PESSIMISM-FREE OFFLINE LEARNING IN GENERAL-SUM GAMES VIA KL REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline multi-agent reinforcement learning in general-sum settings is challenged by the distribution shift between logged datasets and target equilibrium policies. While standard methods rely on manual pessimistic penalties, we demonstrate that KL regularization suffices to stabilize learning and achieve equilibrium recovery. We propose General-sum Anchored Nash Equilibrium (GANE), which recovers regularized Nash equilibria at an accelerated statistical rate of $\tilde{\mathcal{O}}(1/n)$. For computational tractability, we develop General-sum Anchored Mirror Descent (GAMD), an iterative algorithm converging to a Coarse Correlated Equilibrium at the standard rate of $\tilde{\mathcal{O}}(1/\sqrt{n}+1/T)$. These results establish KL regularization as a standalone mechanism for pessimism-free offline learning that achieves equivalent or accelerated rates in multi-player general-sum games.

1 INTRODUCTION

Offline reinforcement learning (RL) provides a framework for developing decision-making policies from pre-collected datasets, making it suitable for applications where active exploration is restricted by safety or financial costs (Levine et al., 2020). While much of the offline literature has historically addressed single-agent optimization, many strategic environments involve interactions among multiple agents with distinct objectives. General-sum games offer a flexible model for these scenarios, capturing strategic complexities in economic markets, multi-agent coordination, and the alignment of large language models through collective human preferences (Ouyang et al., 2022; Ye et al., 2024). In this offline multi-agent setting, the learning objective is to recover an equilibrium policy using only fixed historical data.

A central difficulty in offline general-sum games is distribution shift, which occurs when the learned strategy explores regions of the state-action space that are poorly represented in the logged dataset (Fujimoto et al., 2019; Zhang et al., 2023). Standard methodologies typically address this shift through the principle of explicit pessimism (Jin et al., 2021; Zhang et al., 2023). These approaches incorporate manually designed lower confidence bounds (LCB) or penalty terms to suppress the estimated values of actions that lack sufficient offline data support (Cui & Du, 2022a; Zhong et al., 2022). While theoretically sound, these pessimistic mechanisms often require the construction of complex uncertainty quantifiers over joint action spaces and intensive hyperparameter tuning to balance conservatism with performance (Cui & Du, 2022a; Zhang et al., 2023).

Recently, KL regularization with respect to a fixed reference policy has become a common objective for stabilizing multi-agent learning and enforcing behavioral constraints (Ouyang et al., 2022; Ye et al., 2024). While this framework has gained significant prominence in single-agent settings—for instance, in language model alignment through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022)—its application to multi-player general-sum settings remains relatively scarce. KL-regularized game objectives are motivated by settings such as multi-agent negotiation, where optimizing pure game payoff when fine-tuning language model agents can lead to language drift (Lewis et al., 2017). Regularizing relative to a reference language model penalizes departures from its behavior, making the tradeoff between strategic payoff and reference behavior part of the objective (Ouyang et al., 2022; Munos et al., 2024).

In existing literature that utilizes KL regularization within game-theoretic contexts, the regularization is typically treated solely as an anchor to a reference distribution, while distribution shift is handled

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.

by another separate, explicit pessimism mechanism such as lower confidence bounds (Ye et al., 2024). However, the capacity of KL regularization to act as a standalone, implicit pessimism mechanism that independently stabilizes learning against distribution shift in multi-player general-sum games remains unestablished.

In this work, we establish that KL regularization serves as a standalone, pessimism-free mechanism to stabilize offline learning in general-sum games. To achieve this, we leverage the reference-anchored data coverage framework (Chen & Zhang, 2026), which bypasses the conceptual circularity inherent in prior works by anchoring coverage requirements to a known reference policy rather than an unknown equilibrium. We first introduce General-sum Anchored Nash Equilibrium (GANE), a theoretical framework that recovers regularized Nash equilibria at an accelerated statistical rate of $\tilde{\mathcal{O}}(1/n)$. Our analysis reveals that the independent product structure of Nash policies enables the exact cancellation of first-order estimation errors, allowing GANE to bypass the standard $\tilde{\mathcal{O}}(1/\sqrt{n})$ bottleneck encountered in unregularized offline games (Cui & Du, 2022a; Zhang et al., 2023). Recognizing that computing a Nash equilibrium is often computationally intractable, we further develop General-sum Anchored Mirror Descent (GAMD), a decentralized iterative algorithm. We prove that GAMD recovers a Coarse Correlated Equilibrium at the standard rate of $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/T)$ for n samples and T iterations. Our primary contributions are summarized as follows:

1. **Pessimism-Free Equilibrium Recovery**: We establish that KL regularization serves as an effective alternative to explicit pessimism, allowing for equilibrium recovery in general-sum settings without the overhead of tuning confidence-based bonuses.
2. **Accelerated Rates for Nash Equilibria**: We prove that GANE achieves an $\tilde{\mathcal{O}}(1/n)$ statistical rate. This result demonstrates that the product structure of Nash equilibria can be leveraged to achieve higher statistical efficiency than previously established for unregularized general-sum offline learning.
3. **Tractable Iterative Learning**: We show that GAMD provides a computationally efficient path to recovering Coarse Correlated Equilibria, achieving the standard $\tilde{\mathcal{O}}(1/\sqrt{n})$ statistical rate through simple, decentralized updates without relying on explicit pessimism.

Organization. The remainder of this paper is organized as follows. In Section 2, we formalize the m-player general-sum game, define the KL-regularized value functions, and establish the offline learning model. Section 3 reviews the landscape of concentrability assumptions in multi-agent systems and the reference-anchored coverage requirements. Section 4 introduces the GANE algorithm and presents the accelerated $\tilde{\mathcal{O}}(1/n)$ statistical rate for Nash equilibria. In Section 5, we propose the iterative GAMD algorithm and characterize its optimization and sample complexity guarantees for recovering Coarse Correlated Equilibria. We review related work in Section 6 and conclude the paper in Section 7.

Paper	NE Rate	CCE Rate	Optimization (T)	Pessimism
Cui & Du (2022b)	$\tilde{\mathcal{O}}(1/\sqrt{n})$	—	Oracle	Yes
Zhang et al. (2023)	$\tilde{\mathcal{O}}(1/\sqrt{n})$	$\tilde{\mathcal{O}}(1/\sqrt{n})$	Oracle	Yes
This Work	$\tilde{\mathcal{O}}(\mathbf{1/n})$	$\tilde{\mathcal{O}}(\mathbf{1/\sqrt{n}})$	$\tilde{\mathcal{O}}(\mathbf{1/T})$	**No**

Table 1: Comparison of offline learning guarantees in multi-player general-sum games, with baselines on unregularized games. We report statistical rates in terms of samples n and optimization convergence in terms of iterations T. Our work is the first to establish a fast $\tilde{\mathcal{O}}(1/n)$ rate for Nash Equilibria and a computationally tractable solution for CCE without explicit pessimism. The "Oracle" labels denote a reliance on either black-box global optimization solvers (Cui & Du, 2022b) or global optimization oracles over general function classes (Zhang et al., 2023).

2 Problem Formulation

We consider an m-player general-sum contextual bandit game defined by the tuple $\mathcal{M} = (\mathcal{N}, \mathcal{X}, \{\mathcal{A}_i\}_{i=1}^m, \{r_i^\star\}_{i=1}^m, \rho)$. Here, $\mathcal{N} = [m] := \{1, \ldots, m\}$ denotes the set of players, $\mathcal{X}$ is the context space, and $\mathcal{A}_i$ is the finite action space for player i. In each round, a context $x \sim \rho$

is sampled. All players simultaneously select actions $a_i \in \mathcal{A}_i$, forming a joint action profile $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A} \coloneqq \prod_{i=1}^m \mathcal{A}_i$. Each player i receives a deterministic reward $r_i^\star(x, \boldsymbol{a}) \in [0, 1]$.

A joint policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ maps contexts to a distribution over joint actions. We denote the marginalized policy for player i as $\pi_i(\cdot \mid x)$ and the joint policy of all other players as $\pi_{-i}(\cdot \mid x)$.

2.1 KL-Regularized Objectives

We study objectives that balance game payoff with KL regularization relative to fixed reference policies $\pi^{\text{ref}} = (\pi_1^{\text{ref}}, \ldots, \pi_m^{\text{ref}})$. At each context, policies and deviations in these objectives are supported on their reference policies, since mass outside their supports incurs infinite KL divergence; see also Xiong et al. (2023); Zhang et al. (2025b). For a regularization parameter $\eta > 0$, the Q-function for player i is

$$Q_i^\pi(x, \boldsymbol{a}) \coloneqq r_i^\star(x, \boldsymbol{a}).$$

The KL-regularized value function $V_i^\pi(x)$ is defined as:

$$V_i^\pi(x) = \mathbb{E}_{\boldsymbol{a} \sim \pi(\cdot|x)}[Q_i^\pi(x, \boldsymbol{a})] - \eta^{-1}\mathrm{KL}\big(\pi_i(\cdot|x) \,\|\, \pi_i^{\text{ref}}(\cdot|x)\big). \tag{1}$$

Each player i seeks to maximize their own regularized value function. We define the best-response value against an opponent profile π_{-i} as

$$V_i^{\dagger,\pi_{-i}}(x) \coloneqq \max_{\pi_i'} V_i^{\pi_i', \pi_{-i}}(x).$$

We also denote $C_{\text{shift}} \coloneqq \exp(\eta m)$ as the joint anchoring constant measuring the density shift relative to the reference policy.

2.2 Target Equilibria

We analyze two equilibrium concepts in the regularized m-player game. Conceptually, an equilibrium represents a stable state where no player can unilaterally improve their own regularized value by changing their strategy, assuming all other players' strategies remain fixed. The primary distinction between these concepts lies in the structural constraints placed on the joint policy and the resulting ability for players to coordinate.

Regularized Nash Equilibrium (NE). A joint policy $\pi^\star$ is a regularized NE if it is a **product policy**, factorizing as $\pi^\star(\boldsymbol{a} \mid x) = \prod_{i=1}^m \pi_i^\star(a_i \mid x)$. This captures scenarios where players act completely independently. The equilibrium condition implies that no player i can increase their value by switching to a different independent strategy π_i', given that opponents follow their respective marginal strategies $\pi_{-i}^\star$. For a learned product policy $\hat{\pi} = \prod_i \hat{\pi}_i$, we evaluate the NE Total Exploitability:

$$\mathrm{Gap}_{\text{NE}}(\hat{\pi}) \coloneqq \sum_{i=1}^m \mathbb{E}_{x \sim \rho}\left[V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}_1 \times \cdots \times \hat{\pi}_m}(x)\right].$$

KL regularization induces a reference-weighted softmax best response. At each context, a regularized NE is a fixed point of these responses; with uniform reference policies on the action spaces, it reduces exactly to the canonical logit quantal response equilibrium (QRE) with precision η (McKelvey & Palfrey, 1995). This connects our target to a standard model of bounded rationality in behavioral game theory.

Regularized Coarse Correlated Equilibrium (CCE). A regularized CCE is a broader concept that allows the joint policy $\pi^\star$ to be **any distribution** in the simplex $\Delta(\mathcal{A})$, potentially incorporating correlations between players (e.g., through a shared latent signal). The stability condition here implies that a player cannot improve their expected return by choosing to deviate from the suggested joint distribution before observing their specific action. While every Nash Equilibrium is a CCE, a CCE allows for complex coordination that is forbidden in the NE setting. For a learned (possibly correlated) joint policy $\bar{\pi}$, we evaluate the CCE Total Exploitability:

$$\mathrm{Gap}_{\text{CCE}}(\bar{\pi}) \coloneqq \sum_{i=1}^m \mathbb{E}_{x \sim \rho}\left[\max\left\{0, V_i^{\dagger,\bar{\pi}_{-i}}(x) - V_i^{\bar{\pi}}(x)\right\}\right].$$

While we focus on NE and CCE in this work, we consider the extension to Correlated Equilibrium (CE) a promising direction for future research. Our current analysis serves as a foundational demonstration of how KL regularization alone—without the need for explicit pessimism—suffices to establish both a fast $\tilde{\mathcal{O}}(1/n)$ rate for Nash equilibria and a computationally tractable solution for CCE. By characterizing these two distinct regimes, we lay the groundwork for a broader regularized framework across the entire spectrum of game-theoretic equilibria.

2.3 Offline Learning Model

We consider the offline setting where the learner has access to a static dataset $\mathcal{D} = \{(x_\tau, \boldsymbol{a}_\tau, \boldsymbol{r}_\tau)\}_{\tau=1}^n$ consisting of n tuples. The context-action pairs $(x_\tau, \boldsymbol{a}_\tau)$ are sampled independently from a common distribution μ. The observed rewards $\boldsymbol{r}_\tau = (r_{\tau,1}, \ldots, r_{\tau,m})$ are noisy realizations of the true rewards $r_{\tau,i} = r_i^\star(x_\tau, \boldsymbol{a}_\tau) + \xi_{\tau,i}$, where $\xi_{\tau,i}$ is conditionally independent across samples, zero-mean, and 1-sub-Gaussian given the context-action pairs. We assume access to a fixed finite function class $\mathcal{Q}_i \subset (\mathcal{X} \times \mathcal{A} \to [0,1])$ for each player. Following standard literature (Xie & Jiang, 2021; Zhang et al., 2026), we assume *realizability*, i.e., $r_i^\star \in \mathcal{Q}_i$ for all $i \in \mathcal{N}$. For any estimated function $\hat{Q}_i \in \mathcal{Q}_i$, we define the point-wise regression error as $\mathcal{Z}_i(x, \boldsymbol{a}) := \hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a})$. We write $|\mathcal{Q}| = \max_{i \in [m]} |\mathcal{Q}_i|$.

3 Unilateral Data Coverage

Establishing sample-efficient recovery in offline multi-agent systems requires that the historical dataset $\mathcal{D}$ provides sufficient coverage of the state-action regions relevant to the target objective. In this section, we contrast the traditional unilateral concentrability assumption with the reference-anchored framework used in this work.

3.1 Standard Unilateral Concentrability and the Circularity Problem

The principle of *unilateral concentrability* is often employed to bypass the exponential scaling of joint action spaces in multi-agent environments (Cui & Du, 2022b;a). This principle requires the dataset only to support scenarios where a single agent deviates from a target strategy while all others remain stationary.

Assumption 3.1 (Classical Unilateral Concentrability (Cui & Du, 2022b))**.** Let $\pi^* = (\pi_1^*, \ldots, \pi_m^*)$ be a target equilibrium policy. Let $\Pi_{\text{uni}}(\pi^*) := \bigcup_{i=1}^m \left(\Pi_i \times \{\pi_{-i}^*\}\right)$ be the set of unilateral deviations from π^*. There exists a constant $C_{\text{uni}}^* \geq 1$ such that for any $\pi' \in \Pi_{\text{uni}}(\pi^*)$:

$$\left\| \frac{\rho(x)\pi'(\boldsymbol{a} \mid x)}{\mu(x, \boldsymbol{a})} \right\|_\infty \leq C_{\text{uni}}^*.$$

While this assumption provides favorable scaling ($\sum |\mathcal{A}_i|$ rather than $\prod |\mathcal{A}_i|$), it suffers from the **Curse of the Unknown Optimum**. Because C_{uni}^* is anchored to an unknown optimal equilibrium π^*, the assumption is conceptually circular: one must already know the target equilibrium to determine if the data is sufficient to find it. This makes the condition difficult to verify or estimate in practical strategic settings.

3.2 The Reference-Anchored Unilateral Concentrability Framework

To resolve the circularity problem, we utilize the **Reference-Anchored** coverage framework recently proposed in Chen & Zhang (2026). This condition shifts the theoretical burden from an unknown optimum to a fixed and known reference policy π^{ref}.

Assumption 3.2 (Reference-Anchored Unilateral Concentrability (Chen & Zhang, 2026))**.** Let $\pi^{\text{ref}} = (\pi_1^{\text{ref}}, \ldots, \pi_m^{\text{ref}})$ be the fixed reference joint policy. We define the set of *reference-anchored unilateral deviation policies* $\Pi_{\text{ref}-\text{uni}}$ as the union of profiles where one player deviates while all others adhere strictly to the reference policy:

$$\Pi_{\text{ref}-\text{uni}} := \bigcup_{i=1}^m \left(\Pi_i \times \{\pi_{-i}^{\text{ref}}\}\right).$$

Algorithm 1 General-sum Anchored Nash Equilibrium (GANE)

1: **Input:** Offline dataset $\mathcal{D}$, reference policies $\pi^{\mathrm{ref}} = (\pi_1^{\mathrm{ref}}, \ldots, \pi_m^{\mathrm{ref}})$, regularization parameter $\eta > 0$, function classes $\{\mathcal{Q}_i\}_{i=1}^m$.
2: **for** player $i = 1, \ldots, m$ **do**
3: **Empirical Q-Update:** Estimate the Q-function via least-squares regression:
4: $\hat{Q}_i \leftarrow \mathrm{argmin}_{f \in \mathcal{Q}_i} \sum_{\tau=1}^n \left(f(x_\tau, \boldsymbol{a}_\tau) - r_{\tau,i}\right)^2$.
5: **end for**
6: **Equilibrium Computation:** For every context $x \in \mathcal{X}$, find a joint product policy $\hat{\pi}(\boldsymbol{a} \mid x) = \prod_{i=1}^m \hat{\pi}_i(a_i \mid x)$ that is a Regularized Nash Equilibrium of the empirical game $\{\hat{Q}_i(x, \cdot)\}_{i=1}^m$. Namely, for all $i \in [m]$ and all $\pi_i' \in \Delta(\mathcal{A}_i)$:
7:
$$\mathbb{E}_{\boldsymbol{a} \sim \hat{\pi}(\cdot|x)}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1}\mathrm{KL}(\hat{\pi}_i(\cdot \mid x) \,\|\, \pi_i^{\mathrm{ref}}(\cdot \mid x)) \geq \mathbb{E}_{a_i \sim \pi_i'(\cdot|x), \boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}(\cdot|x)}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1}\mathrm{KL}(\pi_i'(\cdot \mid x) \,\|\, \pi_i^{\mathrm{ref}}(\cdot \mid x)).$$
8: **for** player $i = 1, \ldots, m$ **do**
9: **Value Update:** Update the empirical value function for Player i:
10: $\hat{V}_i(x) \leftarrow \mathbb{E}_{\boldsymbol{a} \sim \hat{\pi}(\cdot|x)}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1}\mathrm{KL}(\hat{\pi}_i(\cdot \mid x) \,\|\, \pi_i^{\mathrm{ref}}(\cdot \mid x))$.
11: **end for**
12: **Output:** The estimated regularized NE joint policy $\hat{\pi}$.

Let μ denote the joint distribution of the offline dataset $\mathcal{D}$. We assume there exists a constant $C_{\mathrm{uni}} \geq 1$ such that for any $\pi' \in \Pi_{\mathrm{ref-uni}}$:

$$\left\| \frac{\rho(x)\pi'(\boldsymbol{a} \mid x)}{\mu(x, \boldsymbol{a})} \right\|_\infty \leq C_{\mathrm{uni}}.$$

By anchoring to π^{ref}, this framework provides two critical advantages: (i) A known coverage anchor, whereas standard unilateral concentrability is anchored to an unknown equilibrium; and (ii) Pessimism-Free Stability, as anchoring to the reference policy via KL regularization allows the algorithm to handle distribution shift without explicit bonuses.

4 Pessimism-Free Nash Equilibria

Building on the regularized objective (1), we investigate whether anchoring to a reference policy can independently mitigate the distribution shift inherent in offline multi-agent settings. We first introduce General-sum Anchored Nash Equilibrium (GANE), an algorithmic framework that leverages KL regularization to stabilize learning on the offline contextual bandit without the overhead of explicit pessimistic bonuses. GANE provides the foundational basis for our analysis, illustrating how the geometric properties of the regularizer interact with the structural properties of Nash equilibria.

GANE operates via empirical risk minimization over the static offline dataset. For every player $i \in [m]$, the algorithm first constructs an empirical estimate of the reward Q-function, $\hat{Q}_i$, by solving a least-squares regression problem over the dataset $\mathcal{D}$ using the function class $\mathcal{Q}_i$. Given these empirical Q-functions, GANE assumes access to a stage-game equilibrium oracle that outputs a joint product policy $\hat{\pi}(\boldsymbol{a} \mid x) = \prod_{i=1}^m \hat{\pi}_i(a_i \mid x)$ forming an exact Regularized NE for the empirical game. Formally, $\hat{\pi}$ must satisfy the condition that no player i can unilaterally improve their regularized value by deviating to any arbitrary policy π_i', assuming the other players marginalize over $\hat{\pi}_{-i}$.

4.1 Error Decomposition and Best-Response Analysis

To analyze the suboptimality of the learned policy $\hat{\pi}$, we decouple the optimization error from the statistical error induced by the finite offline dataset by pivoting through the empirical value functions $\hat{V}$.

Lemma 4.1 (Gap Decomposition). *For any learned joint policy $\hat{\pi}$ and any player $i \in [m]$, the unilateral exploitability gap can be exactly decomposed as:*

$$V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x) = \underbrace{\left(\hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\hat{\pi}}(x)\right)}_{\textit{Term I: Optimization Gap}} + \underbrace{\left(V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)\right)}_{\textit{Term II: Best-Response Error}} + \underbrace{\left(\hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x)\right)}_{\textit{Term III: On-Policy Error}}.$$

The derivation follows algebraically by adding and subtracting the estimated values $\hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)$ and $\hat{V}_i^{\hat{\pi}}(x)$ to the true unilateral exploitability gap of Player i. In the GANE algorithm, since $\hat{\pi}$ is an exact Regularized NE of the empirical game, Term I vanishes identically. To characterize the evaluation errors (Terms II and III), we define the pointwise regression error as

$$\mathcal{Z}_i(x, \boldsymbol{a}) \coloneqq \hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a}).$$

4.2 Statistical Guarantees for GANE

Equipped with these analytical tools and the reference-anchored coverage condition (Assumption 3.2), we establish the primary statistical guarantee for GANE.

Theorem 4.2. *Under Assumption 3.2, and assuming the reward functions are realizable within $\{\mathcal{Q}_i\}_{i=1}^m$, the estimated regularized NE joint policy $\hat{\pi}$ returned by GANE achieves a Total Exploitability bounded by:*

$$\mathrm{Gap}_{\mathrm{NE}}(\hat{\pi}) \le \widetilde{\mathcal{O}}\left(\frac{m\eta C_{\mathrm{uni}} \exp(\eta m) \log(m|\mathcal{Q}|/\delta)}{n}\right).$$

The bound holds with probability at least $1-\delta$ for any $\delta \in (0,1)$ under the sampling model in Section 2.3.

4.3 Proof Sketch of Theorem 4.2

The accelerated $\widetilde{\mathcal{O}}(1/n)$ statistical rate is made possible by a unique symmetry in evaluating product policies. We expand the mechanism below; the detailed proof is provided in Appendix B.1.

Term III evaluation. By expanding the value definitions, the on-policy evaluation error (Term III) evaluates exactly to the expected regression error under the joint policy:

$$\text{Term III}_i(x) = \hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x) = \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)}\left[\mathcal{Z}_i(x, \boldsymbol{a})\right].$$

Term II evaluation. To bound the best-response error (Term II), we leverage the 1-smoothness of the log-partition function associated with the KL-regularizer (see Appendix B.1, Step 1). The difference between the true and empirical best-response values is bounded by the first-order Q-value difference plus a second-order penalty. As derived in Equation (2), this yields:

$$\text{Term II}_i(x) \le \mathbb{E}_{a_i\sim\hat{\pi}_i^\dagger(\cdot|x), \boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}(\cdot|x)}\left[-\mathcal{Z}_i(x, \boldsymbol{a})\right] + \frac{\eta}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2,$$

where $\bar{Q}_i(x, a_i) = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}}[\hat{Q}_i(x, a_i, \boldsymbol{a}_{-i})]$ is the marginalized empirical Q-function.

The Cancellation Mechanism. Crucially, because $\hat{\pi}$ is the empirical equilibrium, the learned marginalized policy $\hat{\pi}_i$ is identically the empirical best response $\hat{\pi}_i^\dagger$. Furthermore, since $\hat{\pi}$ is a product policy, its evaluation distribution matches the trajectory of the best-response error. By summing the two evaluation errors, the first-order linear terms exactly cancel out:

$$\begin{aligned}\text{Term II}_i(x) + \text{Term III}_i(x) &\le \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}}[-\mathcal{Z}_i] + \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}}[+\mathcal{Z}_i] + \frac{\eta}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2 \\ &= \frac{\eta}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2.\end{aligned}$$

This cancellation removes the standard $\tilde{\mathcal{O}}(1/\sqrt{n})$ statistical bottleneck. The remaining suboptimality is bounded strictly by second-order squared error terms. These residuals are controlled by the strong convexity of the regularized objective, which, under the coverage of Assumption 3.2, yields the fast $\widetilde{\mathcal{O}}(1/n)$ rate previously unestablished in general-sum settings.

Algorithm 2 General-sum Anchored Mirror Descent (GAMD)

Input: Offline dataset $\mathcal{D}$, reference policies $\pi^{\text{ref}} = (\pi_1^{\text{ref}}, \dots, \pi_m^{\text{ref}})$, regularization parameter $\eta > 0$, function classes $\{\mathcal{Q}_i\}_{i=1}^m$, number of iterations T.
for player $i = 1, \dots, m$ **do**
  Empirical Q-Update: Estimate the Q-function via least-squares regression:
  $\hat{Q}_i \leftarrow \operatorname{argmin}_{f \in \mathcal{Q}_i} \sum_{\tau=1}^n \left(f(x_\tau, \boldsymbol{a}_\tau) - r_{\tau,i}\right)^2$.
end for
Initialize Policies: For all $i \in [m]$, initialize $\pi_i^{(1)}(\cdot \mid x) = \pi_i^{\text{ref}}(\cdot \mid x)$ for all $x \in \mathcal{X}$.
Self-Play Inner Loop:
for $t = 1, \dots, T$ **do**
  for player $i = 1, \dots, m$ **do**
    Compute Marginalized Q-value: Evaluate the expected Q-value given opponents' current policies:
    $\bar{Q}_i^{(t)}(x, a_i) = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}^{(t)}(\cdot|x)}\left[\hat{Q}_i(x, a_i, \boldsymbol{a}_{-i})\right]$.
    Mirror Descent Update: Update Player i's policy via the closed-form exponentiated gradient:

$$\pi_i^{(t+1)}(a_i \mid x) \propto \left(\pi_i^{\text{ref}}(a_i \mid x)\right)^{\frac{1}{t}} \left(\pi_i^{(t)}(a_i \mid x)\right)^{\frac{t-1}{t}} \exp\left(\frac{\eta}{t} \bar{Q}_i^{(t)}(x, a_i)\right).$$

  end for
end for
Time-Averaged Joint Policy: Define the output policy as the uniform mixture over T iterations:
$\bar{\pi}(\boldsymbol{a} \mid x) = \frac{1}{T} \sum_{t=1}^T \prod_{i=1}^m \pi_i^{(t)}(a_i \mid x)$.
for player $i = 1, \dots, m$ **do**
  Value Update: Update the empirical value function using the time-averaged policy:
  $\hat{V}_i(x) \leftarrow \mathbb{E}_{\boldsymbol{a} \sim \bar{\pi}(\cdot|x)}\left[\hat{Q}_i(x, \boldsymbol{a})\right] - \eta^{-1} \mathrm{KL}\big(\bar{\pi}_i(\cdot \mid x) \,\|\, \pi_i^{\text{ref}}(\cdot \mid x)\big)$.
end for
Output: The learned time-averaged joint policy $\bar{\pi}$.

5 Pessimism-Free Coarse Correlated Equilibria

While GANE establishes a fast statistical rate, Nash equilibrium computation is PPAD-complete (Daskalakis et al., 2009; Chen et al., 2009), making the recovery of an NE computationally intractable. This motivates targeting Coarse Correlated Equilibrium (CCE), which serves as a more tractable objective efficiently reachable via independent learning dynamics.

To provide a scalable alternative, we propose General-sum Anchored Mirror Descent (GAMD). GAMD bypasses the need for an exact equilibrium oracle. Instead, for each context evaluated, it simulates a repeated game for T iterations. In each iteration $t \in [T]$, every player independently updates their policy using a KL-regularized Mirror Descent step in response to the marginalized actions of the other players. This iterative sequence converges to a Coarse Correlated Equilibrium (CCE) of the empirical game.

5.1 Theoretical Guarantees of GAMD

To formalize the convergence of GAMD, we first bound the optimization error of the inner loop. Because the KL-divergence penalty transforms the update into a strongly concave objective, applying Mirror Descent yields an accelerated optimization rate. We specialize the standard convergence results of Online Convex Optimization (Shalev-Shwartz, 2025) to our setting of regularized payoff sequences, where the optimization is anchored to the fixed reference policy π^{ref}. We formalize it in the following lemma (proof provided in Appendix B.2):

Lemma 5.1 (OMD External Regret)**.** *Let $f^{(t)}(\pi) = \langle \pi, Q^{(t)} \rangle - \eta^{-1} \mathrm{KL}(\pi \,\|\, \pi^{\text{ref}})$ be a sequence of objectives where $\|Q^{(t)}\|_\infty \leq 1$. Applying KL-regularized Online Mirror Descent with the decaying*

stepsize schedule $\gamma_t = 1/t$ as in Algorithm 2 yields an average external regret bounded by:

$$\max_{\pi \in \Delta(\mathcal{A})} \frac{1}{T} \sum_{t=1}^{T} f^{(t)}(\pi) - \frac{1}{T} \sum_{t=1}^{T} f^{(t)}(\pi^{(t)}) \leq \mathcal{O}\left(\frac{\eta(1+\log T)}{T}\right).$$

By combining this $\widetilde{\mathcal{O}}(1/T)$ optimization rate with the statistical error induced by finite offline samples, we establish the final suboptimality guarantee for GAMD.

Theorem 5.2 (Sample Complexity of GAMD)**.** *Under Assumption 3.2, and assuming the reward functions are realizable within $\{\mathcal{Q}_i\}_{i=1}^m$, the time-averaged correlated joint policy $\bar{\pi}$ returned by Algorithm 2 achieves a CCE Total Exploitability bounded by:*

$$\mathrm{Gap}_{\mathrm{CCE}}(\bar{\pi}) \leq \widetilde{\mathcal{O}}\left(m\left[\frac{\eta(1+\log T)}{T} + \sqrt{\frac{C_{\mathrm{shift}} C_{\mathrm{uni}} \log(m|\mathcal{Q}|/\delta)}{n}} + \frac{\eta C_{\mathrm{shift}} C_{\mathrm{uni}} \log(m|\mathcal{Q}|/\delta)}{n}\right]\right).$$

The bound holds with probability at least $1-\delta$ for any $\delta \in (0,1)$ under the sampling model in Section 2.3. Consequently, setting the number of iterations to $T \geq \sqrt{n}$ gives a final statistical rate of $\widetilde{\mathcal{O}}(1/\sqrt{n})$.

5.2 Proof Sketch of Theorem 5.2

Using the decomposition from Lemma 4.1, Term I is bounded above by the average external regret of the mirror descent updates across T iterations. By standard Online Convex Optimization, this optimization error decays at $\widetilde{\mathcal{O}}(m\eta/T)$ (Lemma 5.1).

For the statistical errors (Terms II and III), the analysis of GAMD faces a distribution mismatch. The recovered time-averaged policy $\bar{\pi}$ is a *correlated* joint mixture, so the independent product distribution need not equal the actual joint policy: $\bar{\pi}_i^\dagger \times \bar{\pi}_{-i} \neq \bar{\pi}$. Consequently, the linear regression errors $-\mathcal{Z}_i$ (from the best response) and $+\mathcal{Z}_i$ (from on-policy evaluation) are integrated over potentially different distributions and do not algebraically cancel in general:

$$\begin{aligned}&\text{Term II}_i(x) + \text{Term III}_i(x)\\ &\leq \mathbb{E}_{\boldsymbol{a}\sim(\bar{\pi}_i^\dagger, \bar{\pi}_{-i})}[-\mathcal{Z}_i(x,\boldsymbol{a})] + \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}}[\mathcal{Z}_i(x,\boldsymbol{a})]\\ &\quad + \frac{\eta}{2}\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2.\end{aligned}$$

Without cancellation, we bound the absolute magnitude of these linear expectations using Cauchy-Schwarz, $\mathbb{E}[|\mathcal{Z}_i|] \leq \sqrt{\mathbb{E}[\mathcal{Z}_i^2]}$. This relates the linear terms back to the $\widetilde{\mathcal{O}}(1/n)$ population squared regression error, yielding a $\widetilde{\mathcal{O}}(1/\sqrt{n})$ bound. Thus, this analysis bounds the statistical evaluation error at the rate of $\widetilde{\mathcal{O}}(1/\sqrt{n})$. Setting $T \geq \sqrt{n}$ balances the optimization and statistical errors. The complete proof is in Appendix B.3.

6 Related Work

Equilibrium Recovery in Multi-Agent Games. The transition from single-agent RL to multi-agent environments necessitates a shift from value optimality to equilibrium stability, targeting solution concepts such as Nash Equilibrium and Coarse Correlated Equilibrium (CCE) (Roughgarden, 2016; Cui & Du, 2022b;a). To recover these equilibria from fixed datasets, existing methodologies largely adapt the principle of pessimism to the game-theoretic setting, identifying minimax-optimal strategies through conservative value iteration and strategy-wise uncertainty bonuses (Zhong et al., 2022; Cui & Du, 2022b; Zhang et al., 2023; Yan et al., 2024).

The theoretical validity of these pessimistic approaches is grounded in the notion of *unilateral concentrability*—the requirement that the offline logs sufficiently cover any individual agent's deviations

from a target equilibrium (Cui & Du, 2022b). However, anchoring this coverage requirement to an unknown optimal equilibrium π^* introduces a fundamental conceptual circularity, as the very policy required to define data sufficiency is the target of the learning process itself. Our work addresses this by anchoring the learning objective to a fixed and known reference policy π^{ref} (Chen & Zhang, 2026), transforming the coverage requirement into a condition with a known reference anchor that remains grounded in the support of the offline data.

KL Regularization and Behavioral Anchoring. KL regularization relative to a fixed reference policy has emerged as a fundamental tool for enforcing behavioral constraints and incorporating prior knowledge into reinforcement learning frameworks (Xiong et al., 2023; Munos et al., 2024). By penalizing deviations from an anchor distribution, such as a pre-trained model or human demonstrations, this mechanism effectively mitigates policy drift and stabilizes the resulting learning dynamics (Ye et al., 2024; Nayak et al., 2025). The most prominent application of this principle is found in the alignment of large language models through Reinforcement Learning from Human Feedback (RLHF), where a KL divergence penalty ensures the optimized model remains within the trusted region of the initial distribution (Ouyang et al., 2022; Rafailov et al., 2023).

While regularized objectives have been extensively analyzed across both single-agent and multi-agent regimes (Xie et al., 2024; Zhao et al., 2025; Zhang et al., 2025a;b), current theoretical results typically suggest that regularization primarily serves as a numerical stabilizer, yielding sample complexity rates that mirror those of their unregularized counterparts (Ye et al., 2024). Consequently, the capacity of KL regularization to function as a standalone, implicit pessimism mechanism in offline general-sum games has remained unestablished. Our work bridges this gap by demonstrating that regularized KL anchoring suffices for stable equilibrium recovery without explicit bonuses.

Statistical Efficiency and Fast Rates. Most offline learning guarantees characterize the standard $\widetilde{\mathcal{O}}(1/\sqrt{n})$ statistical rate (Jin et al., 2021; Shi et al., 2022). While recent investigations by Zhang et al. (2026); Chen et al. (2026) demonstrated that an accelerated $\widetilde{\mathcal{O}}(1/n)$ rate is achievable in two-player zero-sum games, those results rely on the minimax structure and skew-symmetry inherent to strictly competitive dynamics. In contrast, general-sum games are fundamentally more complex; whereas finding an equilibrium is already PPAD-complete in the two-player general-sum case (Chen et al., 2009), the multi-player setting remains equally intractable while introducing more complex strategic interactions (Papadimitriou, 1994; Daskalakis et al., 2009).

We broaden this theoretical landscape in two directions. First, we establish that the fast $\widetilde{\mathcal{O}}(1/n)$ rate persists for Nash equilibria in general-sum games, proving that independence—rather than zero-sum symmetry—is the key driver of this acceleration. Second, for the computationally tractable CCE, we provide a pessimism-free framework that achieves the standard $\widetilde{\mathcal{O}}(1/\sqrt{n})$ rate. Together, these findings establish a tractable foundation for pessimism-free learning in multi-player general-sum games.

7 Conclusion

In this work, we have demonstrated that KL regularization serves as a robust and standalone mechanism for stabilizing offline multi-agent learning, effectively mitigating distribution shift without the need for explicit pessimistic penalties. By shifting the coverage requirement from unknown optimal policies to a fixed and known reference anchor, we established a reference-anchored theoretical framework that bypasses the conceptual circularity inherent in prior literature. We proposed GANE, which achieves an accelerated $\widetilde{\mathcal{O}}(1/n)$ statistical rate for Nash equilibria by leveraging the product structure of independent policies. Furthermore, we provided GAMD, a decentralized iterative framework that recovers Coarse Correlated Equilibria at the standard $\widetilde{\mathcal{O}}(1/\sqrt{n})$ rate without explicit pessimistic bonuses. Together, these results establish KL regularization as a tractable, pessimism-free alternative for establishing a fast statistical rate in offline game-theoretic learning.

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

Claire Chen and Yuheng Zhang. Fast rates in α-potential games via regularized mirror descent. *arXiv preprint arXiv:2605.00268*, 2026. 2, 4, 9

Claire Chen, Shuze Liu, and Shangtong Zhang. Efficient policy evaluation with safety constraint for reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 34989–35010, 2025. 13

Claire Chen, Yuheng Zhang, Xinyu Liu, Zixuan Xie, Shuze Daniel Liu, and Nan Jiang. Offline two-player zero-sum markov games with kl regularization. *arXiv preprint arXiv:2605.13025*, 2026. 9

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009. 7, 9

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. 1, 2, 4, 8, 13

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. 2, 4, 8, 9

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009. 7, 9

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

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, 13

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv Preprint*, 2020. 1, 13

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2443–2453, 2017. 1

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024. 13

Shuze Liu and Shangtong Zhang. Efficient policy evaluation with offline data informed behavior policy design. In *Proceedings of the International Conference on Machine Learning*, 2024. 13

Shuze Liu, Claire Chen, and Shangtong Zhang. Doubly optimal policy evaluation for reinforcement learning. *ArXiv Preprint*, 2024. 13

Shuze Daniel Liu, Claire Chen, and Shangtong Zhang. Efficient multi-policy evaluation for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18951–18959, 2025. 13

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020. 13

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995. 3

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024. 1, 9

Anupam Nayak, Tong Yang, Osman Yagan, Gauri Joshi, and Yuejie Chi. Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060*, 2025. 9

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. 1, 9

Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994. 9

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 9

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. 13

Tim Roughgarden. *Twenty lectures on algorithmic game theory*. Cambridge University Press, 2016. 8

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2025. 7

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pp. 19967–20025. PMLR, 2022. 9

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021. 13

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *Proceedings of the International Conference on Machine Learning*, 2021. 4

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. 13

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024. 9

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023. 3, 9

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning for offline zero-sum markov games. *Operations research*, 72(6):2430–2445, 2024. 8

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model, 2024. 1, 2, 9

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022. 13

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, 2, 8, 13

Yuheng Zhang, Dian Yu, Tao Ge, Linfeng Song, Zhichen Zeng, Haitao Mi, Nan Jiang, and Dong Yu. Improving llm general preference alignment via optimistic online mirror descent. *arXiv preprint arXiv:2502.16852*, 2025a. 9

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning LLMs with general preferences via no-regret learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. 3, 9

Yuheng Zhang, Claire Chen, and Nan Jiang. Beyond pessimism: Offline learning in kl-regularized games. *arXiv preprint arXiv:2604.06738*, 2026. 4, 9

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025. 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, 8, 13

A Additional Related Work

Pessimism in Offline Reinforcement Learning. Offline reinforcement learning (RL) focuses on policy optimization from pre-collected datasets without the benefit of interactive exploration. The primary theoretical obstacle in this regime is distribution shift, wherein the target policy's visitation frequency deviates from the logged behavior, inducing severe overestimation and extrapolation error (Levine et al., 2020; Liu et al., 2024; Liu & Zhang, 2024; Chen et al., 2025; Liu et al., 2025). To counteract this bias, researchers utilize the principle of *pessimism*, which provides a robust framework for establishing sample-efficient guarantees by penalizing values in poorly explored state-action regions (Liu et al., 2020; Rashidinejad et al., 2021; Jin et al., 2021; Xie et al., 2021; Uehara & Sun, 2021; Zhan et al., 2022). These conservative strategies, often implemented via lower confidence bounds or conservative value iterations, have been proven minimax optimal under various single-policy concentrability settings (Li et al., 2024).

In game-theoretic environments, the complexity of this shift is intensified by strategic interactions, requiring sufficient data coverage of unilateral deviations for each agent to ensure successful equilibrium recovery (Cui & Du, 2022a). Consequently, the current literature on offline Markov games relies heavily on these pessimistic foundations to mitigate information gaps within the static logs (Cui & Du, 2022a; Zhong et al., 2022; Zhang et al., 2023). In contrast, our approach establishes that KL regularization suffices to stabilize learning without relying on such explicit pessimistic mechanisms.

B Proof

B.1 Proof of Theorem 4.2

B.1.1 Decomposition of the Total Exploitability Gap

To bound the suboptimality of the learned joint policy $\hat{\pi}$, we decompose the Nash Equilibrium gap into an empirical optimization error and statistical evaluation errors by pivoting through the estimated value functions $\hat{V}$ and $\hat{Q}$.

Recall the definition of the NE gap (Total Exploitability) for the learned policy $\hat{\pi}$:

$$\mathrm{Gap}_{\mathrm{NE}}(\hat{\pi}) = \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\left[V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x)\right].$$

We define the estimated best-response value for Player i under the empirical game evaluated by $\hat{Q}$ as:

$$\hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x) := \max_{\pi_i'} \hat{V}_i^{\pi_i',\hat{\pi}_{-i}}(x).$$

Here we restate Lemma 4.1:

Lemma 4.1 (Gap Decomposition)**.** *For any learned joint policy $\hat{\pi}$ and any player $i \in [m]$, the unilateral exploitability gap can be exactly decomposed as:*

$$V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x) = \underbrace{\left(\hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\hat{\pi}}(x)\right)}_{\textit{Term I: Optimization Gap}} + \underbrace{\left(V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)\right)}_{\textit{Term II: Best-Response Error}} + \underbrace{\left(\hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x)\right)}_{\textit{Term III: On-Policy Error}}.$$

Bounding the Decomposition Terms: This decomposition isolates the distinct sources of error, allowing us to bound the gap systematically:

- **Term I** represents the empirical external regret of Player i operating strictly on the estimated game $\hat{Q}$. Because the joint policy $\hat{\pi}$ is obtained via an oracle exactly solving the empirical games, this term strictly corresponds to the optimization error and vanishes identically.
- **Term II** and **Term III** represent the statistical evaluation errors. By applying algebraic expansions, these terms can be unrolled into the sum of regression errors over the contextual distributions, respectively. These errors are precisely controlled by our Unilateral Concentrability assumption and converted into squared errors via the KL-divergence strong convexity.

B.1.2 Bounding the Statistical Error for the Idealized Algorithm

Let $\hat{\pi}$ be the joint policy returned by the idealized GANE algorithm (Algorithm 1). Because $\hat{\pi}$ is an exact regularized Nash Equilibrium of the estimated game $\hat{Q}$, no player can unilaterally improve their empirical value. Thus, for all $i \in [m]$, the empirical optimization gap vanishes:

$$\text{Term I}_i = \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\hat{\pi}}(x) = 0.$$

Therefore, the total unilateral exploitability gap is bounded strictly by the statistical evaluation errors:

$$\text{Gap}_{\text{NE}}(\hat{\pi}) \leq \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\left[\text{Term II}_i + \text{Term III}_i\right],$$

where $\text{Term II}_i = V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)$ and $\text{Term III}_i = \hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x)$.

Since $\hat{\pi}$ is an exact NE on the estimated values, the marginalized policy $\hat{\pi}_i$ is the empirical best response to $\hat{\pi}_{-i}$, implying $\hat{V}_i^{\dagger,\hat{\pi}_{-i}} = \hat{V}_i^{\hat{\pi}}$. We analyze the sum of these evaluation errors by first bounding the local value difference of the best response.

Step 1: 1-Smoothness of the Log-Partition Function. Fix a context x. We define the value difference of the best response as $\Delta_i(x) \coloneqq V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)$. For notational clarity, we define the marginalized empirical Q-function as $\bar{Q}_i(x, a_i) \coloneqq \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}}[\hat{Q}_i(x, a_i, \boldsymbol{a}_{-i})]$ and the true best-response Q-value as $Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) \coloneqq \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}}[r_i^\star(x, \boldsymbol{a})]$. Both the true best response and the empirical best response values can be expressed exactly via the log-partition function $\Phi(\theta) = \log\sum_a \exp(\theta_a)$. We define the natural parameters (logits) for the true and empirical best responses respectively:

$$\theta_i^\dagger(x, \cdot) \coloneqq \eta Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) + \log \pi_{\text{ref},i}(\cdot \mid x),$$
$$\hat{\theta}_i(x, \cdot) \coloneqq \eta \bar{Q}_i(x, \cdot) + \log \pi_{\text{ref},i}(\cdot \mid x).$$

The values are exactly $V_i^\dagger(x) = \eta^{-1}\Phi(\theta_i^\dagger)$ and $\hat{V}_i^\dagger(x) = \eta^{-1}\Phi(\hat{\theta}_i)$. Because the log-partition function Φ is 1-smooth with respect to the ℓ_∞-norm, its Bregman divergence (which corresponds to the KL divergence) is strictly bounded by the squared ℓ_∞ distance:

$$\Phi(\theta^\dagger) - \Phi(\hat{\theta}) \leq \langle \nabla\Phi(\hat{\theta}), \theta^\dagger - \hat{\theta}\rangle + \frac{1}{2}\|\theta^\dagger - \hat{\theta}\|_\infty^2.$$

Crucially, the gradient map $\nabla\Phi(\hat{\theta})$ yields exactly the probabilities of the empirical best response, which is the learned policy $\hat{\pi}_i$. Dividing by η, we obtain the local value difference bound:

$$\Delta_i(x) \leq \mathbb{E}_{a_i\sim\hat{\pi}_i(\cdot|x)}\left[Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\right] + \frac{\eta}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot)\right\|_\infty^2. \tag{2}$$

Proof. By definition, the local value difference is given by the scaled difference of the log-partition functions:

$$\begin{aligned}\Delta_i(x) &= V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)\\ &= \eta^{-1}\Phi(\theta_i^\dagger(x, \cdot)) - \eta^{-1}\Phi(\hat{\theta}_i(x, \cdot))\\ &= \eta^{-1}\Big(\Phi(\theta_i^\dagger) - \Phi(\hat{\theta}_i)\Big).\end{aligned}$$

Applying the 1-smoothness property of Φ to the term in the parentheses:

$$\Delta_i(x) \leq \eta^{-1}\left[\langle \nabla\Phi(\hat{\theta}_i), \theta_i^\dagger - \hat{\theta}_i\rangle + \frac{1}{2}\|\theta_i^\dagger - \hat{\theta}_i\|_\infty^2\right]. \tag{3}$$

We now evaluate the components of this upper bound. First, the gradient of the log-partition function $\nabla\Phi(\theta)$ yields the Softmax probabilities. Since $\hat{\theta}_i$ represents the logits of the empirical game, its Softmax is exactly the learned empirical policy:

$$\nabla\Phi(\hat{\theta}_i)_{a_i} = \hat{\pi}_i(a_i|x).$$

Second, we evaluate the difference in logits. By definition, $\theta_i^\dagger(x, a_i) = \eta Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) + \log \pi_{\mathrm{ref},i}(a_i \mid x)$, and similarly for $\hat{\theta}_i$. When we subtract them, the reference policy perfectly cancels out:

$$\begin{aligned}\theta_i^\dagger(x, a_i) - \hat{\theta}_i(x, a_i) &= \Big(\eta Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) + \log \pi_{\mathrm{ref},i}(a_i|x)\Big) - \Big(\eta \bar{Q}_i(x, a_i) + \log \pi_{\mathrm{ref},i}(a_i|x)\Big) \\ &= \eta\Big(Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\Big).\end{aligned}$$

Substituting these evaluated components back into the inner product term of Equation (3):

$$\begin{aligned}\langle \nabla\Phi(\hat{\theta}_i), \theta_i^\dagger - \hat{\theta}_i\rangle &= \sum_{a_i \in \mathcal{A}_i} \hat{\pi}_i(a_i|x) \cdot \eta\Big(Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\Big) \\ &= \eta \mathbb{E}_{a_i \sim \hat{\pi}_i(\cdot|x)}\left[Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\right].\end{aligned}$$

Next, we substitute the logit difference into the squared ℓ_∞ norm term:

$$\begin{aligned}\frac{1}{2}\|\theta_i^\dagger - \hat{\theta}_i\|_\infty^2 &= \frac{1}{2}\left\|\eta\Big(Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot)\Big)\right\|_\infty^2 \\ &= \frac{\eta^2}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot)\right\|_\infty^2.\end{aligned}$$

Finally, plugging these two terms back into Equation (3) and multiplying through by η^{-1}:

$$\begin{aligned}\Delta_i(x) &\le \eta^{-1}\left[\eta \mathbb{E}_{a_i \sim \hat{\pi}_i(\cdot|x)}\left[Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\right] + \frac{\eta^2}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot)\right\|_\infty^2\right] \\ &= \mathbb{E}_{a_i \sim \hat{\pi}_i(\cdot|x)}\left[Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i)\right] + \frac{\eta}{2}\left\|Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot)\right\|_\infty^2.\end{aligned}$$

□

Step 2: Evaluating Term II via the Pointwise Error Identity. We introduce the regression error for Player i:

$$\mathcal{Z}_i(x, \boldsymbol{a}) \coloneqq \hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a}).$$

The Q-value difference expands algebraically as:

$$Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i) = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[-\mathcal{Z}_i(x, \boldsymbol{a})\right]. \tag{4}$$

To see (4), recall the definitions for the true best-response Q-value, the estimated Q-value, and the regression error for Player i:

$$\begin{aligned}Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) &= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[r_i^\star(x, \boldsymbol{a})\right], \\ \bar{Q}_i(x, a_i) &= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[\hat{Q}_i(x, \boldsymbol{a})\right], \\ \mathcal{Z}_i(x, \boldsymbol{a}) &\coloneqq \hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a}).\end{aligned}$$

By expanding the Q-value difference and applying direct substitution, we obtain:

$$\begin{aligned}&Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i) \\ &= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[r_i^\star(x, \boldsymbol{a})\right] - \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[\hat{Q}_i(x, \boldsymbol{a})\right] \\ &= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[r_i^\star(x, \boldsymbol{a}) - \hat{Q}_i(x, \boldsymbol{a})\right]\end{aligned}$$

Grouping the terms strictly isolates the exact definition of $\mathcal{Z}_i$:

$$\begin{aligned}&= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[-\big(\hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a})\big)\right] \\ &= \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[-\mathcal{Z}_i(x, \boldsymbol{a})\right].\end{aligned} \tag{5}$$

Recall that the Total Exploitability of the game is evaluated in expectation over the context distribution $x \sim \rho$. Therefore, to bound the total gap, we evaluate the expected value of our decomposition terms.

For Player i, we define the pointwise value difference of the best response as $\Delta_i(x) \coloneqq V_i^{\dagger,\hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x)$. By direct algebraic substitution, the linear Q-value difference is strictly equal to the negative regression error:

$$Q_i^{\dagger,\hat{\pi}_{-i}}(x, a_i) - \bar{Q}_i(x, a_i) = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}} \left[-\mathcal{Z}_i(x, \boldsymbol{a}) \right].$$

Taking the expectation over the context distribution $x \sim \rho$, Term II is bounded by:

$$\mathbb{E}_{x\sim\rho}[\text{Term II}_i(x)] \leq \mathbb{E}_{x\sim\rho, \boldsymbol{a}\sim(\hat{\pi}_i^\dagger, \hat{\pi}_{-i})} \left[-\mathcal{Z}_i(x, \boldsymbol{a}) \right] + \frac{\eta}{2} \mathbb{E}_{x\sim\rho} \left[\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot) \right\|_\infty^2 \right].$$

Step 3: Evaluating Term III via Pointwise Error Identity. Next, we apply the similar algebraic derivation on Term III.

$$\mathbb{E}_{x\sim\rho}[\text{Term III}_i(x)] = \mathbb{E}_{x\sim\rho} \left[\hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x) \right]$$

We define the local joint policy evaluation error as $\delta_i(x) \coloneqq \hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x)$. Expanding both value functions according to their definitions, the KL regularization terms exactly cancel out because both are evaluated under the identical policy $\hat{\pi}$:

$$\begin{aligned}
\delta_i(x) &= \left(\mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\hat{Q}_i(x, \boldsymbol{a}) \right] - \text{Reg}_i(\hat{\pi}) \right) \\
&\quad - \left(\mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[r_i^\star(x, \boldsymbol{a}) \right] - \text{Reg}_i(\hat{\pi}) \right) \\
&= \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a}) \right]
\end{aligned}$$

Grouping the terms recovers the exact definition of the regression error $\mathcal{Z}_i$:

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\underbrace{\hat{Q}_i(x, \boldsymbol{a}) - r_i^\star(x, \boldsymbol{a})}_{=\mathcal{Z}_i(x,\boldsymbol{a})} \right] \\
&= \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\mathcal{Z}_i(x, \boldsymbol{a}) \right]
\end{aligned}$$

We evaluate this identity in expectation over the context distribution $x \sim \rho$:

$$\begin{aligned}
\mathbb{E}_{x\sim\rho} \left[\delta_i(x) \right] &= \mathbb{E}_{x\sim\rho} \left[\mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\mathcal{Z}_i(x, \boldsymbol{a}) \right] \right] \\
&= \mathbb{E}_{x\sim\rho,\, \boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\mathcal{Z}_i(x, \boldsymbol{a}) \right].
\end{aligned}$$

Thus, taking the expectation over the context distribution, we obtain:

$$\mathbb{E}_{x\sim\rho}[\text{Term III}_i(x)] = \mathbb{E}_{x\sim\rho} \left[\hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x) \right] = \mathbb{E}_{x\sim\rho,\, \boldsymbol{a}\sim\hat{\pi}(\cdot|x)} \left[\mathcal{Z}_i(x, \boldsymbol{a}) \right].$$

When adding the expected Term II and Term III, the linear evaluation errors $-\mathcal{Z}_i$ and $+\mathcal{Z}_i$ completely cancel out. This leaves the total expected gap bounded strictly by the squared errors:

$$\mathbb{E}_{x\sim\rho}\big[\text{Term II}_i(x) + \text{Term III}_i(x)\big] \leq \frac{\eta}{2} \mathbb{E}_{x\sim\rho} \left[\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot) \right\|_\infty^2 \right].$$

Step 4: Unrolling the Linear Error and Cauchy-Schwarz. Taking the absolute value of the Q-value difference identity (5) and applying the triangle inequality yields the pointwise linear bound:

$$\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x, \cdot) - \bar{Q}_i(x, \cdot) \right\|_\infty \leq \max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}} \left[|\mathcal{Z}_i(x, \boldsymbol{a})| \right].$$

Notice that taking the maximum over the action a_i is equivalent to taking the supremum over all possible state-conditioned policies for Player i. Because the opponent's policy remains fixed to $\hat{\pi}_{-i}$,

this shifts the action evaluation distribution from the learned joint policy to a unilateral deviation distribution $(\pi_i', \hat{\pi}_{-i})$. Thus, we can bound the point-wise error by taking the supremum over all such unilateral deviations:

$$\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot) \right\|_\infty \leq \sup_{\pi_i'} \mathbb{E}_{\boldsymbol{a}\sim(\pi_i',\hat{\pi}_{-i})}\left[|\mathcal{Z}_i(x,\boldsymbol{a})|\right],$$

where the inner action a_i is selected by the deviation policy π_i'.

We now square both sides. Applying Jensen's inequality for the inner expectation, we obtain:

$$\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot) \right\|_\infty^2 \leq \sup_{\pi_i'} \mathbb{E}_{\boldsymbol{a}\sim(\pi_i',\hat{\pi}_{-i})}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right] = \max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right].$$

Substituting this squared unrolled error back into the sum of Term II and Term III bounds gives:

$$\begin{aligned}\mathbb{E}_{x\sim\rho}\left[\text{Term II}_i(x) + \text{Term III}_i(x)\right] &\leq \frac{\eta}{2}\mathbb{E}_{x\sim\rho}\left[\left\| Q_i^{\dagger,\hat{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2\right] \\ &\leq \frac{\eta}{2}\mathbb{E}_{x\sim\rho}\left[\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\hat{\pi}_{-i}}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right]\right]. \qquad (6)\end{aligned}$$

Lemma B.1 (Uniform Value Bound)**.** *For any player $i \in [m]$ and any fixed opponent joint policy ν_{-i}, the regularized best-response value function $V_i^{\dagger,\nu_{-i}}(x) = \max_{\pi_i} V_i^{\pi_i,\nu_{-i}}(x)$ satisfies:*

$$0 \leq V_i^{\dagger,\nu_{-i}}(x) \leq 1.$$

Consequently, the uniform infinity norm is strictly bounded: $\left\| V_i^{\dagger,\nu_{-i}} \right\|_\infty \leq 1.$

Proof. For a fixed context x, the best-response value for Player i against ν_{-i} can be written in its exact log-sum-exp form:

$$V_i^{\dagger,\nu_{-i}}(x) = \eta^{-1}\log \sum_{a_i\in\mathcal{A}_i} \pi_{\text{ref},i}(a_i|x)\exp\left(\eta Q_i^{\dagger,\nu_{-i}}(x,a_i)\right),$$

where $Q_i^{\dagger,\nu_{-i}}(x,a_i) = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\nu_{-i}}[r_i^\star(x,\boldsymbol{a})]$. Using the bounds of the deterministic reward $r_i^\star \in [0,1]$, the Q-value is uniformly bounded by $Q_i^{\dagger,\nu_{-i}}(x,a_i) \in [0,1]$.

Because the reference policy $\pi_{\text{ref},i}(\cdot|x)$ is a valid probability distribution that sums to 1, we apply the standard bounds of the log-sum-exp function: $\eta^{-1}\log\sum p_j\exp(\eta q_j) \in [\min q_j, \max q_j]$. Therefore, the best-response value is strictly bounded by the minimum and maximum expected Q-values:

$$\min_{a_i} Q_i^{\dagger,\nu_{-i}}(x,a_i) \leq V_i^{\dagger,\nu_{-i}}(x) \leq \max_{a_i} Q_i^{\dagger,\nu_{-i}}(x,a_i).$$

Since all Q-values are bounded in $[0,1]$, we conclude that $V_i^{\dagger,\nu_{-i}}(x) \in [0,1]$. This exact characterization eliminates any dependence on the reference policy's minimum probability mass, yielding the uniform bound $\left\| V_i^{\dagger,\nu_{-i}} \right\|_\infty \leq 1$. □

To invoke our Reference-Anchored Unilateral Concentrability (Assumption 3.2), we shift the evaluation distribution of the opponents from the learned joint policy $\hat{\pi}_{-i}$ to the fixed reference policy π_{-i}^{ref}, while maintaining the arbitrary deviation policy π_i' for Player i.

To bound the distribution shift, we utilize the first-order optimality of the regularized Nash Equilibrium. For each opponent $j \neq i$, the marginalized policy $\hat{\pi}_j$ is the unique maximizer of the regularized payoff, which implies the Gibbs form:

$$\hat{\pi}_j(a_j \mid x) = \frac{\pi_j^{\text{ref}}(a_j \mid x)\exp\left(\eta\bar{Q}_j(x,a_j)\right)}{\sum_{a_j'}\pi_j^{\text{ref}}(a_j' \mid x)\exp\left(\eta\bar{Q}_j(x,a_j')\right)},$$

where $\bar{Q}_j$ is the marginalized empirical Q-function. Since the empirical Q-functions are bounded in $[0,1]$, $\|\bar{Q}_j\|_\infty \leq 1$. We obtain the point-wise ratio for each action a_j:

$$\frac{\hat{\pi}_j(a_j \mid x)}{\pi_j^{\mathrm{ref}}(a_j \mid x)} = \frac{\exp\left(\eta \bar{Q}_j(x, a_j)\right)}{\mathbb{E}_{a'_j \sim \pi_j^{\mathrm{ref}}} \exp\left(\eta \bar{Q}_j(x, a'_j)\right)} \leq \frac{\exp(\eta)}{1} = \exp(\eta).$$

Since the joint policy of the $m-1$ opponents factorizes at each state, the joint density ratio is bounded by $\exp(\eta(m-1))$. To unify our notation across proofs, we define a single global distribution shift constant bounding the density ratio of all m players:

$$C_{\mathrm{shift}} \coloneqq \exp(\eta m).$$

Thus, the total action likelihood ratio for the $m-1$ opponents is strictly bounded by C_{shift}:

$$\frac{\hat{\pi}_{-i}(\boldsymbol{a}_{-i} \mid x)}{\pi_{-i}^{\mathrm{ref}}(\boldsymbol{a}_{-i} \mid x)} = \prod_{j \neq i} \frac{\hat{\pi}_j(a_j \mid x)}{\pi_j^{\mathrm{ref}}(a_j \mid x)} \leq \exp\left(\eta(m-1)\right) \leq C_{\mathrm{shift}}.$$

To relate the expectations over different context-action distributions, we consider the likelihood ratio of a joint action $\boldsymbol{a}$ under the algorithm's deviation policy $(\pi'_i, \hat{\pi}_{-i})$ versus the reference deviation policy $(\pi'_i, \pi_{-i}^{\mathrm{ref}})$. Because the environment distribution ρ is identical for both, they cancel out in the ratio, leaving only the product of the marginalized opponent policy ratios.

Using this bound to perform a change of measure, we shift the evaluation from the algorithm's trajectory to the covered unilateral deviation trajectory:

$$\mathbb{E}_{x \sim \rho}\left[\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}\left[\mathcal{Z}_i(x, \boldsymbol{a})^2\right]\right] \leq C_{\mathrm{shift}} \cdot \mathbb{E}_{x \sim \rho}\left[\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}^{\mathrm{ref}}}\left[\mathcal{Z}_i(x, \boldsymbol{a})^2\right]\right].$$

To formalize the application of our data coverage assumption on this worst-case action, we explicitly construct the deviation policy that maximizes the expected squared error. Let $\tilde{\pi}_i$ be the deterministic policy that greedily selects the error-maximizing action for Player i at each context: $\tilde{\pi}_i(x) \coloneqq \mathrm{argmax}_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}^{\mathrm{ref}}}[\mathcal{Z}_i(x, a_i, \boldsymbol{a}_{-i})^2]$. The resulting joint policy profile $\tilde{\pi} = (\tilde{\pi}_i, \pi_{-i}^{\mathrm{ref}})$ strictly belongs to the set of reference-anchored unilateral deviations $\Pi_{\mathrm{ref-uni}}$.

By substituting this maximizing policy, we convert the worst-case action into an expectation over the joint distribution $\rho(x)\tilde{\pi}(\boldsymbol{a}|x)$. Applying Assumption 3.2, we shift the evaluation to the offline dataset distribution μ:

$$\begin{aligned}
\mathbb{E}_{x \sim \rho}\left[\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}^{\mathrm{ref}}}\left[\mathcal{Z}_i(x, \boldsymbol{a})^2\right]\right] &= \mathbb{E}_{x \sim \rho, \boldsymbol{a} \sim \tilde{\pi}(\cdot|x)}\left[\mathcal{Z}_i(x, \boldsymbol{a})^2\right] \\
&= \mathbb{E}_{(x,\boldsymbol{a}) \sim \mu}\left[\frac{\rho(x)\tilde{\pi}(\boldsymbol{a}|x)}{\mu(x, \boldsymbol{a})} \mathcal{Z}_i(x, \boldsymbol{a})^2\right] \\
&\leq C_{\mathrm{uni}} \mathbb{E}_{\mu}\left[\mathcal{Z}_i(x, \boldsymbol{a})^2\right].
\end{aligned}$$

Combining this with the opponent distribution shift C_{shift}, together with the previous bound we derived in (6), we obtain the fully shifted bound:

$$\mathbb{E}_{x \sim \rho}\left[\text{Term II}_i + \text{Term III}_i\right] \leq \frac{\eta}{2} C_{\mathrm{shift}} C_{\mathrm{uni}} \mathbb{E}_{\mu}\left[\mathcal{Z}_i^2\right].$$

Using the standard fast-rate statistical guarantee for least-squares regression, the population squared regression error under the data distribution μ is bounded by the function class complexity $|\mathcal{Q}_i|$ and the number of samples n simultaneously for all players with probability at least $1-\delta$:

$$\mathbb{E}_{\mu}\left[\mathcal{Z}_i^2\right] \leq \mathcal{O}\left(\frac{\log(m|\mathcal{Q}_i|/\delta)}{n}\right).$$

By substituting this rate into our unrolled bound, we effectively resolve the expectation over μ by replacing it with its statistical upper limit. We conclude the final suboptimality gap for Player i:

$$\mathbb{E}_{x \sim \rho}\left[\text{Term II}_i + \text{Term III}_i\right] \leq \widetilde{\mathcal{O}}\left(\frac{\eta C_{\mathrm{shift}} C_{\mathrm{uni}} \log|\mathcal{Q}_i|}{n}\right).$$

Final Bound on Total Exploitability. The Total Exploitability of the game is defined as the sum of the unilateral suboptimality across all m players. By summing the bound derived for Player i over the entire set $\mathcal{N} = [m]$, and retaining the distribution shift factor $C_{\text{shift}} = \exp(\eta m)$ while absorbing logarithmic terms into the $\widetilde{\mathcal{O}}$ notation, we obtain:

$$\begin{aligned}\text{Gap}_{\text{NE}}(\hat{\pi}) &= \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\left[V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x)\right] \\ &\leq \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\left[\text{Term II}_i + \text{Term III}_i\right] \\ &\leq \widetilde{\mathcal{O}}\left(\frac{m\eta C_{\text{shift}} C_{\text{uni}} \log|\mathcal{Q}|}{n}\right).\end{aligned}$$

Substituting the bound $C_{\text{shift}} = \exp(\eta m)$, the final statistical rate is:

$$\text{Gap}_{\text{NE}}(\hat{\pi}) \leq \widetilde{\mathcal{O}}\left(\frac{m\eta C_{\text{uni}} \exp(\eta m) \log|\mathcal{Q}|}{n}\right).$$

This result establishes that Algorithm 1 (GANE) achieves a fast statistical rate of $\widetilde{\mathcal{O}}(1/n)$ in the m-player general-sum offline setting. By operating in the Contextual Bandit setting, we completely eliminate the exponential horizon dependencies ($\exp(H)$) that artificially inflate pessimism-free RL bounds, explicitly highlighting the fast statistical rate of product-policy equilibrium recovery.

B.2 Proof of Lemma 5.1

Proof. We drop player and context subscripts throughout and let $\pi^* \in \operatorname{argmax}_{\pi\in\Delta(\mathcal{A})} \sum_{t=1}^{T} f^{(t)}(\pi)$ denote an arbitrary comparator. Denote the cumulative objective by $F_t(\pi) \coloneqq \sum_{k=1}^{t} f^{(k)}(\pi)$, with $F_0 \equiv 0$.

Iterates as Follow-The-Leader. Unrolling the geometric-mean update in Algorithm 2 with initialization $\pi^{(1)} = \pi^{\text{ref}}$ and decaying stepsize $\gamma_t = 1/t$ yields the closed form $\pi^{(t+1)}(\cdot) \propto \pi^{\text{ref}}(\cdot) \cdot \exp\big(\frac{\eta}{t}\sum_{k=1}^{t} Q^{(k)}(\cdot)\big)$, which is precisely the Follow-The-Leader (FTL) iterate on the cumulative objective,

$$\pi^{(t+1)} = \operatorname*{argmax}_{\pi\in\Delta(\mathcal{A})} F_t(\pi). \tag{7}$$

Be-The-Leader. By induction on T, we show

$$\sum_{t=1}^{T} f^{(t)}(\pi^{(t+1)}) \geq \sum_{t=1}^{T} f^{(t)}(\pi^*). \tag{8}$$

The base case $T = 1$ is immediate since $\pi^{(2)} = \operatorname{argmax} f^{(1)}$. For the inductive step, apply the hypothesis at $T-1$ with comparator $\pi^{(T+1)}$:

$$\sum_{t=1}^{T} f^{(t)}(\pi^{(t+1)}) = \sum_{t=1}^{T-1} f^{(t)}(\pi^{(t+1)}) + f^{(T)}(\pi^{(T+1)}) \geq F_{T-1}(\pi^{(T+1)}) + f^{(T)}(\pi^{(T+1)}) = F_T(\pi^{(T+1)}) \geq F_T(\pi^*),$$

where the final inequality uses $\pi^{(T+1)} = \operatorname{argmax} F_T$. Rearranging (8) bounds the regret by the iterate stability:

$$\sum_{t=1}^{T}\left(f^{(t)}(\pi^*) - f^{(t)}(\pi^{(t)})\right) \leq \sum_{t=1}^{T}\left(f^{(t)}(\pi^{(t+1)}) - f^{(t)}(\pi^{(t)})\right). \tag{9}$$

Strong-concavity lower bound. Pinsker's inequality $\mathrm{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2$ shows that the negative entropy is 1-strongly convex with respect to $\|\cdot\|_1$, so each $f^{(t)}$ is η^{-1}-strongly concave and F_t is (t/η)-strongly concave. Applying strong concavity at the maximizers $\pi^{(t)} = \operatorname{argmax} F_{t-1}$ and $\pi^{(t+1)} = \operatorname{argmax} F_t$ gives

$$F_{t-1}(\pi^{(t)}) - F_{t-1}(\pi^{(t+1)}) \geq \tfrac{t-1}{2\eta}\|\pi^{(t+1)} - \pi^{(t)}\|_1^2,$$
$$F_t(\pi^{(t+1)}) - F_t(\pi^{(t)}) \geq \tfrac{t}{2\eta}\|\pi^{(t+1)} - \pi^{(t)}\|_1^2.$$

Summing the two inequalities and using $F_t - F_{t-1} = f^{(t)}$ yields the per-step stability lower bound

$$f^{(t)}(\pi^{(t+1)}) - f^{(t)}(\pi^{(t)}) \geq \tfrac{2t-1}{2\eta}\|\pi^{(t+1)} - \pi^{(t)}\|_1^2. \tag{10}$$

Hölder upper bound. The Gibbs form (7) gives $\log(\pi^{(t)}/\pi^{\mathrm{ref}}) = \frac{\eta}{t-1}\sum_{k=1}^{t-1} Q^{(k)} - \log Z_{t-1}$ for $t \geq 2$, and $\log(\pi^{(1)}/\pi^{\mathrm{ref}}) = 0$ for $t = 1$. Substituting into $\nabla f^{(t)}(\pi) = Q^{(t)} - \eta^{-1}(\log(\pi/\pi^{\mathrm{ref}}) + \mathbf{1})$ and projecting onto the simplex tangent space (i.e., discarding components parallel to $\mathbf{1}$, which vanish against $\pi^{(t+1)} - \pi^{(t)}$ since $\mathbf{1}^\top(\pi^{(t+1)} - \pi^{(t)}) = 0$),

$$\langle \nabla f^{(t)}(\pi^{(t)}), \pi^{(t+1)} - \pi^{(t)} \rangle = \Big\langle Q^{(t)} - \tfrac{1}{t-1}\textstyle\sum_{k=1}^{t-1} Q^{(k)},\ \pi^{(t+1)} - \pi^{(t)} \Big\rangle,$$

with the convention that the second term vanishes when $t = 1$. Because each $\|Q^{(k)}\|_\infty \leq 1$, the centered gradient has ℓ_∞-norm at most 2. Concavity of $f^{(t)}$ combined with Hölder's inequality therefore gives

$$f^{(t)}(\pi^{(t+1)}) - f^{(t)}(\pi^{(t)}) \leq \langle \nabla f^{(t)}(\pi^{(t)}), \pi^{(t+1)} - \pi^{(t)} \rangle \leq 2\|\pi^{(t+1)} - \pi^{(t)}\|_1. \tag{11}$$

Combining the bounds. Equations (10) and (11) imply $\|\pi^{(t+1)} - \pi^{(t)}\|_1 \leq \frac{4\eta}{2t-1}$, and substituting back into (11) gives the per-step stability bound $f^{(t)}(\pi^{(t+1)}) - f^{(t)}(\pi^{(t)}) \leq \frac{8\eta}{2t-1}$. Summing and applying (9),

$$\sum_{t=1}^{T}\Big(f^{(t)}(\pi^*) - f^{(t)}(\pi^{(t)})\Big) \leq \sum_{t=1}^{T}\frac{8\eta}{2t-1} \leq 8\eta(1+\log T).$$

Dividing by T yields the claimed $\mathcal{O}\big(\frac{\eta(1+\log T)}{T}\big)$ bound. For the fitted Q-functions in GAMD, which take values in $[0,1]$, the same proof retains the sharper cumulative bound $2\eta(1+\log T)$.

□

B.3 Proof of Theorem 5.2

In this section, we provide the full sample complexity proof for General-sum Anchored Mirror Descent (GAMD, Algorithm 2). The algorithm outputs a time-averaged joint policy $\bar{\pi}$, where $\bar{\pi}(\boldsymbol{a}|x) = \frac{1}{T}\sum_{t=1}^{T}\prod_{i=1}^{m}\pi_i^{(t)}(a_i|x)$ is a correlated joint distribution representing an approximate Coarse Correlated Equilibrium (CCE).

We evaluate the convergence of the algorithm using the Total Exploitability Gap. For the learned policy $\bar{\pi}$, the expected gap is defined as the sum of the unilateral improvements available to all m players:

$$\mathrm{Gap}_{\mathrm{CCE}}(\bar{\pi}) := \sum_{i=1}^{m}\mathbb{E}_{x\sim\rho}\left[\max\left\{0, V_i^{\dagger,\bar{\pi}_{-i}}(x) - V_i^{\bar{\pi}}(x)\right\}\right].$$

Step 1: Gap Decomposition via Estimated Values

To decouple the optimization error introduced by running a finite number of Mirror Descent steps T from the statistical error caused by finite offline samples n, we pivot through the estimated value functions $\hat{V}$ and $\hat{Q}$ constructed by the algorithm.

For each player $i \in [m]$, we decompose their unilateral exploitability gap identically to Lemma 4.1:

$$V_i^{\dagger,\bar{\pi}_{-i}}(x) - V_i^{\bar{\pi}}(x) = \underbrace{\left(\hat{V}_i^{\dagger,\bar{\pi}_{-i}}(x) - \hat{V}_i^{\bar{\pi}}(x)\right)}_{\text{Term I: Empirical Optimization Gap}} + \underbrace{\left(V_i^{\dagger,\bar{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\bar{\pi}_{-i}}(x)\right)}_{\text{Term II: Evaluation Error of Best Response}} + \underbrace{\left(\hat{V}_i^{\bar{\pi}}(x) - V_i^{\bar{\pi}}(x)\right)}_{\text{Term III: Evaluation Error of Joint Policy}}.$$

STEP 2: BOUNDING THE OPTIMIZATION ERROR (TERM I)

Term I represents the signed empirical deviation gain of the time-averaged policy $\bar{\pi}$ on the estimated bandit defined by $\hat{Q}$. We formally bound this by connecting the exploitability of the average joint policy to the external regret of the independent Mirror Descent iterates.

For any context x, we evaluate Player i's empirical suboptimality against the time-averaged opponents $\bar{\pi}_{-i}$. For notational clarity, we define the marginalized empirical Q-function as $\bar{Q}_i(x, a_i) \coloneqq \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}[\hat{Q}_i(x, a_i, \boldsymbol{a}_{-i})]$. Because the estimated Q-function is linear in the opponents' distribution, we have exact equality for the expected payoff:

$$\bar{Q}_i(x, a_i) = \frac{1}{T}\sum_{t=1}^{T} \bar{Q}_i^{(t)}(x, a_i).$$

The on-policy reward instead evaluates the joint mixture:

$$\mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}(\cdot|x)}[\hat{Q}_i(x, \boldsymbol{a})] = \frac{1}{T}\sum_{t=1}^{T} \bar{Q}_i^{(t)}(x, \pi_i^{(t)}).$$

Furthermore, by the convexity of the KL-divergence, the regularizer of the average policy is bounded by the average of the regularizers:

$$\mathrm{KL}(\bar{\pi}_i(\cdot|x) \,\|\, \pi_i^{\mathrm{ref}}(\cdot|x)) \le \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(\pi_i^{(t)}(\cdot|x) \,\|\, \pi_i^{\mathrm{ref}}(\cdot|x)).$$

Applying these properties, the signed empirical deviation gain for Player i at context x is bounded by their average external regret against the historical sequence of policies:

$$\begin{aligned}
\text{Term I}_i(x) &= \max_{\pi_i} \left[\bar{Q}_i(x, \pi_i) - \eta^{-1}\mathrm{KL}(\pi_i \| \pi_i^{\mathrm{ref}})\right] \\
&\quad - \left[\mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}(\cdot|x)}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1}\mathrm{KL}(\bar{\pi}_i \| \pi_i^{\mathrm{ref}})\right] \\
&\le \max_{\pi_i \in \Delta(\mathcal{A}_i)} \frac{1}{T}\sum_{t=1}^{T} \underbrace{\left(\bar{Q}_i^{(t)}(x, \pi_i) - \eta^{-1}\mathrm{KL}(\pi_i \| \pi_i^{\mathrm{ref}})\right)}_{\coloneqq f_{i,x}^{(t)}(\pi_i)} \\
&\quad - \frac{1}{T}\sum_{t=1}^{T} \underbrace{\left(\bar{Q}_i^{(t)}(x, \pi_i^{(t)}) - \eta^{-1}\mathrm{KL}(\pi_i^{(t)} \| \pi_i^{\mathrm{ref}})\right)}_{= f_{i,x}^{(t)}(\pi_i^{(t)})}.
\end{aligned}$$

At each iteration t, Algorithm 2 independently updates the policies $\pi_i^{(t)}$ via KL-regularized Online Mirror Descent (OMD) on the contextual objectives $f_{i,x}^{(t)}$. To bound the external regret of this update sequence, we rely on foundational results from the Online Convex Optimization (OCO) literature. While standard online learning algorithms typically suffer an $\tilde{\mathcal{O}}(\sqrt{T})$ cumulative regret bound, the explicit presence of the KL-divergence penalty in our formulation strictly transforms the objective

into a strongly concave function (when maximizing payoff). As established in classical OCO texts, applying no-regret learning to a strongly concave objective with the decaying stepsize schedule $\gamma_t = 1/t$ matching the update in Algorithm 2 attains a cumulative regret of $\tilde{\mathcal{O}}(\eta(1+\log T))$.

By the standard online-to-offline reduction for normal-form games, the total empirical optimization gap is bounded by the expected context-wise average regret. Using Lemma 5.1:

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[\text{Term I}_i] &\le \mathbb{E}_{x\sim\rho}\left[\max_{\pi_i}\frac{1}{T}\sum_{t=1}^{T}\left(f_{i,x}^{(t)}(\pi_i) - f_{i,x}^{(t)}(\pi_i^{(t)})\right)\right] \\ &\le \mathcal{O}\left(\frac{\eta(1+\log T)}{T}\right) = \widetilde{\mathcal{O}}\left(\frac{\eta}{T}\right).\end{aligned}$$

The same argument holds before averaging over contexts. Since the marginalized fitted Q-functions take values in $[0,1]$, Appendix B.2 gives, for every x,

$$\text{Term I}_i(x) \le \frac{2\eta(1+\log T)}{T}.$$

Summing this optimization error over all m players yields a total empirical optimization gap that strictly decays at a fast rate of $\widetilde{\mathcal{O}}(m\eta/T)$.

Step 3: Unrolling the Statistical Errors (Terms II and III)

We now bound the statistical evaluation errors (Term II and Term III). We introduce the regression error for Player i evaluated under the algorithm's estimated Q-functions:

$$\mathcal{Z}_i(x,\boldsymbol{a}) \coloneqq \hat{Q}_i(x,\boldsymbol{a}) - r_i^\star(x,\boldsymbol{a}).$$

Evaluating Term II (Best Response Error). For Term II, we evaluate the suboptimality of the estimated best response. Let $\Delta_i^\dagger(x) \coloneqq V_i^{\dagger,\bar{\pi}_{-i}}(x) - \hat{V}_i^{\dagger,\bar{\pi}_{-i}}(x)$. By applying the 1-smoothness of the log-partition function exactly as derived in the idealized proof (Equation 2), we obtain the local bound:

$$\Delta_i^\dagger(x) \le \mathbb{E}_{a_i\sim\bar{\pi}_i^\dagger(\cdot|x)}\left[Q_i^{\dagger,\bar{\pi}_{-i}}(x,a_i) - \bar{Q}_i(x,a_i)\right] + \frac{\eta}{2}\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2,$$

where $\bar{\pi}_i^\dagger$ is the empirical best-response policy.

By direct algebraic substitution, the linear Q-value difference is strictly equal to the negative regression error:

$$Q_i^{\dagger,\bar{\pi}_{-i}}(x,a_i) - \bar{Q}_i(x,a_i) = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}\left[-\mathcal{Z}_i(x,\boldsymbol{a})\right].$$

Taking the expectation over the context distribution $x\sim\rho$, Term II is bounded by:

$$\mathbb{E}_{x\sim\rho}[\text{Term II}_i(x)] \le \mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\bar{\pi}_{-i})}\left[-\mathcal{Z}_i(x,\boldsymbol{a})\right] + \frac{\eta}{2}\mathbb{E}_{x\sim\rho}\left[\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2\right].$$

Evaluating Term III (Joint Policy Error). Term III measures the evaluation error of the algorithm's joint policy: $\hat{V}_i^{\bar{\pi}}(x) - V_i^{\bar{\pi}}(x)$. Because the identical policy $\bar{\pi}$ is used for both the empirical and true values, the KL-regularization terms exactly cancel out. Direct substitution yields:

$$\mathbb{E}_{x\sim\rho}[\text{Term III}_i(x)] = \mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\bar{\pi}}\left[\mathcal{Z}_i(x,\boldsymbol{a})\right].$$

Step 4: The Distribution Mismatch and Absolute Errors

In the analysis of exact Nash Equilibria, the joint policy is a pure product distribution ($\hat{\pi} = \hat{\pi}_i \times \hat{\pi}_{-i}$) and the learned policy is identically the best response ($\hat{\pi}_i = \hat{\pi}_i^\dagger$). This structural alignment causes the evaluation distributions to match perfectly, allowing the linear errors $-\mathcal{Z}_i$ and $+\mathcal{Z}_i$ to cancel entirely.

However, the time-averaged policy $\bar{\pi}$ produced by GAMD can be a *correlated* joint distribution. Consequently, the independent product of the best response and the marginalized opponents need not coincide with the actual correlated joint policy:

$$\bar{\pi}_i^\dagger(a_i|x) \times \bar{\pi}_{-i}(\boldsymbol{a}_{-i}|x) \neq \bar{\pi}(\boldsymbol{a}|x) \quad \text{in general.}$$

Because the expectations for the linear regression errors need not be taken over the same distribution, they need not cancel. To establish a rigorous upper bound, we must bound their absolute values. By the triangle inequality, we maintain the exact squared error term:

$$\text{Term II}_i(x) + \text{Term III}_i(x) \leq \mathbb{E}_{\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\bar{\pi}_{-i})}\left[|\mathcal{Z}_i(x,\boldsymbol{a})|\right] + \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}}\left[|\mathcal{Z}_i(x,\boldsymbol{a})|\right] + \frac{\eta}{2}\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2.$$

Step 5: Bounding the L_∞ Error and Applying Cauchy-Schwarz

To bridge the pointwise absolute errors to our least-squares statistical oracle, we apply Cauchy-Schwarz ($\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$).

For the squared Q-value difference, we must unroll the L_∞ norm into the pointwise squared regression errors. Taking the absolute value of the Q-value difference identity yields:

$$\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty \leq \max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}\left[|\mathcal{Z}_i(x,\boldsymbol{a})|\right].$$

Squaring both sides and applying Jensen's inequality for the inner expectation bounds the L_∞ norm strictly by the expected squared regression error under a maximally adverse unilateral deviation:

$$\left\|Q_i^{\dagger,\bar{\pi}_{-i}}(x,\cdot) - \bar{Q}_i(x,\cdot)\right\|_\infty^2 \leq \max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right].$$

Summing the expectations over all players and applying Cauchy-Schwarz to the linear terms, we isolate the statistical gap strictly in terms of squared errors:

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E}_{x\sim\rho}[\text{Term II}_i(x) + \text{Term III}_i(x)] \\
&\leq \sum_{i=1}^m \left(\sqrt{\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\bar{\pi}_{-i})}\left[\mathcal{Z}_i^2\right]} + \sqrt{\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\bar{\pi}}\left[\mathcal{Z}_i^2\right]}\right) \\
&\quad + \frac{\eta}{2}\sum_{i=1}^m \mathbb{E}_{x\sim\rho}\left[\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right]\right].
\end{aligned}$$

Step 6: Distribution Shift and Data Coverage

We bound each expected term by shifting the evaluation distribution to the offline dataset distribution μ via the Reference-Anchored Unilateral Concentrability (Assumption 3.2).

Crucially, because GAMD explicitly constructs the correlated policy via exponentiated Mirror Descent updates, we can strictly bound the density ratio of every individual iterate $\pi_i^{(t)}$ relative to the reference policy. Initialized at $\pi_i^{(1)} = \pi_i^{\text{ref}}$, the decaying-stepsize closed-form update (Step 12, Algorithm 2) unrolls to the Gibbs form anchored to the reference policy:

$$\pi_i^{(t+1)}(a_i|x) \propto \pi_i^{\text{ref}}(a_i|x)\exp\left(\eta\tilde{Q}_i^{(t)}(x,a_i)\right), \quad \tilde{Q}_i^{(t)}(x,a_i) := \frac{1}{t}\sum_{k=1}^t \bar{Q}_i^{(k)}(x,a_i).$$

Since the empirical Q-functions are bounded in $[0,1]$, we have $\|\bar{Q}_i^{(k)}\|_\infty \leq 1$, so the arithmetic average $\tilde{Q}_i^{(t)} \in [0,1]$. Evaluating the normalization constant yields the exact pointwise bound for any action:

$$\frac{\pi_i^{(t+1)}(a_i|x)}{\pi_i^{\text{ref}}(a_i|x)} = \frac{\exp\left(\eta\tilde{Q}_i^{(t)}(x,a_i)\right)}{\mathbb{E}_{a'\sim\pi_i^{\text{ref}}}\exp\left(\eta\tilde{Q}_i^{(t)}(x,a')\right)} \leq \frac{\exp(\eta)}{1} = \exp(\eta).$$

Because the marginalized opponent policy $\bar{\pi}_{-i}$ and the joint policy $\bar{\pi}$ are convex combinations of products of these iterates, they preserve the corresponding product bounds. To unify our notation across the m players, we define a single global distribution shift constant:

$$C_{\text{shift}} \coloneqq \exp(\eta m).$$

For the unilateral deviation trajectories, we encounter two distinct expectations: one over the empirical best response profile $(\bar{\pi}_i^\dagger, \bar{\pi}_{-i})$ and one involving a worst-case maximization $\max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}[\cdot]$. In both cases, shifting the $m-1$ opponents from $\bar{\pi}_{-i}$ to the reference policy π_{-i}^{ref} produces a joint density ratio bounded by $\exp(\eta(m-1)) \le C_{\text{shift}}$.

For the first term, the profile $(\bar{\pi}_i^\dagger, \pi_{-i}^{\text{ref}})$ strictly belongs to $\Pi_{\text{ref}-\text{uni}}$, allowing us to directly apply the concentrability coefficient C_{uni}:

$$\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\bar{\pi}_{-i})}\left[\mathcal{Z}_i^2\right] \le C_{\text{shift}}\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\pi_{-i}^{\text{ref}})}\left[\mathcal{Z}_i^2\right] \le C_{\text{shift}}C_{\text{uni}}\mathbb{E}_{\mu}\left[\mathcal{Z}_i^2\right].$$

For the second term, to formalize the application of our data coverage assumption on the maximization, we construct a deterministic greedy policy $\tilde{\pi}_i(x) \coloneqq \operatorname{argmax}_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}^{\text{ref}}}[\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i})^2]$. The resulting joint profile $\tilde{\pi} = (\tilde{\pi}_i, \pi_{-i}^{\text{ref}})$ also strictly belongs to $\Pi_{\text{ref}-\text{uni}}$, yielding:

$$\begin{aligned}
\mathbb{E}_{x\sim\rho}\left[\max_{a_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right]\right] &\le C_{\text{shift}}\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\tilde{\pi}(\cdot|x)}\left[\mathcal{Z}_i(x,\boldsymbol{a})^2\right] \\
&= C_{\text{shift}}\mathbb{E}_{(x,\boldsymbol{a})\sim\mu}\left[\frac{\rho(x)\tilde{\pi}(\boldsymbol{a}|x)}{\mu(x,\boldsymbol{a})}\mathcal{Z}_i(x,\boldsymbol{a})^2\right] \\
&\le C_{\text{shift}}C_{\text{uni}}\mathbb{E}_{\mu}\left[\mathcal{Z}_i^2\right].
\end{aligned}$$

Similarly, for the on-policy evaluation over the correlated joint policy $\bar{\pi}$, we shift the actions of all m players to the joint reference policy $\pi^{\text{ref}} = \prod_{j=1}^m \pi_j^{\text{ref}}$. The total density ratio is strictly bounded by $\prod_{j=1}^m \exp(\eta) = \exp(\eta m) = C_{\text{shift}}$. Because the joint reference policy π^{ref} is explicitly covered by the unilateral assumption set $\Pi_{\text{ref}-\text{uni}}$ (representing the trivial case where the deviation policy equals the reference policy), we directly apply C_{uni} to bound the joint evaluation:

$$\mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\bar{\pi}}\left[\mathcal{Z}_i^2\right] \le C_{\text{shift}}C_{\text{uni}}\mathbb{E}_{\mu}\left[\mathcal{Z}_i^2\right].$$

By applying the fast-rate guarantee for least-squares regression, the population squared regression error is bounded by $\mathcal{O}\left(\frac{\log(m|\mathcal{Q}_i|/\delta)}{n}\right)$ simultaneously for all players on the event of probability at least $1-\delta$ used in Appendix B.1. Substituting this statistical rate into the square-root and linear components yields:

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E}_{x\sim\rho}[\text{Term II}_i(x) + \text{Term III}_i(x)] \\
&\le \sum_{i=1}^m 2\sqrt{C_{\text{shift}}C_{\text{uni}}\mathcal{O}\left(\frac{\log(m|\mathcal{Q}_i|/\delta)}{n}\right)} \\
&\quad + \sum_{i=1}^m \mathcal{O}\left(\frac{\eta C_{\text{shift}}C_{\text{uni}}\log(m|\mathcal{Q}_i|/\delta)}{n}\right).
\end{aligned}$$

STEP 7: FINAL BOUND SYNTHESIS

Combining the pointwise bound from Step 2 with Steps 4–5 gives a nonnegative upper bound for each player's deviation gain:

$$\begin{aligned}&\max\left\{0, V_i^{\dagger,\bar{\pi}_{-i}}(x) - V_i^{\bar{\pi}}(x)\right\}\\&\leq \frac{2\eta(1+\log T)}{T} + \mathbb{E}_{\boldsymbol{a}\sim(\bar{\pi}_i^\dagger,\bar{\pi}_{-i})}[|\mathcal{Z}_i(x,\boldsymbol{a})|] + \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}}[|\mathcal{Z}_i(x,\boldsymbol{a})|]\\&\quad + \frac{\eta}{2}\max_{a_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\bar{\pi}_{-i}}[\mathcal{Z}_i(x,\boldsymbol{a})^2].\end{aligned}$$

Taking the context expectation, summing over players, and using the Cauchy–Schwarz and coverage bounds in Steps 5–6 yields

$$\begin{aligned}\mathrm{Gap}_{\mathrm{CCE}}(\bar{\pi}) \leq \frac{2m\eta(1+\log T)}{T} + 2\sum_{i=1}^{m}\sqrt{C_{\mathrm{shift}}C_{\mathrm{uni}}\mathbb{E}_\mu[\mathcal{Z}_i^2]}\\+ \frac{\eta}{2}C_{\mathrm{shift}}C_{\mathrm{uni}}\sum_{i=1}^{m}\mathbb{E}_\mu[\mathcal{Z}_i^2].\end{aligned}$$

Summing the empirical optimization error from Step 2 and the combined statistical evaluation error from Step 6, we obtain the final bound on the Total Exploitability Gap on the same high-probability event. Because the bounds on the absolute linear errors decay at $\tilde{\mathcal{O}}(1/\sqrt{n})$, for fixed parameters they asymptotically dominate the bound on the $\tilde{\mathcal{O}}(1/n)$ squared Q-difference term.

$$\begin{aligned}&\mathrm{Gap}_{\mathrm{CCE}}(\bar{\pi}) = \sum_{i=1}^{m}\mathbb{E}_{x\sim\rho}\left[\max\left\{0, \text{Term I}_i(x) + \text{Term II}_i(x) + \text{Term III}_i(x)\right\}\right]\\&\leq \widetilde{\mathcal{O}}\left(\frac{m\eta}{T}\right) + \widetilde{\mathcal{O}}\left(m\sqrt{\frac{C_{\mathrm{shift}}C_{\mathrm{uni}}\log(m|\mathcal{Q}|/\delta)}{n}}\right)\\&\quad + \widetilde{\mathcal{O}}\left(\frac{m\eta C_{\mathrm{shift}}C_{\mathrm{uni}}\log(m|\mathcal{Q}|/\delta)}{n}\right).\end{aligned}$$

Here, $C_{\mathrm{shift}} \coloneqq \exp(\eta m)$, and C_{uni} is the coefficient for unilateral coverage. Setting the number of Mirror Descent iterations such that $T \geq \sqrt{n}$ ensures that the empirical optimization error is of no larger order than the statistical evaluation error. The final sample complexity is therefore:

$$\mathrm{Gap}_{\mathrm{CCE}}(\bar{\pi}) \leq \widetilde{\mathcal{O}}\left(\frac{m(\eta+\sqrt{C_{\mathrm{shift}}C_{\mathrm{uni}}})}{\sqrt{n}} + \frac{m\eta C_{\mathrm{shift}}C_{\mathrm{uni}}}{n}\right).$$

This result demonstrates that while GAMD attains a fast $\widetilde{\mathcal{O}}(m\eta/T)$ empirical optimization rate, this analysis of the distribution mismatch for a correlated joint policy gives the statistical upper bound displayed above.