reward functions with near-optimal empirical risk:

$$\mathcal{F}(\beta) \stackrel{\text{def}}{=} \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}}(f) - \mathcal{L}_{\mathcal{D}}(\widehat{r}) \leq \beta\},$$

where $\beta > 0$ is a tolerance parameter scaled by the statistical complexity of $\mathcal{F}$. To protect against adversarial exploitation of poorly covered regions, the trainer constructs the globally pessimistic reward predictor by taking the lower envelope of this space:

$$\underline{r}(s, a) \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}(\beta)} f(s, a). \tag{4}$$

5.2 Asymmetric Policy Mirror Descent

With the pessimistic critic $\underline{r}$ and belief $\widehat{b}$ established, the decentralized actors π_{θ_1} and π_{θ_2} update their parameters via iterative mirror descent. At each iteration t, the algorithm evaluates the conditionally expected local Q-values for both players against the opponent's current strategy:

$$Q_{1,t}(c, x^1, a^1) = \mathbb{E}_{x^2 \sim \widehat{b}(\cdot \mid c, x^1), a^2 \sim \pi_{\theta_{2,t}}} \left[\underline{r}(c, x^1, x^2, a^1, a^2)\right], \tag{5}$$

$$Q_{2,t}(c, x^2, a^2) = \mathbb{E}_{x^1 \sim \widehat{b}(\cdot \mid c, x^2), a^1 \sim \pi_{\theta_{1,t}}} \left[\underline{r}(c, x^1, x^2, a^1, a^2)\right]. \tag{6}$$

Here $\widehat{b}(x^{-i} \mid c, x^i)$ denotes the conditional induced by $\widehat{b}$. The actors then take mirror-descent steps against these Q-values. Any update satisfying the external-regret condition of Assumption 5.3 is admissible; Appendix B.4 gives one parameterised implementation (Algorithm 3).

5.3 Theoretical Analysis

To establish the sample complexity of PPA-PMD, we adopt the standard assumption that the chosen function classes are sufficiently expressive to contain the ground-truth distributions (Xie & Jiang, 2021; Zhang et al., 2023).

Assumption 5.1 (Realizability)**.** The true joint reward function is contained in the regressor class, $r \in \mathcal{F}$. The true public belief distribution is contained in the density class, $b(\cdot \mid c) \in \mathcal{B}$ for all $c \in \mathcal{C}$. Every candidate is uniformly bounded on the same scale as the reward: $\mathcal{F} \subseteq [-1, 1]^{\mathcal{S}\times\mathcal{A}}$.

To handle general function approximation, we require the dataset to capture the structural variance of the version space. Let d^{π_1,π_2} be the joint state-action distribution. For the equilibrium $\pi_{1,\star}$ and its empirical best-response $\widehat{\pi}_2 = \operatorname{argmin}_{\pi_2} L_c(\pi_{1,\star}, \pi_2)$, we have:

Assumption 5.2 (Prescriptive Uniform FA Concentrability)**.** There exists a finite constant $\mathscr{C}_{\mathcal{F}} < \infty$ such that, for any fixed unilateral prescriptive deviation $\pi' \in \Pi_{\text{pres-uni}}$ and *every* subset $\mathcal{G} \subseteq \mathcal{F}$ with $r \in \mathcal{G}$ (in particular $\mathcal{G} = \mathcal{F}(\beta)$),

$$\mathbb{E}_{s,a\sim d^{\pi'}}\left[\sup_{f\in\mathcal{G}}(f(s,a) - r(s,a))^2\right] \le \mathscr{C}_{\mathcal{F}} \cdot \sup_{f\in\mathcal{G}} \mathbb{E}_{s,a\sim d^{\rho}}\left[(f(s,a) - r(s,a))^2\right],$$

where $\Pi_{\text{pres-uni}}$ denotes reference-anchored unilateral prescriptive deviations from the equilibrium $\pi_{1,\star}$, including the unilateral deviation pair $(\pi_{1,\star}, \widehat{\pi}_2)$.

Assumption 5.2 generalizes prescriptive unilateral coverage (Assumption 3.2) to function space, following the convention used in offline games (Zhang et al., 2023). This formulation is analogous to the integral concentrability coefficients used to handle version-space diameters in offline RL (Uehara & Sun, 2021; Yin et al., 2022) and enables the pointwise version-space pessimism in (4). Because it bounds the version-space envelope rather than each regressor's error, $\mathscr{C}_{\mathcal{F}}$ can also reflect the richness of $\mathcal{F}$.

We assume the actor updates satisfy a standard external-regret guarantee with respect to the pessimistic contextual objective.

Assumption 5.3 (Actor No-Regret Guarantee)**.** The iterates $(\pi_{1,t}, \pi_{2,t})_{t=1}^{T}$ produced by the actor updates admit external regret bounds $R_1(T)$ and $R_2(T)$ that hold uniformly over public contexts: for every $c \in \mathcal{C}$,

$$\sup_{\pi_1\in\Pi_1} \sum_{t=1}^{T} [L_c(\pi_1, \pi_{2,t}) - L_c(\pi_{1,t}, \pi_{2,t})] \le R_1(T),$$
$$\sup_{\pi_2\in\Pi_2} \sum_{t=1}^{T} [L_c(\pi_{1,t}, \pi_{2,t}) - L_c(\pi_{1,t}, \pi_2)] \le R_2(T).$$

Assumption 5.3 is a standard external-regret condition (Freund & Schapire, 1999; Cesa-Bianchi & Lugosi, 2006). In particular, multiplicative-weights updates satisfy $R_i(T) \le 2\sqrt{T \log(\max_i |\mathcal{A}_i|)}$, as shown in Appendix A.

We now state the performance guarantee.

Theorem 5.4 (Function-Approximation Exploitability Gap)**.** *Suppose the reward regressor class $\mathcal{F}$ and the belief density class $\mathcal{B}$ are finite. Under Assumptions 5.1, 5.2, and 5.3, if we set the version space tolerance to $\beta = \frac{136 \log(2|\mathcal{F}|/\delta)}{n}$, then with probability at least $1 - \delta$, the exploitability gap of the policy $\bar{\pi}_1$ produced by these updates satisfies:*

$$Gap(\bar{\pi}_1) \le 2\sqrt{\frac{272\mathscr{C}_{\mathcal{F}} \log(2|\mathcal{F}|/\delta)}{n}} + 4\sqrt{\frac{2\log(2|\mathcal{B}|/\delta)}{n}} + \frac{R_1(T) + R_2(T)}{T}.$$

For multiplicative-weights actor updates, $R_i(T) \le 2\sqrt{T \log(\max_i |\mathcal{A}_i|)}$, so the last term is $O(\sqrt{\log(\max_i |\mathcal{A}_i|)/T})$ and the bound becomes $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/\sqrt{T})$. The parameterised mini-batch implementation in Algorithm 3 is a practical variant; its stochastic optimisation error is not analysed here (Appendix B.4).

Proof Sketch. Its full proof is in Appendix B.5. The proof decomposes the gap into value, optimization, and pessimism errors. The pessimism error is bounded by the diameter of $\mathcal{F}(\beta)$ and the L_1 distance of the empirical belief $\widehat{b}$. That L_1 error is controlled by a finite-class Hellinger likelihood-ratio bound. We then bound the version-space squared error on the behavior measure via a fast-rate uniform Bernstein inequality, which requires setting $\beta \propto \log(|\mathcal{F}|)/n$ to apply a union

bound over the finite class $\mathcal{F}$. We transfer this statistical error to the target trajectory by applying Cauchy–Schwarz directly to the random variable $\sup_{f \in \mathcal{F}(\beta)} (f - r)^2$, followed by the uniform-ratio assumption $\mathscr{C}_{\mathcal{F}}$ in a single step. Substituting the actor regret of Assumption 5.3 gives the stated $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/\sqrt{T})$ rate.

6 Related Work

The Principle of Pessimism in Offline Learning. Offline reinforcement learning (RL) is fundamentally challenged by distribution shift, where the learner must infer an optimal policy from a fixed behavior dataset that may lack coverage of critical state–action regions (Levine et al., 2020). To provide statistical guarantees in the face of such uncertainty, the principle of *pessimism* has become the standard theoretical tool (Jin et al., 2021; Rashidinejad et al., 2021). By constructing lower confidence bounds (LCB) or conservative value estimates, pessimistic algorithms prevent the learner from being misled by overestimation bias in out-of-distribution regions (Xie et al., 2021; Uehara & Sun, 2021). While this framework has been shown to be minimax optimal for single-agent environments under various concentrability assumptions (Jin et al., 2021; Rashidinejad et al., 2021), these results typically assume that the learner has access to the full state of the environment.

Offline Learning for Two-Player Zero-Sum Games. Learning minimax equilibria in competitive environments extends the challenges of offline RL to strategic interactions (Littman, 1994). In two-player zero-sum games, the identification of a Nash Equilibrium (NE) requires the dataset to cover not only the equilibrium path but also any possible unilateral deviations by an adversary (Cui & Du, 2022a). Recent work has formalized this via the *unilateral concentrability* coefficient, proving it to be a necessary and sufficient condition for solvability in offline Markov games (Cui & Du, 2022b; Zhong et al., 2022). These approaches typically extend pessimism to the game-theoretic setting by constructing conservative value iterations for the minimax objective (Zhong et al., 2022; Zhang et al., 2023). However, existing literature largely assumes a fully observed state space where all players possess perfect information (Cui & Du, 2022a; Zhong et al., 2022; Zhang et al., 2023). Our analysis provides, to our knowledge, the first theoretical analysis of offline equilibrium learning in zero-sum games with asymmetric information. We introduce *unilateral prescriptive concentrability* and show that information structure changes offline coverage through the equilibrium it induces.

Information Asymmetry and Bayesian Strategic Games. Strategic interactions involving hidden information are formally characterized as Bayesian games or games with incomplete information (Harsanyi, 1967). Such models are fundamental to Operations Research and Economics, underpinning the design of complex auction mechanisms (Milgrom & Weber, 1982) and automated negotiation protocols (Roughgarden, 2010) where agents possess private signals hidden from their rivals. In modern multi-agent RL, this asymmetric structure is commonly navigated via the Centralized Training with Decentralized Execution (CTDE) paradigm (Lowe et al., 2017). While the existence of game values and the characterization of optimal strategies in these environments have been established from a planning perspective (Kartik & Nayyar, 2021), the problem of recovering these decentralized strategies from fixed, full-state historical logs remains largely unexplored. Our work bridges the gap between these game-theoretic foundations and offline RL by providing the first statistical framework that leverages the public belief to identify minimax-optimal policies directly from logged data.

7 Conclusion

In this paper, we studied offline learning in two-player zero-sum contextual games with public and private information, a structure common in economic interactions such as auctions, procurement, and negotiation. We introduced *unilateral prescriptive concentrability*, a coverage notion defined over the prescriptive policy space, and developed pessimistic learning methods that account for both statistical uncertainty and the information constraints of decentralized execution. For finite games, `PPA-Exact` achieves an $\tilde{\mathcal{O}}(1/\sqrt{n})$ exploitability rate, matching the standard sample-size dependence for fully observed minimax games, while our policy mirror descent framework for general function approximation, `PPA-PMD`, achieves $\tilde{\mathcal{O}}(1/\sqrt{n}+1/\sqrt{T})$ with no-regret actor updates. Extending these results to multi-stage games with dynamic beliefs and strategic signaling is a natural direction for future work.

AI USE STATEMENT

Generative AI tools were used solely for grammar correction. They were not used to conduct experiments, generate results, or for research ideation, methodology, theoretical development or proofs, code, substantive data analysis, or interpretation of results. All AI-assisted work was reviewed by the authors, who take responsibility for the final content.

REFERENCES

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012. 13

Robert J Aumann, Michael Maschler, and Richard E Stearns. *Repeated games with incomplete information*. MIT press, 1995. 1

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 8, 13

Qiwen Cui and Simon S Du. When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022a. 2, 4, 9

Qiwen Cui and Simon S Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Advances in Neural Information Processing Systems*, 35:11739–11751, 2022b. 9

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *ArXiv Preprint*, 2011. 2

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

Claude Formanek, Asad Jeewa, Jonathan Shock, and Arnu Pretorius. Off-the-grid marl: Datasets with baselines for offline multi-agent reinforcement learning. *arXiv preprint arXiv:2302.00521*, 2023. 3

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999. 8

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the International Conference on Machine Learning*, 2019. 1

John C Harsanyi. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967. 1, 9

Nicholas R Jennings, Peyman Faratin, Alessio R Lomuscio, Simon Parsons, Carles Sierra, and Michael Wooldridge. Automated negotiation: prospects, methods and challenges. *International Journal of Group Decision and Negotiation*, 10(2):199–215, 2001. 1

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021. 1, 9

Morton I Kamien, Yair Tauman, and Shmuel Zamir. On the value of information in a strategic conflict. *Games and Economic Behavior*, 2(2):129–153, 1990. 1

Dhruva Kartik and Ashutosh Nayyar. Upper and lower values in zero-sum stochastic games with asymmetric information. *Dynamic Games and Applications*, 11(2):363–388, 2021. 4, 5, 9

Sarit Kraus. *Strategic negotiation in multiagent environments*. MIT press, 2001. 1

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020. 1

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv Preprint*, 2020. 1, 9

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994. 1, 9

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017. 3, 9

Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.04402*, 2021. 3

Linghui Meng, Muning Wen, Chenyang Le, Xiyun Li, Dengpeng Xing, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, Yaodong Yang, et al. Offline pre-trained multi-agent decision transformer. *Machine Intelligence Research*, 20(2):233–248, 2023. 3

Paul R Milgrom and Robert J Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pp. 1089–1122, 1982. 1, 9

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981. 1

Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International conference on machine learning*, pp. 17221–17237. PMLR, 2022. 3

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021. 1, 9

Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010. 1, 9

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953. 1

Maurice Sion. On general minimax theorems. 1958. 22

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015. 2

Wei-Cheng Tseng, Tsun-Hsuan Johnson Wang, Yen-Chen Lin, and Phillip Isola. Offline multi-agent reinforcement learning with knowledge distillation. *Advances in Neural Information Processing Systems*, 35:226–237, 2022. 3

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021. 8, 9

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961. 1

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019. 1

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *Proceedings of the International Conference on Machine Learning*, 2021. 7

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021. 1, 7, 9

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022. 8

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *International Conference on Machine Learning*, pp. 40804–40829. PMLR, 2023. 1, 7, 8, 9

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022. 1, 2, 9

A Optimization via No-Regret Dynamics

A.1 Iterative Minimax Optimization

`PPA-Exact` solves each contextual pessimistic game by linear programming. We also consider an iterative alternative based on no-regret dynamics (Cesa-Bianchi & Lugosi, 2006). The public–private structure decomposes the minimax objective into local decision problems, so each player can update its policy by multiplicative weights (Arora et al., 2012) conditioned on its private observation. This yields an $O(T^{-1/2})$ optimization guarantee and provides the optimization framework used in Section 5.

A.2 Local Q-Values and Minimax Decomposition

For a fixed public context $c \in \mathcal{C}$, the expected pessimistic objective $L_c(\pi_1, \pi_2)$ can be rewritten in terms of local Q-values.

When Player 2 plays policy π_2, the conditional expected pessimistic reward for Player 1 (their local Q-value) is:

$$Q_1(x^1, a^1; \pi_2) \stackrel{\text{def}}{=} \sum_{x^2 \in \mathcal{X}_2, a^2 \in \mathcal{A}_2} \widehat{b}(x^2 \mid c, x^1)\pi_2(a^2 \mid c, x^2)\underline{r}(c, x^1, x^2, a^1, a^2). \tag{7}$$

Similarly, Player 2's local Q-value against π_1 is:

$$Q_2(x^2, a^2; \pi_1) \stackrel{\text{def}}{=} \sum_{x^1 \in \mathcal{X}_1, a^1 \in \mathcal{A}_1} \widehat{b}(x^1 \mid c, x^2)\pi_1(a^1 \mid c, x^1)\underline{r}(c, x^1, x^2, a^1, a^2). \tag{8}$$

By simulating simultaneous no-regret dynamics using Multiplicative Weights Updates independently at each private state, we can recover an $O(1/\sqrt{T})$-approximate saddle point of the empirical pessimistic objective L_c, for the time-averaged iterates, without ever explicitly constructing the full prescriptive LP. We demonstrate this method in Algorithm 2.

Algorithm 2 Public-Private Asymmetric No-Regret Learning (PPA-Learning)

Input: Public context c, empirical belief $\widehat{b}$, clipped pessimistic rewards $\underline{r}$, iterations T.
Initialize: Uniform policies $\pi_{1,1}(a^1 \mid c, x^1) = 1/|\mathcal{A}_1|$ and $\pi_{2,1}(a^2 \mid c, x^2) = 1/|\mathcal{A}_2|$.
Set Learning Rate: $\eta = \sqrt{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})/T}$.
for $t = 1, \ldots, T$ **do**
  Compute Local Q-values:
    For all x^1, a^1: compute $Q_{1,t}(x^1, a^1) = Q_1(x^1, a^1; \pi_{2,t})$ via Eq. (7)
    For all x^2, a^2: compute $Q_{2,t}(x^2, a^2) = Q_2(x^2, a^2; \pi_{1,t})$ via Eq. (8)
  Mirror Descent Policy Update:
    $\pi_{1,t+1}(a^1 \mid c, x^1) \propto \pi_{1,t}(a^1 \mid c, x^1) \exp\left(\eta Q_{1,t}(x^1, a^1)\right)$
    $\pi_{2,t+1}(a^2 \mid c, x^2) \propto \pi_{2,t}(a^2 \mid c, x^2) \exp\left(-\eta Q_{2,t}(x^2, a^2)\right)$
end for
Output: The average strategy for Player 1: $\bar{\pi}_{1,c} = \frac{1}{T}\sum_{t=1}^{T} \pi_{1,t}$.

A.3 Optimization Guarantee and Unified Exploitability Bound

The iterative solver approximates the empirical minimax solution at rate $O(T^{-1/2})$; combining this optimization guarantee with the statistical analysis of `PPA-Exact` gives the following result.

Theorem A.1. *Suppose the dataset size satisfies* $n \geq \frac{8}{d_{\min}^{\rho}} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)$. *Let* $\bar{\pi}_1$ *be the policy obtained by substituting Algorithm 2 into PPA-Exact, run for* T *iterations. Under Assumption 3.2, with probability at least* $1 - \delta$*:*

$$Gap(\bar{\pi}_1) \leq 2(2 + 2\kappa_b)\sqrt{\frac{C^{\star}_{Pres}|\mathcal{S}||\mathcal{A}|\iota}{n}} + 4\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}.$$

where $\kappa_b = \frac{\sqrt{|\mathcal{X}_1||\mathcal{X}_2|}+\sqrt{2}}{\sqrt{2}}$ *and* $\iota \stackrel{\text{def}}{=} \log(4|\mathcal{S}||\mathcal{A}|/\delta)$.

Its proof is in Appendix B.3.

B PROOF

B.1 COVERAGE SEPARATION: PROOFS FOR SECTION 3

Throughout this subsection we adopt the convention that $q(s,a)/d^\rho(s,a) = +\infty$ whenever $q(s,a) > 0 = d^\rho(s,a)$, and $0/0 = 0$; all statements hold verbatim under this convention. Every behavior density appearing in Section 3 — in Proposition 3.3 and, for every $t > 0$, the sharpness family of Proposition B.2 — has full support, so the convention is never invoked there; it is stated because C_Σ in Proposition B.2 is defined over all of $\mathcal{S} \times \mathcal{A}$ and would otherwise be ambiguous.

B.1.1 PROOF OF LEMMA 3.4

We prove the following precise form of Lemma 3.4, which also records the common value of the two suprema.

Lemma B.1 (Precise Form of Lemma 3.4)**.** *Fix any reference policy $\bar\pi_1 : \mathcal{S} \to \Delta(\mathcal{A}_1)$, prescriptive or fully observed, and read the induced density with the conditioning each policy admits,*

$$d^{\bar\pi_1,\nu}(s,a) \stackrel{\text{def}}{=} \rho(s)\,\bar\pi_1(a^1 \mid s)\,\nu(a^2 \mid \cdot), \qquad \nu(a^2 \mid \cdot) = \begin{cases} \nu(a^2 \mid c, x^2), & \nu \in \Pi_2, \\ \nu(a^2 \mid s), & \nu \in \Pi_2^{\textit{full}}, \end{cases}$$

so that a prescriptive deviator is embedded in $\Pi_2^{\textit{full}}$ by $\nu(a^2 \mid s) \stackrel{\text{def}}{=} \nu(a^2 \mid c, x^2)$. Then, for an arbitrary logged density d^ρ, not necessarily of decentralised product form,

$$\sup_{\nu\in\Pi_2} \max_{s,a} \frac{d^{\bar\pi_1,\nu}(s,a)}{d^\rho(s,a)} = \sup_{\nu\in\Pi_2^{\textit{full}}} \max_{s,a} \frac{d^{\bar\pi_1,\nu}(s,a)}{d^\rho(s,a)} = \max_{s,a} \frac{\rho(s)\,\bar\pi_1(a^1 \mid s)}{d^\rho(s,a)},$$

and symmetrically for Player 1's deviations against a fixed $\bar\pi_2 : \mathcal{S} \to \Delta(\mathcal{A}_2)$. When $\bar\pi_1 \in \Pi_1$ is prescriptive, $\bar\pi_1(a^1 \mid s)$ reduces to $\bar\pi_1(a^1 \mid c, x^1)$.

Proof. Fix a reference policy $\bar\pi_1 : \mathcal{S} \to \Delta(\mathcal{A}_1)$ and set

$$g(s,a) \stackrel{\text{def}}{=} \frac{\rho(s)\,\bar\pi_1(a^1 \mid s)}{d^\rho(s,a)}, \qquad G \stackrel{\text{def}}{=} \max_{s\in\mathcal{S},a\in\mathcal{A}} g(s,a),$$

where the maximum is attained because $\mathcal{S} \times \mathcal{A}$ is finite. For *any* ν, prescriptive or full-information, the induced density factorises as $d^{\bar\pi_1,\nu}(s,a) = \rho(s)\bar\pi_1(a^1 \mid s)\nu(a^2 \mid \cdot)$ — note that no property of the reference policy beyond this factorisation is used — so

$$\frac{d^{\bar\pi_1,\nu}(s,a)}{d^\rho(s,a)} \;\le\; g(s,a) \;\le\; G.$$

Writing $N \stackrel{\text{def}}{=} \rho(s)\bar\pi_1(a^1 \mid s)$, $D \stackrel{\text{def}}{=} d^\rho(s,a)$ and $t \stackrel{\text{def}}{=} \nu(a^2 \mid \cdot) \in [0,1]$, this holds in each of the three cases separately: if $D > 0$ the ratio is $g(s,a)\,t \le g(s,a)$; if $D = 0$ and $Nt > 0$ then $N > 0$, so $g(s,a) = G = +\infty$; and if $D = 0$ and $Nt = 0$ the ratio is 0 by the convention $0/0 = 0$. Taking $\max_{s,a}$ and then the supremum over either policy class gives the upper bound G for both.

For the matching lower bound, let $(s^\dagger, a^\dagger) = (c^\dagger, x^{1\dagger}, x^{2\dagger}, a^{1\dagger}, a^{2\dagger})$ attain G, and define $\nu^\dagger$ by

$$\nu^\dagger(\cdot \mid c^\dagger, x^{2\dagger}) = \delta_{a^{2\dagger}}, \qquad \nu^\dagger(\cdot \mid c, x^2) \text{ arbitrary for } (c,x^2) \neq (c^\dagger, x^{2\dagger}).$$

Crucially $\nu^\dagger \in \Pi_2$: it conditions only on the local observation (c, x^2), and its behaviour at the hidden types $x^1 \neq x^{1\dagger}$ that share the observation $(c^\dagger, x^{2\dagger})$ is irrelevant, because the criterion is a maximum over a *single* state-action tuple. Then

$$\sup_{\nu\in\Pi_2} \max_{s,a} \frac{d^{\bar\pi_1,\nu}(s,a)}{d^\rho(s,a)} \;\ge\; \frac{d^{\bar\pi_1,\nu^\dagger}(s^\dagger,a^\dagger)}{d^\rho(s^\dagger,a^\dagger)} = g(s^\dagger,a^\dagger)\cdot 1 = G.$$

Hence the prescriptive supremum equals G. Since $\Pi_2 \subseteq \Pi_2^{\text{full}}$, the full-information supremum is sandwiched between G and G as well. If $G = +\infty$ the same $\nu^\dagger$ makes both sides $+\infty$. The argument for Player 1's deviations against a fixed $\bar\pi_2 : \mathcal{S} \to \Delta(\mathcal{A}_2)$ is identical after exchanging the roles of the players, with $\mu^\dagger(\cdot \mid c^\dagger, x^{1\dagger}) = \delta_{a^{1\dagger}}$. Note that no property of d^ρ beyond nonnegativity was used, and nothing about which coordinates of s the reference policy is measurable with respect to; this is what makes the statement cover fully observed reference policies as well. □

B.1.2 AN INTEGRATED COVERAGE CRITERION

The following proposition is the robustness check referred to in Section 3: it shows that the equality of Lemma 3.4 is not an artifact of the pointwise L_∞ criterion.

Proposition B.2 (The equality at a common reference policy is not an artifact of the pointwise criterion)**.** *For a deviation class P define the integrated coefficient $C_\Sigma(\bar\pi_1; P) \stackrel{\text{def}}{=} \sup_{\nu\in P}\sum_{s,a} d^{\bar\pi_1,\nu}(s,a)/d^\rho(s,a)$, the sum running over all of $\mathcal{S}\times\mathcal{A}$ under Convention 1 of Appendix B.2, so that target mass placed off* $\mathrm{supp}(d^\rho)$ *makes the coefficient $+\infty$. For an arbitrary behavior density d^ρ, not necessarily of decentralised product form, the two coefficients are infinite simultaneously, and whenever they are finite*

$$C_\Sigma(\bar\pi_1;\Pi_2) \;\le\; C_\Sigma(\bar\pi_1;\Pi_2^{\textit{full}}) \;\le\; \min\{|\mathcal{X}_1|,|\mathcal{A}_2|\}\cdot C_\Sigma(\bar\pi_1;\Pi_2),$$

where $\min\{|\mathcal{X}_1|,|\mathcal{A}_2|\}$ *cannot be improved: for every $k\ge 2$ there are instances with $|\mathcal{X}_1|=|\mathcal{A}_2|=k$ and a family of* full-support *behavior densities along which the ratio tends to k. However, whenever the logged behavior is itself decentralised, $d^\rho(s,a)=\rho(s)\mu_1(a^1\mid c,x^1)\mu_2(a^2\mid c,x^2)$, the two coincide:*

$$C_\Sigma(\bar\pi_1;\Pi_2) = C_\Sigma(\bar\pi_1;\Pi_2^{\textit{full}}) \qquad \textit{for every reference policy } \bar\pi_1 \textit{ and every } \rho,\mu_1,\mu_2.$$

Proof. Fix any reference policy $\bar\pi_1:\mathcal{S}\to\Delta(\mathcal{A}_1)$, prescriptive or fully observed, and write $g(s,a)=\rho(s)\bar\pi_1(a^1\mid s)/d^\rho(s,a)$, read under Convention 1 and with the mixed-pair conditioning of Lemma 3.4. Both coefficients are $+\infty$ exactly when some (s,a) has $\rho(s)\bar\pi_1(a^1\mid s)>0=d^\rho(s,a)$: a prescriptive ν can place unit mass on that a^2 at (c,x^2), and a full-information ν can do no more, so the two are infinite together and the asserted identities hold trivially. Assume henceforth that both are finite, so that g is real-valued wherever the reference policy puts mass. Since the objective $\sum_{s,a} g(s,a)\nu(a^2\mid\cdot)$ is linear in ν over a product of simplices, the supremum is attained at a deterministic ν, and grouping the sum by the coordinates each class may condition on gives the closed forms

$$C_\Sigma(\bar\pi_1;\Pi_2) = \sum_{c,x^2}\max_{a^2}\sum_{x^1}\sum_{a^1} g(c,x^1,x^2,a^1,a^2),$$
$$C_\Sigma(\bar\pi_1;\Pi_2^{\text{full}}) = \sum_{c,x^1,x^2}\max_{a^2}\sum_{a^1} g(c,x^1,x^2,a^1,a^2).$$

Since a maximum of a sum never exceeds the sum of the maxima, $C_\Sigma(\bar\pi_1;\Pi_2)\le C_\Sigma(\bar\pi_1;\Pi_2^{\text{full}})$, and writing $M(x^1)=\max_{a^2}\sum_{a^1} g$ and $S=\max_{a^2}\sum_{x^1,a^1} g$ we have $M(x^1)\le S$ for each x^1, so the ratio is at most $|\mathcal{X}_1|$. It is also at most $|\mathcal{A}_2|$: choosing $j(x^1)\in\operatorname{argmax}_{a^2}\sum_{a^1} g$ and grouping the hidden types by their selected action,

$$\sum_{x^1} M(x^1) = \sum_{a^2\in\mathrm{image}(j)}\;\sum_{x^1:j(x^1)=a^2} M(x^1) \le |\mathrm{image}(j)|\cdot S \le \min\{|\mathcal{X}_1|,|\mathcal{A}_2|\}\cdot S,$$

since for each fixed a^2 the inner sum is at most S by definition of S.

The factor $\min\{|\mathcal{X}_1|,|\mathcal{A}_2|\}$ cannot be improved. Take $|\mathcal{C}|=1$, $|\mathcal{X}_2|=1$, $|\mathcal{X}_1|=k\ge 2$, $\mathcal{A}_1=\{\alpha\}$ a singleton (so $\bar\pi_1$ is forced), $\mathcal{A}_2=[k]$, and ρ uniform on $\mathcal{X}_1$, so that $\min\{|\mathcal{X}_1|,|\mathcal{A}_2|\}=k$. For $t\in(0,1/k]$ let the logged density concentrate on the graph $a^2=x^1$ with weight t and spread the rest uniformly,

$$d_t^\rho(x^1,\alpha,a^2) = \frac{1}{k}\left[t\,\mathbf{1}\{a^2=x^1\} + \frac{1-t}{k-1}\,\mathbf{1}\{a^2\ne x^1\}\right],$$

which is a probability distribution with *full support*, since $t+(k-1)\frac{1-t}{k-1}=1$ for each x^1. Here $g=\rho(x^1)\bar\pi_1(\alpha)/d_t^\rho$, so $\sum_{a^1} g$ equals $1/t$ at $a^2=x^1$ and $(k-1)/(1-t)$ otherwise. The restriction $t\le 1/k$ is exactly the condition $1/t\ge(k-1)/(1-t)$ that makes the diagonal entry the largest, whence

$$C_\Sigma(\bar\pi_1;\Pi_2^{\text{full}}) = \sum_{x^1}\max_{a^2}\sum_{a^1} g = \frac{k}{t}, \qquad C_\Sigma(\bar\pi_1;\Pi_2) = \max_{a^2}\sum_{x^1}\sum_{a^1} g = \frac{1}{t}+\frac{(k-1)^2}{1-t},$$

the latter because the inner sum takes the same value for every a^2. Therefore

$$\frac{C_\Sigma(\bar{\pi}_1;\Pi_2^{\text{full}})}{C_\Sigma(\bar{\pi}_1;\Pi_2)} = \frac{k}{1+\dfrac{t(k-1)^2}{1-t}} \;\longrightarrow\; k = \min\{|\mathcal{X}_1|,|\mathcal{A}_2|\} \quad \text{as } t \to 0^+,$$

strictly below k for every $t \in (0, 1/k]$, and tending to k.

Now impose $d^\rho(s,a) = \rho(s)\mu_1(a^1 \mid c, x^1)\mu_2(a^2 \mid c, x^2)$. On $\operatorname{supp}(d^\rho)$ the prior cancels and

$$g(s,a) = \frac{\bar{\pi}_1(a^1 \mid s)}{\mu_1(a^1 \mid c, x^1)\,\mu_2(a^2 \mid c, x^2)},$$

so, setting $G(s) \overset{\text{def}}{=} \sum_{a^1:\mu_1(a^1|c,x^1)>0} \bar{\pi}_1(a^1 \mid s)/\mu_1(a^1 \mid c, x^1) \geq 0$ and $h(c, x^2, a^2) \overset{\text{def}}{=} \mathbf{1}\{\mu_2(a^2 \mid c, x^2) > 0\}/\mu_2(a^2 \mid c, x^2)$, the inner sum factorises as

$$\sum_{a^1} g(c, x^1, x^2, a^1, a^2) = \mathbf{1}\{\rho(c, x^1, x^2) > 0\}\, G(s)\, h(c, x^2, a^2).$$

The dependence on a^2 enters only through h, which does not depend on x^1. Hence for every x^1 the maximisers over a^2 are the same — namely the least-logged actions $\operatorname{argmin}_{a^2:\mu_2>0} \mu_2(a^2 \mid c, x^2)$ — and, because $\mathbf{1}\{\rho > 0\}G(s) \geq 0$ does not depend on a^2, it can be taken outside the maximization over a^2,

$$\sum_{x^1} \max_{a^2} \sum_{a^1} g = \Big[\sum_{x^1} \mathbf{1}\{\rho > 0\}G(s)\Big] \max_{a^2} h(c, x^2, a^2) = \max_{a^2} \sum_{x^1} \sum_{a^1} g$$

for every (c, x^2). Summing over (c, x^2) gives the claimed equality. □

B.1.3 Proof of Proposition 3.3

Proposition 3.3 is the case $\mu(0) = 1/2$, $\mu(j) = 1/(2K)$ of the following statement, which we prove.

Proposition B.3 (Coverage Separation for General Symmetric Loggers)**.** *For every integer $K \geq 3$ there is a public–private zero-sum game, constructed in the proof below, whose prescriptive and fully observed games each admit a unique Nash equilibrium, and whose equilibria do not depend on the logged behavior. For* any *symmetric, type-independent behavior policy $\mu_1 = \mu_2 = \mu$ with full support on $\{0, 1, \ldots, K\}$,*

$$C^\star_{\textit{Pres}} = \frac{1}{\mu(0)\,\min_a \mu(a)}, \qquad C^\star_{\textit{Full}} = \frac{1}{\min_{j\in[K]} \mu(j) \cdot \min_a \mu(a)}, \qquad \frac{C^\star_{\textit{Full}}}{C^\star_{\textit{Pres}}} = \frac{\mu(0)}{\min_{j\in[K]} \mu(j)}. \tag{9}$$

In particular, $\mu(0) = 1/2$ and $\mu(j) = 1/(2K)$ give $C^\star_{\textit{Pres}} = 4K$ and $C^\star_{\textit{Full}} = 4K^2$, an improvement by the factor K; while $\mu(0) = 1/(K^2+1)$ and $\mu(j) = K/(K^2+1)$ give $C^\star_{\textit{Pres}} = (K^2+1)^2$ and $C^\star_{\textit{Full}} = (K^2+1)^2/K$, a deterioration by the same factor. Consequently neither $C^\star_{\textit{Pres}} \leq C^\star_{\textit{Full}}$ nor its reverse holds for every behavior distribution.

Proof. Fix $K \geq 3$. Let $|\mathcal{C}| = 1$, let X^1, X^2 be independent and uniform on $[K]$ so that $\rho(x^1, x^2) = 1/K^2$, and let $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1, \ldots, K\}$. With $u(a, x) = \mathbf{1}\{a = 0\} + 2 \cdot \mathbf{1}\{a = x\}$ we have $u \in \{0, 1, 2\}$, hence

$$r(x^1, x^2, a^1, a^2) = \tfrac{1}{2}\left[u(a^1, x^2) - u(a^2, x^1)\right] \in [-1, 1],$$

as required by the model. The logged behavior is $\mu_i(0 \mid x^i) = \mu(0)$ and $\mu_i(j \mid x^i) = \mu(j)$ for $j \in [K]$, with $\mu(0) + \sum_{j\in[K]} \mu(j) = 1$ and $\mu > 0$ everywhere, so d^ρ has full support. We keep μ general throughout and specialise only at the end.

Step 1: both games have a unique equilibrium, independent of μ. The reward is additively separable, $r = \frac{1}{2}u(a^1, x^2) - \frac{1}{2}u(a^2, x^1)$, so each player's optimal action is independent of the opponent's strategy and it suffices to maximise the player's own term. In particular the equilibria depend only on u, r and ρ, never on the behavior policy.

Under prescriptive execution Player 1 observes only (c, x^1), and since $X^2 \perp X^1$ is uniform on $[K]$,

$$\mathbb{E}\left[u(0, X^2)\right] = 1, \qquad \mathbb{E}\left[u(j, X^2)\right] = 2\,\mathbb{P}(X^2 = j) = \frac{2}{K} \quad \text{for } j \in [K].$$

Because $K \geq 3$ gives $2/K < 1$, the action $a^1 = 0$ is the strict maximiser at every type, hence strictly dominant in Π_1. Player 2 minimises r, equivalently maximises $u(a^2, X^1)$, so by the same computation $a^2 = 0$ is strictly dominant in Π_2. The prescriptive game therefore has the unique Nash equilibrium $\pi_\star^{\text{Pres}} = (\delta_0, \delta_0)$, and strict dominance rules out mixed equilibria as well.

Under full information Player 1 observes x^2 and compares $u(0, x^2) = 1$, $u(x^2, x^2) = 2$ (note $x^2 \in [K]$, so $x^2 \neq 0$), and $u(j, x^2) = 0$ for $j \notin \{0, x^2\}$. Thus $a^1 = x^2$ is strictly dominant, and symmetrically $a^2 = x^1$; the fully observed game has the unique equilibrium $\pi_\star^{\text{Full}} = (\delta_{x^2}, \delta_{x^1})$.

Step 2: a closed form for both coefficients. Because $d^\rho(s, a) = \rho(s)\mu_1(a^1 \mid x^1)\mu_2(a^2 \mid x^2)$, the prior cancels in every density ratio:

$$\frac{d^{\pi_{1,\star},\nu}(s, a)}{d^\rho(s, a)} = \frac{\pi_{1,\star}(a^1 \mid \cdot)}{\mu_1(a^1 \mid x^1)} \cdot \frac{\nu(a^2 \mid \cdot)}{\mu_2(a^2 \mid x^2)}.$$

Taking the supremum over ν and the maximum over (s, a), and using that for each fixed (x^1, x^2) the deviator may place unit mass on the least-covered action, gives

$$\sup_\nu \max_{s,a} \frac{d^{\pi_{1,\star},\nu}(s, a)}{d^\rho(s, a)} = \max_{x^1, x^2} \left\{ \left[\max_{a^1} \frac{\pi_{1,\star}(a^1 \mid \cdot)}{\mu_1(a^1 \mid x^1)} \right] \cdot \frac{1}{\min_{a^2} \mu_2(a^2 \mid x^2)} \right\},$$

where by Lemma 3.4 it is immaterial whether the supremum is taken over Π_2 or over Π_2^{full}, at either equilibrium; for the fully observed equilibrium policy $\pi_{1,\star}^{\text{Full}} \notin \Pi_1$ this is the fully observed case of that lemma. Since μ is type-independent, the second factor equals $1/\min_a \mu(a)$ for every x^2.

The two equilibrium policies differ only in which action the first bracket evaluates μ_1 at. At the prescriptive equilibrium $\pi_{1,\star} = \delta_0$ it is $1/\mu(0)$. At the fully observed equilibrium $\pi_{1,\star}(a^1 \mid s) = \mathbf{1}\{a^1 = x^2\}$ it is $1/\mu(x^2)$, maximised over $x^2 \in [K]$ at $1/\min_{j \in [K]} \mu(j)$. Neither expression depends on x^1.

The game is symmetric under exchanging the players, so the second argument of the $\max\{\cdot, \cdot\}$ in (1) and (2) takes the same value as the first, and both equilibria are unique, so each outer minimisation is over a singleton. Therefore

$$C_{\text{Pres}}^\star = \frac{1}{\mu(0) \min_a \mu(a)}, \qquad C_{\text{Full}}^\star = \frac{1}{\min_{j \in [K]} \mu(j) \cdot \min_a \mu(a)},$$

and the common factor $1/\min_a \mu(a)$ cancels in the ratio, leaving $C_{\text{Full}}^\star / C_{\text{Pres}}^\star = \mu(0)/\min_{j \in [K]} \mu(j)$. This is (9).

Step 3: the two instantiations. Take $\mu(0) = 1/2$ and $\mu(j) = 1/(2K)$, a valid policy since $1/2 + K/(2K) = 1$. Then $\min_a \mu(a) = 1/(2K)$, so $C_{\text{Pres}}^\star = 1/[(1/2)(1/(2K))] = 4K$ and $C_{\text{Full}}^\star = 1/[(1/(2K))(1/(2K))] = 4K^2$, a ratio of K.

Now take $\mu(0) = 1/(K^2+1)$ and $\mu(j) = K/(K^2+1)$, valid since $1/(K^2+1) + K^2/(K^2+1) = 1$. Here $\mu(0) < \mu(j)$ for $K \geq 2$, so $\min_a \mu(a) = \mu(0) = 1/(K^2+1)$, giving $C_{\text{Pres}}^\star = (K^2+1)^2$ and $C_{\text{Full}}^\star = (K^2+1)^2/K$, a ratio of $1/K$.

Hence $C_{\text{Full}}^\star / C_{\text{Pres}}^\star$ equals K in the first case and $1/K$ in the second, so neither ordering holds universally. □

B.2 Proof of Theorem 4.1

Proof. **Convention 1 (density ratios).** For a target distribution q we set $q(s, a)/d^\rho(s, a) \stackrel{\text{def}}{=} +\infty$ if $q(s, a) > 0 = d^\rho(s, a)$ and $\stackrel{\text{def}}{=} 0$ if $q(s, a) = d^\rho(s, a) = 0$. Finiteness of $C_{\text{Pres}}^\star$ then implies $d^{\pi_{1,\star},\nu} \ll d^\rho$ for every $\nu \in \Pi_2$, and symmetrically, so sums over $\mathcal{S} \times \mathcal{A}$ may be restricted to $\text{supp}(d^\rho)$.

Convention 2 (unobserved contexts). If $N(c) = 0$, then $\widehat{b}(\cdot \mid c)$ is the fallback $b_0(\cdot \mid c)$ of Section 4.

Throughout, $\iota = \log(4|\mathcal{S}||\mathcal{A}|/\delta) \geq \log 4 > 1$.

Before proving Theorem 4.1, we present and prove the following lemma.

Lemma B.4 (L_1 Concentration of Empirical Beliefs)**.** *For a fixed public context $c \in \mathcal{C}$, let $b(\cdot \mid c)$ be the true distribution over $\mathcal{X}_1 \times \mathcal{X}_2$ and $\widehat{b}(\cdot \mid c)$ be the empirical distribution derived from $n = N(c)$ i.i.d. samples, where the statement is to be read* conditionally on the realised context-count vector $(N(c))_{c\in\mathcal{C}}$*; see the remark following the proof. For any $\delta' \in (0,1)$, it holds with probability at least $1-\delta'$ that:*

$$\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1 \leq \mathcal{E}_b(c) \stackrel{\text{def}}{=} \begin{cases} \sqrt{\frac{|\mathcal{X}_1||\mathcal{X}_2|}{n}} + \sqrt{\frac{2\log(1/\delta')}{n}}, & n > 0, \\ 2, & n = 0. \end{cases}$$

Proof of Lemma B.4. Let $K = |\mathcal{X}_1||\mathcal{X}_2|$ be the cardinality of the joint private state space. Consider the function $f(X_1, \ldots, X_n) = \|b - \widehat{b}\|_1$, where $X_i \in \{1, \ldots, K\}$.

Step 1: Sensitivity Analysis. Changing a single sample X_i to X_i' alters the empirical frequency of at most two categories by $1/n$. Thus, the L_1 norm satisfies the bounded differences property with $c_i = 2/n$:

$$|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}.$$

Step 2: Concentration around the Expectation. By McDiarmid's Inequality, for any $\epsilon > 0$:

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X)] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(-\frac{2\epsilon^2}{n(4/n^2)}\right) = \exp\left(-\frac{n\epsilon^2}{2}\right).$$

Taking $\epsilon = \sqrt{2\log(1/\delta')/n}$ makes the tail bound equal to δ'.

Step 3: Bounding the Expected L_1 Deviation. It is a standard result for empirical distributions on a discrete set of size K that $\mathbb{E}[\|b - \widehat{b}\|_1] \leq \sqrt{\frac{K-1}{n}} \leq \sqrt{\frac{K}{n}}$.

Step 4: Exact High-Probability Bound. Combining the expectation bound and the concentration term via the triangle inequality, we obtain the exact bound:

$$\|b - \widehat{b}\|_1 \leq \sqrt{\frac{K}{n}} + \sqrt{\frac{2\log(1/\delta')}{n}}.$$

For $n = N(c) = 0$ no sampling has occurred and $\widehat{b}(\cdot \mid c) = b_0(\cdot \mid c)$ by Convention 2; there the bound is the trivial one, $\|b - \widehat{b}\|_1 \leq 2$, which is why $\mathcal{E}_b(c)$ is defined separately in that case. This serves as our exact definition for $\mathcal{E}_b(c)$. □

Remark B.5 (Removing the conditioning). Conditional on the context counts $(N(c'))_{c'\in\mathcal{C}}$, the type pairs observed at each context c are i.i.d. draws from $b(\cdot \mid c)$, so the lemma applies conditionally. A union bound over $c \in \mathcal{C}$ at level $\delta' = \delta/(4|\mathcal{C}|)$ then holds for every realisation of the counts, and hence unconditionally.

To ensure a global failure probability of at most δ, we allocate $\delta/2$ to belief estimation and $\delta/2$ to the count-concentration event. Let $\iota = \log(4|\mathcal{S}||\mathcal{A}|/\delta)$.

By Lemma B.4, we have $\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1 \leq \mathcal{E}_b(c)$. Recall that $\beta(s,a) = \sqrt{2}\sqrt{\frac{\iota}{N(s,a)\vee 1}}$. By setting the belief concentration failure probability to $\delta' = \frac{\delta}{4|\mathcal{C}|}$, we ensure the union bound holds and that $\log(1/\delta') \leq \iota$, the latter because $|\mathcal{S}||\mathcal{A}| = |\mathcal{C}||\mathcal{X}_1||\mathcal{X}_2||\mathcal{A}| \geq |\mathcal{C}|$. We bound the context-level belief error by the point-wise uncertainty bonus in two cases. If $N(c) > 0$, then $N(c) \geq N(s,a)$ for every latent state-action pair whose public component is c, and $\iota > 1$ gives $\sqrt{|\mathcal{X}_1||\mathcal{X}_2|} \leq \sqrt{|\mathcal{X}_1||\mathcal{X}_2|\,\iota}$, so we can reliably bound the context-level belief error using the point-wise uncertainty bonus:

$$\mathcal{E}_b(c) \leq \left(\sqrt{|\mathcal{X}_1||\mathcal{X}_2|} + \sqrt{2}\right)\sqrt{\frac{\iota}{N(s,a)\vee 1}} = \frac{\sqrt{|\mathcal{X}_1||\mathcal{X}_2|} + \sqrt{2}}{\sqrt{2}}\beta(s,a) \stackrel{\text{def}}{=} \kappa_b \beta(s,a), \tag{10}$$

for all $s \in \mathcal{S}, a \in \mathcal{A}$. If instead $N(c) = 0$, then $N(s,a) = 0$ at every such pair, so $\beta(s,a) = \sqrt{2\iota}$ and $\kappa_b \beta(s,a) = (\sqrt{|\mathcal{X}_1||\mathcal{X}_2|} + \sqrt{2})\sqrt{\iota} \geq (1+\sqrt{2})\sqrt{\iota} > 2 = \mathcal{E}_b(c)$, using only $|\mathcal{X}_1||\mathcal{X}_2| \geq 1$ and $\iota > 1$; so (10) holds in both cases. Here, we defined

$$\kappa_b \doteq \frac{\sqrt{|\mathcal{X}_1||\mathcal{X}_2|} + \sqrt{2}}{\sqrt{2}}.$$

To establish the *approximate pessimism* property of the objective L_c relative to the true expected reward r_c, we utilize the clipped pessimistic reward $\underline{r}(s,a) = \max\{\widehat{r}(s,a) - \beta(s,a), -1\}$. Because the true reward satisfies $r \in [-1,1]$, this projection ensures that $\|\underline{r}\|_\infty \leq 1$. The required pessimism property holds *deterministically*, with no reward-concentration event, because the model of Section 2 specifies rewards through a function $r : \mathcal{S} \times \mathcal{A} \to [-1,1]$.

Lemma B.6 (Deterministic Pessimism of the Clipped Reward)**.** *Assume rewards are deterministic, so that $r^{(k)} = r(s^{(k)}, a^{(k)})$ for every logged transition, and $\iota = \log(4|\mathcal{S}||\mathcal{A}|/\delta) \geq 1$. Then for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, surely,*

$$\underline{r}(s,a) \leq r(s,a) \qquad \textit{and} \qquad 0 \leq r(s,a) - \underline{r}(s,a) \leq 2\beta(s,a).$$

Proof. If $N(s,a) > 0$, determinism gives $\widehat{r}(s,a) = r(s,a)$ exactly, so $\widehat{r}(s,a) - \beta(s,a) \leq r(s,a)$; clipping at -1 preserves this because $r(s,a) \geq -1$, and $r(s,a) - \underline{r}(s,a) \leq \beta(s,a) \leq 2\beta(s,a)$. If $N(s,a) = 0$, then $\widehat{r}(s,a) = 0$ and $\beta(s,a) = \sqrt{2\iota} > 1$, so $\widehat{r}(s,a) - \beta(s,a) < -1$ and hence $\underline{r}(s,a) = -1 \leq r(s,a)$, while $r(s,a) - \underline{r}(s,a) = r(s,a) + 1 \leq 2 \leq 2\beta(s,a)$. Note $\iota \geq \log 4 > 1$ always holds, since $\delta < 1$ and $|\mathcal{S}||\mathcal{A}| \geq 1$. □

Consequently no failure probability need be budgeted for the reward estimate; the entire $\delta/2$ statistical budget is available to the belief estimate.

For any policy pair (π_1, π_2):

$$\begin{aligned} r_c(\pi_1,\pi_2) - L_c(\pi_1,\pi_2) &= \sum_{x,a} br\pi_1\pi_2 - \sum_{x,a} \widehat{b}\underline{r}\pi_1\pi_2 \\ &= \sum_{x,a} b(r-\underline{r})\pi_1\pi_2 + \sum_{x,a}(b-\widehat{b})\underline{r}\pi_1\pi_2 \\ &\leq \mathbb{E}_{s,a\sim d^{\pi_1,\pi_2}|c}[2\beta(s,a)] + \|b-\widehat{b}\|_1\|\underline{r}\|_\infty \\ &\qquad \text{(By } r - \underline{r} \leq 2\beta \text{ and Hölder's)} \\ &\leq \mathbb{E}_{s,a\sim d^{\pi_1,\pi_2}|c}[2\beta(s,a)] + \mathcal{E}_b(c). \\ &\qquad \text{(By } \|b-\widehat{b}\|_1 \leq \mathcal{E}_b(c) \text{ and } \|\underline{r}\|_\infty \leq 1) \end{aligned} \tag{11}$$

Recall that the true contextual value is $V^\star(c) = \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2)$. Note that for all $\pi_2 \in \Pi_2$:

$$\begin{aligned} L_c(\widehat{\pi}_1,\pi_2) &= \sum_{x,a} \widehat{b}\underline{r}\widehat{\pi}_1\pi_2 \\ &\leq \sum_{x,a} \widehat{b}r\widehat{\pi}_1\pi_2 \qquad \text{(Since } \underline{r} \leq r\text{)} \\ &= \sum_{x,a} br\widehat{\pi}_1\pi_2 + \sum_{x,a}(\widehat{b}-b)r\widehat{\pi}_1\pi_2 \\ &\leq r_c(\widehat{\pi}_1,\pi_2) + \|\widehat{b}-b\|_1 \left\| \sum_a r\widehat{\pi}_1\pi_2 \right\|_\infty \qquad \text{(Hölder's Inequality)} \\ &\leq r_c(\widehat{\pi}_1,\pi_2) + \mathcal{E}_b(c) \qquad \text{(Since } \|\widehat{b}-b\|_1 \leq \mathcal{E}_b(c), \|r\|_\infty \leq 1\text{)} \\ \implies \min_{\pi_2\in\Pi_2} L_c(\widehat{\pi}_1,\pi_2) - \mathcal{E}_b(c) &\leq \min_{\pi_2\in\Pi_2} r_c(\widehat{\pi}_1,\pi_2) = r_c(\widehat{\pi}_1,\nu_c^\star), \qquad \text{(Monotonicity of min)} \end{aligned}$$

where $\nu_c^\star = \operatorname{argmin}_{\pi_2\in\Pi_2} r_c(\widehat{\pi}_1,\pi_2)$. Therefore we have $-r_c(\widehat{\pi}_1,\nu_c^\star) \leq -\min_{\pi_2\in\Pi_2} L_c(\widehat{\pi}_1,\pi_2) + \mathcal{E}_b(c)$.

Using (11), the contextual exploitability gap satisfies:

$$\begin{aligned}
V^\star(c) - r_c(\widehat{\pi}_1, \nu_c^\star) &\le \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \min_{\pi_2 \in \Pi_2} L_c(\widehat{\pi}_1, \pi_2) + \mathcal{E}_b(c) \\
&\le \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \min_{\pi_2 \in \Pi_2} L_c(\pi_{1,\star}, \pi_2) + \mathcal{E}_b(c) \\
&\qquad \text{(Since } \widehat{\pi}_1 = \operatorname{argmax}_{\pi_1} \min_{\pi_2} L_c\text{)} \\
&\le r_c(\pi_{1,\star}, \widehat{\pi}_2) - L_c(\pi_{1,\star}, \widehat{\pi}_2) + \mathcal{E}_b(c) \quad \text{(Where } \widehat{\pi}_2 = \operatorname{argmin}_{\pi_2} L_c(\pi_{1,\star}, \pi_2)\text{)} \\
&\le \mathbb{E}_{s,a \sim d^{\pi_{1,\star}, \widehat{\pi}_2 | c}} [2\beta(s,a)] + 2\mathcal{E}_b(c) \qquad \text{(By (11))} \\
&\le \mathbb{E}_{s,a \sim d^{\pi_{1,\star}, \widehat{\pi}_2 | c}} [(2 + 2\kappa_b)\beta(s,a)]\,. \qquad \text{(Since } \mathcal{E}_b(c) \le \kappa_b \beta(s,a) \text{ point-wise)}
\end{aligned} \tag{12}$$

To bound the global gap, we integrate over the public context marginal $c \sim \rho_{\mathcal{C}}(c)$:

$$\begin{aligned}
Gap(\widehat{\pi}_1) &\le \mathbb{E}_{c \sim \rho_{\mathcal{C}}} \left[\mathbb{E}_{s,a \sim d^{\pi_{1,\star}, \widehat{\pi}_2 | c}} [(2 + 2\kappa_b)\beta(s,a)] \right] \\
&= (2 + 2\kappa_b) \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d^{\pi_{1,\star}, \widehat{\pi}_2}(s,a)\beta(s,a).
\end{aligned} \tag{13}$$

Define the count concentration event $\mathcal{E}_{\text{count}} \overset{\text{def}}{=} \{\forall (s,a) \in \text{supp}(d^\rho) : N(s,a) \ge \frac{1}{2} n d^\rho(s,a)\}$, which is the event actually established below and, by Convention 1, the only one the bound requires. By the multiplicative Chernoff bound, for a specific (s,a) such that $d^\rho(s,a) > 0$:

$$\mathbb{P}\left(N(s,a) < \frac{1}{2} n d^\rho(s,a) \right) \le \exp\left(-\frac{n d^\rho(s,a)}{8} \right).$$

Applying a union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ in the support of d^ρ:

$$\mathbb{P}(\mathcal{E}_{\text{count}}^c) \le |\mathcal{S}||\mathcal{A}| \exp\left(-\frac{n d^\rho_{\min}}{8} \right).$$

To ensure $\mathbb{P}(\mathcal{E}_{\text{count}}) \ge 1 - \delta/2$, we require $n \ge \frac{8}{d^\rho_{\min}} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta} \right)$. Under $\mathcal{E}_{\text{count}}$, the uncertainty bonus satisfies $\beta(s,a)^2 \le \frac{4\iota}{n d^\rho(s,a)}$. Substituting this into the global gap (13) and applying the Cauchy-Schwarz inequality:

$$\begin{aligned}
&Gap(\widehat{\pi}_1) \\
&\le (2 + 2\kappa_b) \sqrt{\sum_{s,a} d^{\pi_{1,\star}, \widehat{\pi}_2}(s,a)} \sqrt{\sum_{s,a} d^{\pi_{1,\star}, \widehat{\pi}_2}(s,a)\beta(s,a)^2} \qquad \text{(Cauchy-Schwarz)} \\
&\le (2 + 2\kappa_b)\sqrt{1} \cdot \sqrt{\sum_{s,a} d^{\pi_{1,\star}, \widehat{\pi}_2}(s,a) \frac{4\iota}{n d^\rho(s,a)}} \qquad \text{(Conditioned on } \mathcal{E}_{\text{count}}\text{)} \\
&= 2(2 + 2\kappa_b) \sqrt{\frac{\iota}{n} \sum_{s,a} \frac{d^{\pi_{1,\star}, \widehat{\pi}_2}(s,a)}{d^\rho(s,a)}} \\
&\le 2(2 + 2\kappa_b) \sqrt{\frac{C^\star_{\text{Pres}} |\mathcal{S}||\mathcal{A}| \iota}{n}} \qquad \text{(By Assumption 3.2)}
\end{aligned} \tag{14}$$

The final bound establishes the $\tilde{O}\left((1 + \kappa_b) \sqrt{\frac{C^\star_{\text{Pres}} |\mathcal{S}||\mathcal{A}|}{n}} \right)$ rate. □

B.3 Proof of Theorem A.1

Proof. We first establish the local regret guarantee for the Mirror Descent update using the Kullback-Leibler (KL) divergence as a potential function.

Lemma B.7 (Local MWU Regret). *Let $Q_t \in [-1,1]$ for all $t \in [T]$ and assume $|\mathcal{A}| \geq 2$, the case $|\mathcal{A}| = 1$ being immediate since the regret is then identically zero. For any fixed private state x and any fixed comparator distribution $\pi_\star(\cdot \mid x)$, the update rule $\pi_{t+1}(a \mid x) \propto \pi_t(a \mid x) \exp(\eta Q_t(x,a))$ with uniform initialization $\pi_1(a \mid x) = 1/|\mathcal{A}|$ and learning rate $\eta > 0$ guarantees:*

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{a}\pi_\star(a \mid x)Q_t(x,a) - \frac{1}{T}\sum_{t=1}^{T}\sum_{a}\pi_t(a \mid x)Q_t(x,a) \leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta}{2}.$$

Proof of Lemma B.7. Let $\Phi_t \stackrel{\text{def}}{=} \mathrm{KL}(\pi_\star(\cdot \mid x) \,\|\, \pi_t(\cdot \mid x))$. Since $\pi_{t+1}(a \mid x) = \pi_t(a \mid x)\, e^{\eta Q_t(x,a)}/Z_t$ with $Z_t \stackrel{\text{def}}{=} \sum_{a'} \pi_t(a' \mid x)\, e^{\eta Q_t(x,a')}$,

$$\Phi_t - \Phi_{t+1} = \sum_a \pi_\star(a \mid x) \log \frac{\pi_{t+1}(a \mid x)}{\pi_t(a \mid x)} = \eta \sum_a \pi_\star(a \mid x)\, Q_t(x,a) - \log Z_t.$$

Because $Q_t(x,\cdot) \in [-1,1]$, Hoeffding's lemma applied to $Q_t(x, A_t)$ with $A_t \sim \pi_t(\cdot \mid x)$ gives $\log Z_t \leq \eta \sum_a \pi_t(a \mid x)\, Q_t(x,a) + \eta^2/2$ for every $\eta > 0$, so

$$\Phi_t - \Phi_{t+1} \geq \eta \sum_a \big(\pi_\star(a \mid x) - \pi_t(a \mid x)\big) Q_t(x,a) - \frac{\eta^2}{2}.$$

Summing over $t \in [T]$ and using $\Phi_{T+1} \geq 0$ and, for the uniform initialization, $\Phi_1 = \log|\mathcal{A}| + \sum_a \pi_\star(a \mid x) \log \pi_\star(a \mid x) \leq \log|\mathcal{A}|$, we obtain $\eta \sum_{t=1}^{T} \sum_a (\pi_\star(a \mid x) - \pi_t(a \mid x))\, Q_t(x,a) \leq \log|\mathcal{A}| + \eta^2 T/2$. Dividing by ηT proves the claim.

□

Let $\widehat{V}(c) \stackrel{\text{def}}{=} \max_{\pi_1} \min_{\pi_2} L_c(\pi_1, \pi_2)$ be the exact idealized minimax value. Since the clipped rewards satisfy $\underline{r} \in [-1,1]$, the conditional expectations $Q_{1,t}$ and $Q_{2,t}$ are bounded in $[-1,1]$.

By applying Lemma B.7 to Player 1 at each private state $x^1 \in \mathcal{X}_1$ using the shared algorithm learning rate $\eta = \sqrt{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})/T}$, the right-hand side of the lemma evaluates to:

$$\frac{\log|\mathcal{A}_1|}{\eta T} + \frac{\eta}{2} \leq \frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{\eta T} + \frac{\eta}{2} = \frac{3}{2}\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}} \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}.$$

We lift these local bounds to the global objective by taking the expectation over the marginal belief $\widehat{b}(x^1 \mid c)$:

$$\sum_{x^1} \widehat{b}(x^1 \mid c) \left[\frac{1}{T}\sum_{t=1}^{T}\sum_{a^1}(\pi_1(a^1 \mid c, x^1) - \pi_{1,t}(a^1 \mid c, x^1))Q_{1,t}(x^1, a^1)\right] \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}. \tag{15}$$

By the definition of $Q_{1,t}$ and $\widehat{b}(x^1 \mid c)\,\widehat{b}(x^2 \mid c, x^1) = \widehat{b}(x^1, x^2 \mid c)$, we have $\sum_{x^1,a^1} \widehat{b}(x^1 \mid c)\, \pi_1(a^1 \mid c, x^1)\, Q_{1,t}(x^1, a^1) = L_c(\pi_1, \pi_{2,t})$ for every $\pi_1 \in \Pi_1$; applying this to π_1 and to $\pi_{1,t}$ in (15) gives the global regret bound for Player 1:

$$\frac{1}{T}\sum_{t=1}^{T} L_c(\pi_1, \pi_{2,t}) - \frac{1}{T}\sum_{t=1}^{T} L_c(\pi_{1,t}, \pi_{2,t}) \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}} \stackrel{\text{def}}{=} \epsilon_1. \tag{16}$$

For Player 2, who seeks to minimize the reward (maximize cost), Lemma B.7 is applied to the sequence of negative local Q-values $-Q_{2,t}(x^2, a^2)$. For any fixed comparator $\pi_2 \in \Pi_2$, the local cost-regret bound at each state $x^2 \in \mathcal{X}_2$ using the shared learning rate η is:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{a^2}(\pi_2(a^2 \mid c, x^2) - \pi_{2,t}(a^2 \mid c, x^2))(-Q_{2,t}(x^2, a^2)) \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}.$$

Distributing the negative sign into the policy difference and taking the expectation over the marginal belief $\widehat{b}(x^2 \mid c)$ yields:

$$\sum_{x^2} \widehat{b}(x^2 \mid c) \left[\frac{1}{T} \sum_{t=1}^{T} \sum_{a^2} (\pi_{2,t}(a^2 \mid c, x^2) - \pi_2(a^2 \mid c, x^2)) Q_{2,t}(x^2, a^2) \right] \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}.$$

In the same way, $\sum_{x^2,a^2} \widehat{b}(x^2 \mid c)\, \pi_2(a^2 \mid c, x^2)\, Q_{2,t}(x^2, a^2) = L_c(\pi_{1,t}, \pi_2)$ for every $\pi_2 \in \Pi_2$, including $\pi_2 = \pi_{2,t}$, which gives the second global regret guarantee:

$$\frac{1}{T} \sum_{t=1}^{T} L_c(\pi_{1,t}, \pi_{2,t}) - \frac{1}{T} \sum_{t=1}^{T} L_c(\pi_{1,t}, \pi_2) \leq 2\sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}} \stackrel{\text{def}}{=} \epsilon_2. \tag{17}$$

Since L_c is bilinear, $\frac{1}{T}\sum_t L_c(\pi_1, \pi_{2,t}) = L_c(\pi_1, \bar{\pi}_2)$ and $\frac{1}{T}\sum_t L_c(\pi_{1,t}, \pi_2) = L_c(\bar{\pi}_1, \pi_2)$ for all $\pi_1 \in \Pi_1$ and $\pi_2 \in \Pi_2$, where $\bar{\pi}_i \stackrel{\text{def}}{=} \frac{1}{T}\sum_t \pi_{i,t}$. Since (16) and (17) hold for any fixed policies, it follows that:

$$L_c(\pi_1, \bar{\pi}_2) - \epsilon_1 \leq \frac{1}{T} \sum_{t=1}^{T} L_c(\pi_{1,t}, \pi_{2,t}) \leq L_c(\bar{\pi}_1, \pi_2) + \epsilon_2, \quad \forall \pi_1 \in \Pi_1, \pi_2 \in \Pi_2.$$

Since this inequality holds for the entire policy space, we can take the supremum over π_1 on the left and the infimum over π_2 on the right to obtain the saddle-point optimality bound:

$$\max_{\pi_1 \in \Pi_1} L_c(\pi_1, \bar{\pi}_2) \leq \min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) + \epsilon_1 + \epsilon_2. \tag{18}$$

By the bilinearity of L_c and the fact that the policy spaces $\Pi_1(c)$ and $\Pi_2(c)$ are compact and convex, the conditions for Sion's Minimax Theorem (Sion, 1958) are satisfied. Thus, there exists a unique value for the empirical game:

$$\widehat{V}(c) \stackrel{\text{def}}{=} \min_{\pi_2 \in \Pi_2} \max_{\pi_1 \in \Pi_1} L_c(\pi_1, \pi_2). \tag{19}$$

From the definition of the minimum, it holds for the specific average policy $\bar{\pi}_2 \in \Pi_2$ that:

$$\widehat{V}(c) \leq \max_{\pi_1 \in \Pi_1} L_c(\pi_1, \bar{\pi}_2). \tag{20}$$

Substituting the bound (20) into the left-hand side of (18) yields the required relationship between the empirical value and the performance of our iterative policy:

$$\widehat{V}(c) \leq \min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) + \epsilon_1 + \epsilon_2.$$

Rearranging this inequality isolates the optimization error relative to the minimax value:

$$-\min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) \leq -\widehat{V}(c) + \epsilon_1 + \epsilon_2. \tag{21}$$

Finally, we substitute (21) into the contextual exploitability gap expansion from (12). Here and for the remainder of this proof, $\nu_c^\star \in \operatorname{argmin}_{\pi_2 \in \Pi_2} r_c(\bar{\pi}_1, \pi_2)$ denotes the best response against the iterative policy $\bar{\pi}_1$, so that $r_c(\bar{\pi}_1, \nu_c^\star) = \min_{\pi_2 \in \Pi_2} r_c(\bar{\pi}_1, \pi_2)$.

$$\begin{aligned} V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star) &\leq \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) + \mathcal{E}_b(c) \\ &\leq \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \widehat{V}(c) + \mathcal{E}_b(c) + \epsilon_1 + \epsilon_2. \end{aligned}$$

Since $\widehat{V}(c) = \max_{\pi_1} \min_{\pi_2} L_c(\pi_1, \pi_2) \geq \min_{\pi_2} L_c(\pi_{1,\star}, \pi_2)$, we have:

$$\begin{aligned} V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star) &\leq \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \min_{\pi_2 \in \Pi_2} L_c(\pi_{1,\star}, \pi_2) + \mathcal{E}_b(c) + \epsilon_1 + \epsilon_2 \\ &\leq \mathbb{E}_{s,a \sim d^{\pi_{1,\star}, \hat{\pi}_2 | c}} [(2 + 2\kappa_b)\beta(s,a)] + \epsilon_1 + \epsilon_2, \end{aligned}$$

where the final inequality follows from the statistical pessimism derivation in (12).

Algorithm 3 Public-Private Asymmetric Policy Mirror Descent (PPA-PMD)

Input: Offline dataset $\mathcal{D}$, function classes $\mathcal{F}$ and $\mathcal{B}$, uncertainty tolerance β, iterations T, step size $\eta > 0$, and initial parameters $\theta_{1,1}, \theta_{2,1}$ whose policies have full support.

Phase 1: Offline Supervised Estimation (The Critic)

Estimate belief $\widehat{b} \in \mathcal{B}$ via MLE ▷ Eq. (3)

Fit empirical reward $\widehat{r} \in \mathcal{F}$ and construct $\underline{r}(s,a) = \inf_{f \in \mathcal{F}(\beta)} f(s,a)$ ▷ Eq. (4)

Phase 2: Decentralized Actor Optimization

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

  Sample mini-batch of public contexts $c \sim \mathcal{D}$.

  Critic Evaluation: Compute local Q-values $Q_{1,t}$ and $Q_{2,t}$ ▷ Eqs. (5), (6)

  Policy Mirror Descent Update: expectations over the mini-batch contexts c and $x^i \sim \widehat{b}(x^i \mid c)$

    $\theta_{1,t+1} \leftarrow \operatorname{argmin}_{\theta_1} \mathbb{E}_{c,x^1}[-\eta \langle \pi_{\theta_1}, Q_{1,t} \rangle + \mathrm{KL}(\pi_{\theta_1} \| \pi_{\theta_{1,t}})]$

    $\theta_{2,t+1} \leftarrow \operatorname{argmin}_{\theta_2} \mathbb{E}_{c,x^2}[\eta \langle \pi_{\theta_2}, Q_{2,t} \rangle + \mathrm{KL}(\pi_{\theta_2} \| \pi_{\theta_{2,t}})]$

end for

Output: Average strategy for Player 1: $\bar{\pi}_1 = \frac{1}{T} \sum_{t=1}^{T} \pi_{\theta_{1,t}}$.

To bound the global gap, we integrate over the public context distribution $c \sim \rho_{\mathcal{C}}$. Because ϵ_1 and ϵ_2 are deterministic optimization constants independent of the context, the expectation perfectly splits:

$$\begin{aligned} Gap(\bar{\pi}_1) &= \mathbb{E}_{c \sim \rho_{\mathcal{C}}} \left[V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star) \right] \\ &\leq \mathbb{E}_{c \sim \rho_{\mathcal{C}}} \left[\mathbb{E}_{s,a \sim d^{\pi_{1,\star}, \widehat{\pi}_{2|c}}} [(2 + 2\kappa_b)\beta(s,a)] \right] + \epsilon_1 + \epsilon_2. \end{aligned}$$

Applying the global Cauchy-Schwarz integration derived in Equation (14) to bound the statistical expectation term, we obtain:

$$Gap(\bar{\pi}_1) \leq 2(2 + 2\kappa_b) \sqrt{\frac{C^\star_{\text{Pres}} |\mathcal{S}||\mathcal{A}| \iota}{n}} + \epsilon_1 + \epsilon_2. \tag{22}$$

To complete the proof, we substitute the explicitly symmetric optimization errors defined in (16) and (17). Because we utilized the shared learning rate bounded by the maximum action space, ϵ_1 and ϵ_2 are identical, yielding an exact sum:

$$\epsilon_1 + \epsilon_2 = 4 \sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}. \tag{23}$$

Substituting this sum into (22) yields the final exploitability guarantee:

$$Gap(\bar{\pi}_1) \leq 2(2 + 2\kappa_b) \sqrt{\frac{C^\star_{\text{Pres}} |\mathcal{S}||\mathcal{A}| \iota}{n}} + 4 \sqrt{\frac{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}{T}}.$$

This establishes the $\tilde{O}(1/\sqrt{n} + 1/\sqrt{T})$ convergence rate for the iterative asymmetric bandit solver.

□

B.4 Parameterized PPA-PMD Implementation

The procedure below instantiates the framework of Section 5 with parameterised actors and mini-batched contexts.

Algorithm 3 is a parameterized minibatch implementation of the framework; Theorem 5.4 applies when its actor updates satisfy the external-regret condition of Assumption 5.3.

B.5 Proof of Theorem 5.4

Proof. We first establish the standard statistical guarantees for our offline estimation phase.

Lemma B.8 (Offline Estimation Guarantees)**.** *Under Assumption 5.1, setting the version space tolerance* $\beta = \frac{136 \log(2|\mathcal{F}|/\delta)}{n}$*, with probability at least* $1 - \delta$*, the true reward resides in the version space* $r \in \mathcal{F}(\beta)$*, and for all* $f \in \mathcal{F}(\beta)$:

$$\mathbb{E}_{s,a\sim d^\rho}\left[(f(s,a)-r(s,a))^2\right] \le 2\beta.$$

Furthermore, the Maximum Likelihood Estimate $\widehat{b} \in \mathcal{B}$ guarantees an expected L_1 error bounded by $\mathcal{E}_{\mathcal{B}} \stackrel{\text{def}}{=} 2\sqrt{2\log(2|\mathcal{B}|/\delta)/n}$ such that:

$$\mathbb{E}_{c\sim\rho_C}\left[\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1\right] \le \mathcal{E}_{\mathcal{B}}.$$

Proof of Lemma B.8. The proof proceeds in two parts, allocating a failure probability of $\delta/2$ to the reward regression and $\delta/2$ to the belief estimation, yielding a global guarantee with probability at least $1-\delta$ via a union bound.

Part 1: Reward estimation. This part establishes $\sup_{f\in\mathcal{F}(\beta)} \mathbb{E}_{s,a\sim d^\rho}[(f-r)^2] \le 2\beta$ on the population behavior measure, which together with Assumption 5.2 will yield the target-distribution bound in the main proof body. A uniform fast-rate ERM bound follows from a Bernstein concentration combined with a union bound over $\mathcal{F}$; realizability and the version-space tolerance β then convert the empirical bound into a true-risk guarantee. No concentrability assumption enters this part.

Let $\mathcal{L}(f) = \mathbb{E}_{s,a\sim d^\rho}[(f(s,a)-r(s,a))^2]$ denote the true expected squared error under the population behavior measure, and $\mathcal{L}_{\mathcal{D}}(f) = \frac{1}{n}\sum_{k=1}^{n}(f(s^{(k)},a^{(k)}) - r^{(k)})^2$ denote the empirical risk. Consider the random variable $X = (f(s,a)-r(s,a))^2$ with $(s,a)\sim d^\rho$. Since functions in $\mathcal{F}$ and the true reward r are bounded in $[-1,1]$, we have $X \in [0,4]$. The expectation is $\mathbb{E}[X] = \mathcal{L}(f)$. The variance is bounded by the uncentered second moment:

$$\mathbb{V}(X) \le \mathbb{E}[X^2] = \mathbb{E}[(f(s,a)-r(s,a))^4] \le 4\mathbb{E}[(f(s,a)-r(s,a))^2] = 4\mathcal{L}(f).$$

Applying Bernstein's inequality for a fixed $f \in \mathcal{F}$:

$$\mathbb{P}\left(\mathcal{L}(f) - \mathcal{L}_{\mathcal{D}}(f) \ge t\right) \le \exp\left(\frac{-nt^2/2}{\mathbb{V}(X) + 4t/3}\right) \le \exp\left(\frac{-nt^2/2}{4\mathcal{L}(f) + 4t/3}\right).$$

Equating the right-hand side to $\frac{\delta}{2|\mathcal{F}|}$ and letting $\iota_{\mathcal{F}} \stackrel{\text{def}}{=} \log(2|\mathcal{F}|/\delta)$, we solve the quadratic $\frac{nt^2}{2} - \frac{4}{3}t\iota_{\mathcal{F}} - 4\mathcal{L}(f)\iota_{\mathcal{F}} = 0$ for its positive root and apply $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$:

$$t = \frac{1}{n}\left[\frac{4}{3}\iota_{\mathcal{F}} + \sqrt{\frac{16}{9}\iota_{\mathcal{F}}^2 + 8n\mathcal{L}(f)\iota_{\mathcal{F}}}\right] \le \frac{8\iota_{\mathcal{F}}}{3n} + \sqrt{\frac{8\mathcal{L}(f)\iota_{\mathcal{F}}}{n}}.$$

Using the AM-GM inequality variant $\sqrt{xy} \le \frac{x}{4} + y$ with $x = \mathcal{L}(f)$ and $y = \frac{8\iota_{\mathcal{F}}}{n}$, we bound the square root term:

$$\sqrt{\mathcal{L}(f)\cdot\frac{8\iota_{\mathcal{F}}}{n}} \le \frac{1}{4}\mathcal{L}(f) + \frac{8\iota_{\mathcal{F}}}{n}.$$

Substituting this back yields the deviation bound:

$$t \le \frac{1}{4}\mathcal{L}(f) + \left(\frac{8}{3}+8\right)\frac{\iota_{\mathcal{F}}}{n} = \frac{1}{4}\mathcal{L}(f) + \frac{32\iota_{\mathcal{F}}}{3n}.$$

By applying a union bound over all $f \in \mathcal{F}$, it holds uniformly with probability at least $1-\delta/2$ that:

$$\begin{aligned}\mathcal{L}(f) - \mathcal{L}_{\mathcal{D}}(f) &\le \frac{1}{4}\mathcal{L}(f) + \frac{32\iota_{\mathcal{F}}}{3n}\\ \frac{3}{4}\mathcal{L}(f) &\le \mathcal{L}_{\mathcal{D}}(f) + \frac{32\iota_{\mathcal{F}}}{3n}\\ \mathcal{L}(f) &\le \frac{4}{3}\mathcal{L}_{\mathcal{D}}(f) + \frac{128\iota_{\mathcal{F}}}{9n}.\end{aligned} \tag{24}$$

Under Assumption 5.1, the true reward satisfies $r \in \mathcal{F}$. Because the environment rewards are deterministic, $r(s^{(k)},a^{(k)}) = r^{(k)}$ identically, meaning the empirical risk of the true reward is exactly zero ($\mathcal{L}_{\mathcal{D}}(r) = 0$). Since $\widehat{r} = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}_{\mathcal{D}}(f)$, it must be that $\mathcal{L}_{\mathcal{D}}(\widehat{r}) \le \mathcal{L}_{\mathcal{D}}(r) = 0$, implying $\mathcal{L}_{\mathcal{D}}(\widehat{r}) = 0$.

The version space is defined as $\mathcal{F}(\beta) = \{f \in \mathcal{F} : \mathcal{L}_\mathcal{D}(f) - \mathcal{L}_\mathcal{D}(\widehat{r}) \leq \beta\}$. Since $\mathcal{L}_\mathcal{D}(r) - \mathcal{L}_\mathcal{D}(\widehat{r}) = 0 \leq \beta$, it trivially holds that $r \in \mathcal{F}(\beta)$ for any $\beta \geq 0$. For any arbitrary $f \in \mathcal{F}(\beta)$, we have $\mathcal{L}_\mathcal{D}(f) \leq \beta$. Substituting this into (24):

$$\mathcal{L}(f) \leq \frac{4}{3}\beta + \frac{128\iota_\mathcal{F}}{9n}.$$

Given the lemma's specification of $\beta = \frac{136\iota_\mathcal{F}}{n}$, we observe that $\frac{128\iota_\mathcal{F}}{9n} \leq \frac{2}{3}\left(\frac{136\iota_\mathcal{F}}{n}\right) = \frac{2}{3}\beta$. Therefore:

$$\mathcal{L}(f) \leq \frac{4}{3}\beta + \frac{2}{3}\beta = 2\beta,$$

which satisfies the required risk guarantee for the reward estimator.

Part 2: Belief estimation. The empirical belief is obtained via MLE (3): $\widehat{b} = \operatorname{argmax}_{b\in\mathcal{B}} \frac{1}{n}\sum_{k=1}^{n} \log b(x^{1,(k)}, x^{2,(k)} \mid c^{(k)})$.

We bound the MLE error through the squared Hellinger distance. Introduce the joint distributions on $\mathcal{C} \times \mathcal{X}_1 \times \mathcal{X}_2$

$$P_b(c,x) \stackrel{\text{def}}{=} \rho_\mathcal{C}(c)\, b(x \mid c), \qquad P_q(c,x) \stackrel{\text{def}}{=} \rho_\mathcal{C}(c)\, q(x \mid c) \quad \text{for } q \in \mathcal{B},$$

and write $H^2(P,Q) \stackrel{\text{def}}{=} 1 - \sum_z \sqrt{P(z)Q(z)}$ for the squared Hellinger distance. The n observations $(c^{(k)}, x^{1,(k)}, x^{2,(k)})$ are i.i.d. draws from P_b, and the MLE (3) maximises the joint likelihood over $\mathcal{B}$ because the factor $\rho_\mathcal{C}(c)$ is common to all candidates.

Step 1: pointwise likelihood-ratio bound. Fix $q \in \mathcal{B}$ and let $L_n(q) = \prod_{k=1}^{n} q(x^{(k)} \mid c^{(k)})$. By Markov's inequality applied to the square-root likelihood ratio, and then independence,

$$\mathbb{P}_b\left(L_n(q) \geq L_n(b)\right) = \mathbb{P}_b\left(\sqrt{L_n(q)/L_n(b)} \geq 1\right) \leq \mathbb{E}_b\left[\sqrt{L_n(q)/L_n(b)}\right] = \left(\sum_z \sqrt{P_b(z)P_q(z)}\right)^n,$$

and since $\sum_z \sqrt{P_b P_q} = 1 - H^2(P_b, P_q) \leq \exp(-H^2(P_b, P_q))$, the right-hand side is at most $\exp\left(-nH^2(P_b, P_q)\right)$.

Step 2: union bound. Under Assumption 5.1 we have $b \in \mathcal{B}$, so the maximiser $\widehat{b}$ satisfies $L_n(\widehat{b}) \geq L_n(b)$ by definition. Applying Step 1 to every $q \in \mathcal{B}$ with $H^2(P_b, P_q) > \log(2|\mathcal{B}|/\delta)/n$ and taking a union bound, with probability at least $1 - \delta/2$ no such candidate can beat b, whence

$$H^2(P_b, P_{\widehat{b}}) \leq \frac{\log(2|\mathcal{B}|/\delta)}{n}.$$

Step 3: conversion to L_1. Because P_b and $P_{\widehat{b}}$ share the public-context marginal $\rho_\mathcal{C}$,

$$\mathbb{E}_{c\sim\rho_\mathcal{C}}\left[\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1\right] = \sum_{c,x} \left|P_b(c,x) - P_{\widehat{b}}(c,x)\right| = \|P_b - P_{\widehat{b}}\|_1.$$

The standard comparison $\|P - Q\|_1 = 2\, d_{\mathrm{TV}}(P,Q) \leq 2\sqrt{2}\, H(P,Q)$ then gives

$$\mathbb{E}_{c\sim\rho_\mathcal{C}}\left[\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1\right] \leq 2\sqrt{2H^2(P_b, P_{\widehat{b}})} \leq 2\sqrt{\frac{2\log(2|\mathcal{B}|/\delta)}{n}} \stackrel{\text{def}}{=} \mathcal{E}_\mathcal{B}.$$

Combining Part 1 and Part 2 via a union bound, both guarantees hold simultaneously with probability at least $1 - \delta$, completing the proof. □

Now we prove Theorem 5.4. We begin by establishing the approximate pessimism of the FA objective L_c relative to the true reward r_c. By Lemma B.8, $r \in \mathcal{F}(\beta)$. Therefore, the constructed critic $\underline{r}(s,a) = \inf_{f\in\mathcal{F}(\beta)} f(s,a)$ satisfies $\underline{r}(s,a) \leq r(s,a)$. The point-wise underestimation error is bounded by the version space diameter:

$$r(s,a) - \underline{r}(s,a) \leq \sup_{f\in\mathcal{F}(\beta)} f(s,a) - \inf_{f\in\mathcal{F}(\beta)} f(s,a) \leq 2 \sup_{f\in\mathcal{F}(\beta)} |f(s,a) - r(s,a)|.$$

Let $\nu_c^\star$ denote the true best-response policy of Player 2 against the learner's iterative strategy $\bar{\pi}_1$, defined as:

$$\nu_c^\star \in \operatorname*{argmin}_{\pi_2 \in \Pi_2} r_c(\bar{\pi}_1, \pi_2).$$

Under this definition, the contextual exploitability gap at context c is $V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star)$.

To bound the exploitability gap of the iterative policy $\bar{\pi}_1$, we first decompose it by adding and subtracting the empirical minimax value $\widehat{V}(c)$ and the empirical performance of the iterative policy:

$$V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star) = \underbrace{V^\star(c) - \widehat{V}(c)}_{\text{Value Error}} + \underbrace{\widehat{V}(c) - \min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2)}_{\text{Optimization Error}} + \underbrace{\min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) - r_c(\bar{\pi}_1, \nu_c^\star)}_{\text{Pessimism Error}}. \tag{25}$$

We bound each term individually. First, for the **Value Error**, recall that $\widehat{V}(c) = \max_{\pi_1} \min_{\pi_2} L_c(\pi_1, \pi_2)$ (19). By the property of the maximum, $\widehat{V}(c) \geq \min_{\pi_2} L_c(\pi_{1,\star}, \pi_2)$. Thus:

$$V^\star(c) - \widehat{V}(c) \leq \min_{\pi_2 \in \Pi_2} r_c(\pi_{1,\star}, \pi_2) - \min_{\pi_2 \in \Pi_2} L_c(\pi_{1,\star}, \pi_2).$$

Second, the **Optimization Error** is bounded by the no-regret property of the iterative solver derived in (21):

$$\widehat{V}(c) - \min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) \leq \epsilon_1 + \epsilon_2.$$

Third, for **Pessimism Error**, we evaluate the consistency error of the iterative policy $\bar{\pi}_1$. By the definition of the best-response $\nu_c^\star = \operatorname{argmin}_{\pi_2} r_c(\bar{\pi}_1, \pi_2)$, it follows that:

$$\min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) - r_c(\bar{\pi}_1, \nu_c^\star) \leq \max_{\pi_2 \in \Pi_2} \left(L_c(\bar{\pi}_1, \pi_2) - r_c(\bar{\pi}_1, \pi_2) \right).$$

For any fixed π_2, we decompose the difference into belief and reward estimation errors by adding and subtracting $\sum_{x,a} b(x \mid c) \underline{r}(s,a) \bar{\pi}_1 \pi_2$:

$$\begin{aligned} L_c(\bar{\pi}_1, \pi_2) - r_c(\bar{\pi}_1, \pi_2) &= \sum_{x,a} (\widehat{b} - b) \underline{r} \bar{\pi}_1 \pi_2 + \sum_{x,a} b(\underline{r} - r) \bar{\pi}_1 \pi_2 \\ &\leq \sum_{x,a} |b - \widehat{b}| \cdot |\underline{r}| \bar{\pi}_1 \pi_2 + 0 \qquad \text{(Since } \underline{r} \leq r \text{ implies } (\underline{r} - r) \leq 0) \\ &\leq \sum_{x} |b(x \mid c) - \widehat{b}(x \mid c)| \underbrace{\sum_{a} \bar{\pi}_1 \pi_2}_{=1} = \|b - \widehat{b}\|_1. \end{aligned}$$

Thus, the consistency error is bounded purely by the belief estimation accuracy:

$$\min_{\pi_2 \in \Pi_2} L_c(\bar{\pi}_1, \pi_2) - r_c(\bar{\pi}_1, \nu_c^\star) \leq \|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1. \tag{26}$$

Using the decomposition from (11), for any policy pair (π_1, π_2) and context c :

$$\begin{aligned} & r_c(\pi_1, \pi_2) - L_c(\pi_1, \pi_2) \\ =& \sum_{x,a} b(r - \underline{r}) \pi_1 \pi_2 + \sum_{x,a} (b - \widehat{b}) \underline{r} \pi_1 \pi_2 \\ \leq& \mathbb{E}_{s,a \sim d^{\pi_1, \pi_2 | c}} \left[2 \sup_{f \in \mathcal{F}(\beta)} |f(s,a) - r(s,a)| \right] + \|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1. \end{aligned} \tag{27}$$

The second term is at most $\|b(\cdot \mid c) - \widehat{b}(\cdot \mid c)\|_1$ because $\|\underline{r}\|_\infty \leq 1$, exactly as in (26). In the first term, $r(s,a) - \underline{r}(s,a) = \sup_{f \in \mathcal{F}(\beta)} (r(s,a) - f(s,a)) \leq \sup_{f \in \mathcal{F}(\beta)} |f(s,a) - r(s,a)|$, which in particular implies the bound with $2 \sup_{f \in \mathcal{F}(\beta)} |f - r|$ in (27).

Substituting the individual bounds for the three terms of (25) back into the contextual decomposition yields:

$$V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star) \le \mathbb{E}_{s,a\sim d^{\pi_{1,\star},\widehat{\pi}_2}|c}\left[2\sup_{f\in\mathcal{F}(\beta)}|f(s,a)-r(s,a)|\right] + 2\|b(\cdot\mid c) - \widehat{b}(\cdot\mid c)\|_1 + \epsilon_{\text{opt}},$$

where $\widehat{\pi}_2 = \operatorname{argmin}_{\pi_2} L_c(\pi_{1,\star}, \pi_2)$. Here $\epsilon_{\text{opt}} \stackrel{\text{def}}{=} \epsilon_1 + \epsilon_2$, and Assumption 5.3 gives the saddle-point optimization error $\epsilon_{\text{opt}} = (R_1(T) + R_2(T))/T$ in (18). For multiplicative weights, Lemma B.7 gives $R_i(T) \le 2\sqrt{T\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})}$, so $\epsilon_{\text{opt}} \le 4\sqrt{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})/T}$ as in (23).

To reach the global exploitability guarantee, we integrate this contextual gap over the public context distribution $c \sim \rho_{\mathcal{C}}$:

$$\begin{aligned}\text{Gap}(\bar{\pi}_1) &= \mathbb{E}_{c\sim\rho_{\mathcal{C}}}\left[V^\star(c) - r_c(\bar{\pi}_1, \nu_c^\star)\right] \\ &\le \mathbb{E}_{c,s,a\sim d^{\pi_{1,\star},\widehat{\pi}_2}}\left[2\sup_{f\in\mathcal{F}(\beta)}|f(s,a)-r(s,a)|\right] + 2\mathbb{E}_{c\sim\rho_{\mathcal{C}}}\|b-\widehat{b}\|_1 + \epsilon_{\text{opt}}.\end{aligned}$$

We now apply our concentration results. For the belief error, Part 2 of Lemma B.8 directly gives

$$2\mathbb{E}_{c\sim\rho_{\mathcal{C}}}\|b-\widehat{b}\|_1 \le 2\mathcal{E}_{\mathcal{B}} = 4\sqrt{\frac{2\log(2|\mathcal{B}|/\delta)}{n}},$$

since Part 2 bounds the L_1 error directly via the Hellinger distance.

For the reward error, we apply Cauchy–Schwarz directly to the nonnegative random variable $\sup_{f\in\mathcal{F}(\beta)}(f-r)^2$, keeping the supremum *inside* the expectation; the uniform-ratio assumption (Assumption 5.2) then transfers the resulting expected squared error from $d^{\pi_{1,\star},\widehat{\pi}_2}$ to the behavior distribution d^ρ in a single step, after which Lemma B.8 (Part 1) closes the chain:

$$\begin{aligned}\mathbb{E}_{c,s,a\sim d^{\pi_{1,\star},\widehat{\pi}_2}}\left[2\sup_{f\in\mathcal{F}(\beta)}|f(s,a)-r(s,a)|\right] &\le 2\sqrt{\mathbb{E}_{d^{\pi_{1,\star},\widehat{\pi}_2}}\left[\sup_{f\in\mathcal{F}(\beta)}(f(s,a)-r(s,a))^2\right]} \\ &\qquad\qquad\text{(Cauchy–Schwarz / Jensen)} \\ &\le 2\sqrt{\mathscr{C}_{\mathcal{F}}\cdot\sup_{f\in\mathcal{F}(\beta)}\mathbb{E}_{s,a\sim d^\rho}[(f(s,a)-r(s,a))^2]} \\ &\qquad\qquad\text{(Assumption 5.2)} \\ &\le 2\sqrt{2\,\mathscr{C}_{\mathcal{F}}\,\beta}. \qquad\text{(Lemma B.8, Part 1)}\end{aligned}$$

Summing these integrated components and substituting $\beta = 136\log(2|\mathcal{F}|/\delta)/n$ from Lemma B.8 (Part 1) and $\mathcal{E}_{\mathcal{B}} = 2\sqrt{2\log(2|\mathcal{B}|/\delta)/n}$ completes the proof of Theorem 5.4:

$$\text{Gap}(\bar{\pi}_1) \le 2\sqrt{\frac{272\,\mathscr{C}_{\mathcal{F}}\log(2|\mathcal{F}|/\delta)}{n}} + 4\sqrt{\frac{2\log(2|\mathcal{B}|/\delta)}{n}} + \frac{R_1(T)+R_2(T)}{T}.$$

With multiplicative weights the optimization term is at most $4\sqrt{\log(\max\{|\mathcal{A}_1|, |\mathcal{A}_2|\})/T}$.

□